Effektiv und nutzerfreundlich

Einsatz von semantischen Technologien
und Usability-Methoden
zur Verbesserung der medizinischen Literatursuche

Waldemar Dzeyk

Köln 2010

Das Projekt wurde gefördert durch Bund und Länder

im Rahmen des Paktes für Forschung und Innovation

Verlag: Deutsche Zentralbibliothek für Medizin
Gleueler Str. 60
50931 Köln

ISBN 978-3-9808397-5-4

Danksagung

Ein so umfangreiches Vorhaben wie das MorphoSaurus-Projekt kann nur gelingen, wenn es auf vielen Schultern verteilt ist. Ich möchte mich an dieser Stelle deshalb bei allen bedanken, die zum Gelingen des Projekts beigetragen haben. Zunächst bei meinem sehr engagierten Evaluationsteam: Maarit Stoor und Stefanie Paschke sowie bei den Studentinnen Natascha Dahmen und Katja Köhl, die zeitweise im Projekt mitarbeiteten. Des Weiteren gilt der Dank auch Averbis in Person von Kornél Markó und Philipp Daumke, die kompetent und schnell alle Rückmeldungen des Teams umgesetzt haben. Ich möchte mich bei allen Mitarbeitern der ZB MED bedanken, die auf die eine oder andere Weise am Erfolg des MorphoSaurus-Projekts beteiligt waren, besonders auch beim EDV-Team sowie bei Herrn Korwitz, dem Direktor der ZB MED. Mein besonderer Dank gilt den Teilnehmer/innen der Usability-Untersuchungen, die uns viele neue Einsichten eröffnet haben. Nicht zuletzt möchte ich mich bei Barbara Cogel und Stefanie Paschke für die Durchsicht der Arbeit herzlich bedanken. Auch wäre ohne die finanzielle Unterstützung aus dem „Pakt für Forschung und Innovation“ das Projekt und die vorliegende Monografie nicht möglich gewesen.

Zusammenfassung

In der vorliegenden Arbeit werden die Ergebnisse des *MorphoSaurus-Projekts* der Deutschen Zentralbibliothek für Medizin (ZB MED)[1] vorgestellt. Ziel des Forschungsprojekts war die substanzielle Verbesserung des Information-Retrievals der medizinischen Suchmaschine MEDPILOT[2] mithilfe computerlinguistischer Ansätze sowie die Optimierung der Gebrauchstauglichkeit (Usability) der Suchmaschinenoberfläche.

Das Projekt wurde in Kooperation mit der Averbis GmbH[3] aus Freiburg im Zeitraum von Juni 2007 bis Dezember 2008 an der ZB MED in Köln durchgeführt. Ermöglicht wurde die Realisierung des Projekts durch eine Förderung des Paktes für Forschung und Innovation[4]. Während Averbis die MorphoSaurus-Technologie zur Verarbeitung problematischer Sprachaspekte von Suchanfragen einbrachte und wesentliche Datenbanken der ZB MED in ein Testsystem mit moderner Suchmaschinentechnologie implementierte, evaluierte ein Team der ZB MED das Potenzial dieser Technologie.

Neben einem Vergleich der Leistungsfähigkeit zwischen der bisherigen MEDPILOT-Suche und der neuen Sucharchitektur wurde ein Benchmarking mit konkurrierenden Suchmaschinen wie PubMed, Scirus, Google und Google Scholar sowie GoPubMed durchgeführt. Für die Evaluation wurden verschiedene Testkollektionen erstellt, deren Items bzw. Suchphrasen aus einer Inhaltsanalyse realer Suchanfragen des MEDPILOT-Systems gewonnen wurden. Eine Überprüfung der Relevanz der Treffer der Testsuchmaschine als wesentliches Kriterium für die Qualität der Suche zeigte folgendes Ergebnis: Durch die Anwendung der MorphoSaurus-Technologie[5] ist eine im hohen Maße unabhängige Verarbeitung fremdsprachlicher medizinischer Inhalte möglich geworden. Darüber hinaus zeigt die neue Technik insbesondere dort ihre Stärken, wo es um die gleichwertige Verarbeitung von Laien- und Expertensprache, die Analyse von Komposita, Synonymen und grammatikalischen Varianten geht. Zudem sind Module zur Erkennung von Rechtschreibfehlern und zur Auflösung von

[1] www.zbmed.de

[2] www.medpilot.de

[3] www.averbis.de

[4] www.bmbf.de

[5] MorphoSaurus ist ein Akronym für "MORPHem-theSAURUS": ein Thesaurus, der aus Morphemen bzw. Subwörtern besteht.

Akronymen und medizinischen Abkürzungen implementiert worden, die eine weitere Leistungssteigerung des Systems versprechen.

Ein Vergleich auf der Basis von MEDLINE-Daten zeigte: Den Suchmaschinen MEDPILOT, PubMed, GoPubMed und Scirus war die Averbis-Testsuchumgebung klar überlegen. Die Trefferrelevanz war größer, es wurden insgesamt mehr Treffer gefunden und die Anzahl der Null-Treffer-Meldungen war im Vergleich zu den anderen Suchmaschinen am geringsten.

Bei einem Vergleich unter Berücksichtigung aller verfügbaren Quellen gelang es mithilfe der MorphoSaurus-Technik – bei wesentlich geringerem Datenbestand – ähnlich gute Resultate zu erzielen, wie mit den Suchmaschinen Google oder Google Scholar. Die Ergebnisse der Evaluation lassen den Schluss zu, dass durch den MorphoSaurus-Ansatz die Leistungsfähigkeit von Google oder Google Scholar im Bereich der medizinischen Literatursuche durch eine Erweiterung der vorhandenen Datenbasis sogar deutlich übertroffen werden kann.

Zusätzlich zu den Retrieval-Tests wurde eine Usability-Untersuchung der Testsuchmaschine mit Probanden aus der Medizin durchgeführt. Die Testpersonen attestierten dem Suchinterface eine hohe Gebrauchstauglichkeit und Nützlichkeit. Der szenariobasierte Usability-Test hat zudem gezeigt, dass die Testpersonen bzw. User[6] die integrierten Unterstützungsmaßnahmen zur Erhöhung der Benutzerfreundlichkeit während der Suche als sehr positiv und nützlich bewerten. In der Testsuchmaschine wurde diese Unterstützung z. B. durch das *Aufklappen und Präsentieren von verwandten MeSH- und ICD-10-Begriffen* realisiert. Die Einführung eines *Schiebereglers* zur effektiven Eingrenzung des Suchraums wurde ebenfalls überwiegend positiv bewertet. Zudem wurden nach Abschicken der Suchanfrage sogenannte *Verwandte Suchbegriffe* aus verschiedenen medizinischen Teilbereichen angezeigt. Diese Facetten-Funktion diente der Eingrenzung bzw. Verfeinerung der Suche und wurde von den Testpersonen mehrheitlich als ein sinnvolles Hilfsangebot bewertet.

Insgesamt stellt das MorphoSaurus-Projekt – mit seinem spezifischen Ansatz – ein gelungenes Beispiel für die Innovationsfähigkeit von Bibliotheken im Bereich der öffentlichen Informationsversorgung dar. Durch die mögliche Anpassung der MorphoSaurus-Technologie mittels fachspezifischer Thesauri ist zudem eine hohe Anschlussfähigkeit für Suchmaschinenprojekte anderer Inhaltsdomänen gegeben.

[6] Werden Personenbezeichnungen aus Gründen der besseren Lesbarkeit lediglich in der männlichen oder weiblichen Form verwendet, so schließt dies das jeweils andere Geschlecht mit ein.

Inhaltsverzeichnis

1 Einleitung

Die Menge verfügbarer wissenschaftlicher Literatur wächst in einem kaum zu überschauenden Ausmaß. Nach Price (1963) soll sich der Umfang wissenschaftlicher Literatur alle 15 Jahre verdoppeln. Nach bisherigem Wissensstand ist dies auch für die Entwicklung der medizinischen Literatur anzunehmen. So wächst die größte medizinische Datenbank MEDLINE [7] pro Werktag etwa um 2000 bis 4000 Artikel[8]. Ende 2008 enthielt MEDLINE über 18,5 Mio. Einträge. Auf der Seite der Nutzer ist ebenfalls eine stetige Zunahme der Nachfrage nach medizinischer Literatur zu verzeichnen: Monatlich werden in MEDLINE über 70 Millionen Recherchen durchgeführt[9].

Angesichts dieser Informationsmenge gibt es einen großen Bedarf an Orientierungs- und Unterstützungsangeboten bei der Suche nach qualifizierter wissenschaftlicher Information. Aufgrund der technischen Entwicklungen in den letzten Jahren profitieren die Nutzer medizinischer Informationsdienste dabei von der steigenden Leistungsfähigkeit webbasierter Suchmaschinen. Medizinische Informationen werden vor allem in Wissenschaft, Klinik und Praxis erzeugt und nachgefragt. Der effiziente und qualifizierte Zugriff auf relevante Literatur kann dazu beitragen, Innovationsprozesse in der medizinischen Forschung und im klinischen Alltag zu beschleunigen, indem stets aktualisierte Informationen bereitgehalten und zugänglich gemacht werden. Darüber hinaus stellen Suchmaschinen eine wichtige Entscheidungs- und Orientierungshilfe dar: Sie können Mediziner im Prozess der Diagnosefindung unterstützen sowie bei der Auswahl einer geeigneten Therapie und Nachsorge von Krankheiten. Zudem trägt das medizinische Information-Retrieval dazu bei, Kosten im Gesundheitssystem zu reduzieren, indem es hilft, Entscheidungen zu vermeiden, die aufgrund veralteten Wissens getroffen werden (vgl. dazu Hersh, 2004).

Innerhalb der letzten Dekade hat sich die im webbasierten Information-Retrieval eingesetzte Suchmaschinentechnologie radikal gewandelt. Mit dem Siegeszug von Google wurde

[7] MEDLINE (MEDdical Literature Analysis and Retrieval System OnLINE) ist mit rund 18,5 Millionen Artikeln aus etwa 4800 Zeitschriften die umfangreichste bibliografische Datenbank in der Medizin (Stand: 12.2008). Online unter http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed

[8] http://www.nlm.nih.gov/pubs/factsheets/medline.html

[9] Monatlicher Durchschnitt in 2007 (http://www.nlm.nih.gov/bsd/medline_growth.html)

die algorithmische, indexbasierte Suche mehr und mehr zum technologischen Standard. Die Entwicklung im Bereich der wissenschaftlichen Informationsversorgung trägt diesem Trend Rechnung. Daher sind in den letzten Jahren verstärkt Bemühungen zur technologischen Neuausrichtung von „Virtuellen Fachbibliotheken“ unternommen worden.

Da sich das Angebot dieser elektronischen Fachinformationen in der Regel aus einer Zusammenstellung vielfältiger Quellen speist, basierten diese Systeme bisher häufig auf dem Prinzip der Metasuche. Angestoßen durch eine Suchanfrage werden hier sukzessive bzw. parallel viele unterschiedliche Quellen abgefragt und anschließend in einer Trefferliste zusammengeführt. Dabei sind diese Systeme nach heutigen Maßstäben eher langsam und erfüllen die Erwartung hinsichtlich eines intelligenten Information-Retrievals kaum. Zunehmend werden diese Suchdienste jedoch in moderne Suchmaschinenarchitekturen überführt (vgl. hierzu z. B. die Entwicklung von vascoda bei Krause & Mayr, 2006).

Zu den wesentlichen Vorteilen moderner, algorithmenbasierter Suchumgebungen zählen ihre große Schnelligkeit aufgrund eines gemeinsamen Indexes der zur Verfügung stehenden Quellen sowie die verbesserten Möglichkeiten und Algorithmen zum Auffinden relevanter Dokumente (z. B. der „Pagerank“[10] bei Google).

Alternative Suchmaschinenansätze finden sich etwa im Bereich der sozialen Suchdienste, die sich durch eine Verknüpfung von technischen und sozialen Komponenten auszeichnen. Dabei wird das jeweilige Ranking nicht allein durch einen Algorithmus bestimmt, sondern durch eine Community von Freiwilligen, welche die Inhalte klassifizieren (bzw. ‚taggen‘) und so Einfluss auf die Gewichtung von gesuchten Inhalten nehmen (z. B. durch ‚social bookmarking‘ oder durch sogenannte ‚tag clouds‘).

Ein weiterer Ansatz mit großem Zukunftspotenzial ist die sogenannte *semantische Suche*. Hierbei wird versucht, die Inhalte des Webs so aufzubereiten, dass diese in ihren wesentlichen

[10] „Der PageRank-Algorithmus ist eine spezielle Methode, die Linkpopularität einer Seite bzw. eines Dokumentes festzulegen. Das Grundprinzip lautet: Je mehr Links auf eine Seite verweisen, umso höher ist das Gewicht dieser Seite. Je höher das Gewicht der verweisenden Seiten ist, desto größer ist der Effekt. Das Ziel des Verfahrens ist es, die Links dem Gewicht entsprechend zu sortieren, um so eine Ergebnisreihenfolge bei einer Suchabfrage herzustellen, d.h. Links zu wichtigeren Seiten weiter vorne in der Ergebnisliste anzuzeigen“ Seite „PageRank“. In: Wikipedia, Die freie Enzyklopädie. Bearbeitungsstand: 18. Juli 2009, 10:58 UTC. URL: http://de.wikipedia.org/w/index.php?title=PageRank&oldid=62335774 (Abgerufen: 24. Juli 2009, 13:34 UTC).

semantischen Dimensionen maschinenlesbar werden. Dazu werden die Inhalte von den Autoren der Webquellen so mit Metabeschreibungen angereichert, dass die zentralen Inhaltskonzepte und ihre Relationen (in Form von Ontologien) zueinander von Suchmaschinen erfasst werden können. Auf diese Weise wird es möglich, die Relevanz der in Suchmaschinen angezeigten Treffer erheblich zu steigern, weil sich hierdurch die Unschärfe der Trefferrückmeldungen effizient reduzieren lässt (vgl. Berners-Lee, Hendler & Lassila, 2001). Semantische Suchmaschinen stehen jedoch erst am Anfang ihrer Entwicklung. Mit erfolgreichen marktreifen Systemen ist erst in einigen Jahren zu rechnen. Ein aktueller Überblick über semantische Suchansätze im biomedizinischen Bereich findet sich z. B. bei Dietze und Schroeder (2009). Einzelheiten zu den jeweiligen Suchmaschinenansätzen sollen an dieser Stelle nicht weiter ausgeführt werden. Eine allgemeine Übersicht über die aktuelle Suchmaschinenlandschaft geben Maaß, Skusa, Heß und Pietsch (2009) oder Grießbaum, Bekavac und Rittberger (2009).

Universalsuchmaschinen wie z. B. Google oder Yahoo versuchen potenziell, das gesamte Web zu erfassen. Sie bieten zwar auch einen Zugang zu wissenschaftlichen Informationen, doch ist durch ihre Orientierung am Durchschnittsuser ihre Fähigkeit zur domänenspezifischen Suche sehr begrenzt. Wie Lewandowski (2009) feststellt, besteht, trotz der Erfolge der algorithmenbasierten Suchmaschinen, die Notwendigkeit, für bestimmte Bereiche des Webs Spezialsuchmaschinen einzusetzen. Insbesondere für den Wissenschaftsbereich lassen sich einige Defizite bei den großen Suchmaschinen feststellen: Inhalte des sogenannten „Invisible Webs“ werden von den Universalsuchmaschinen nicht oder nur unvollständig gefunden. Dabei handelt es sich zumeist um Inhalte von Datenbanken, wie sie z. B. in Form von bibliografischen Einträgen bei vielen Bibliotheken vorliegen. Zudem stößt die Analyse der Nutzeranfragen schnell an die Grenzen dieser Suchmaschinen. Gerade durch neuere Entwicklungen in der Computerlinguistik ist es aber seit kurzem möglich, eine Reihe spezieller Probleme des medizinischen Information-Retrievals viel effizienter zu bewältigen. Dazu gehören z. B. die korrekte Verarbeitung von Synonymen und lexikalischen Varianten, das Auflösen von Abkürzungen, die gleichwertige Verarbeitung von experten- oder laiensprachlichen Suchanfragen sowie die Fähigkeit zur Verarbeitung fremdsprachiger Suchterme (vgl. Zaiß, Graubner, Ingenerf, Leiner, Lochmann, Schopen, Schrade & Schulz, 2004).

Im Jahr 2007 ist deshalb an der Deutschen Zentralbibliothek für Medizin (ZB MED) im Rahmen der Förderung durch den „Pakt für Forschung und Innovation“[11] das Forschungsprojekt *MorphoSaurus* angestoßen worden. Primäres Ziel dieses Projektes war die Verbesserung der MEDPILOT-Suche mithilfe moderner Suchmaschinentechnologie und computerlinguistischer Ansätze. MEDPILOT ist ein Serviceangebot der ZB MED. In seiner bisherigen Form handelt es sich um eine medizinische Metasuchmaschine (Stand Dezember 2008) bzw. ein Fachportal zur Recherche medizinischer Informationen. MEDPILOT bietet den Zugang zu zahlreichen kostenlosen, aber auch zu kostenpflichtigen Informationsquellen bis hin zum Volltext (www.medpilot.de).

Unter dem Titel „Optimierung der MEDPILOT-Recherche: Mehrsprachigkeit und Normalisierung sprachlicher Varianten“ (bzw. MorphoSaurus-Projekt) wurde im Juni 2007 eine Kooperation zwischen der ZB MED und der Averbis GmbH eingegangen. Zu den Projektaufgaben dieser auf computerlinguistische Suchtechniken spezialisierten Firma gehörte zum einen die Überführung der ZB MED-Datenbanken MEDLINE, CC MED[12] und ZB MED-OPAC[13] in einen gemeinsamen Index auf der Basis moderner Suchmaschinentechnologie (Lucene[14]) sowie die Implementation der Averbis-Suchtechnik bzw. MorphoSaurus-Technologie (vgl. Markó, Schulz & Hahn, 2005) in ein Testsystem. Zum anderen war es die Aufgabe des Projektpartners, die Ergebnisse und das Feedback des bei der ZB MED ange-

[11] http://www.bmbf.de

[12] CC MED (Current Contents Medizin) ist eine Literaturdatenbank der ZB MED, die Artikel ausgewerteter deutschsprachiger oder in Deutschland verlegter Zeitschriften zu medizinischen und gesundheitsrelevanten Themenbereichen nachweist. Es sind insbesondere Zeitschriften enthalten, die nicht in den Datenbanken MEDLINE und Embase ausgewertet werden (Umfang ca. 551.000 Einträge, Stand: 24.11.08).

[13] Der ZB MED-OPAC ist der Online-Katalog der Deutschen Zentralbibliothek für Medizin und umfasst ca. 746.000 bibliografische Angaben. Er bildet den Gesamtbestand der Bibliothek ab und enthält Angaben über Monografien, Zeitschriften und andere Medien (Stand: 24.11.08).

[14] Bei Lucene handelt es sich um eine Open-Source-Java-Bibliothek, die ein sehr performantes und gut skalierbares Suchmaschinen-Framework bereitstellt und von vielen Entwicklern unterstützt wird. Die MorphoSaurus-Technologie arbeitet aber auch mit anderen Frameworks wie z. B. FAST zusammen.

siedelten Evaluationsteams[15] zu verarbeiten und während der Projektlaufzeit kontinuierlich verbesserte Versionen von MEDPILOT-Testsuchumgebungen zu entwickeln.

Zu den Aufgaben des ZB MED-Evaluationsteams gehörte zunächst die Durchführung einer Logfile-Analyse[16] des bisherigen MEDPILOT-Suchsystems. Diese hatte zum Ziel, den Istzustand der medizinischen Literatursuche bzw. den Informationsbedarf der MEDPILOT-User zu beschreiben und als Basis für Verbesserungsvorschläge nutzbar zu machen. In diesem Zusammenhang wurde eine Inhaltsanalyse der von den Nutzern verwendeten Suchterme (Queries) durchgeführt sowie eine Analyse der formalen Aspekte der Suchanfragen (z.B. Länge der Queries, Gebrauch von Booleschen Operatoren usw.).

Zur Überprüfung der Leistungsfähigkeit der Averbis-Suchmaschinentechnologie wurden spezielle Testkollektionen entwickelt, mit denen die verschiedenen Aspekte sprachlich problematischer Suchanfragen überprüft und evaluiert werden konnten. Im weiteren Verlauf der Evaluation wurde schließlich eine repräsentative Testkollektion mit 100 Suchanfragen entwickelt, die speziell für den Vergleich der Leistungsfähigkeit zwischen der Averbis-Suchmaschine und den potenziellen Konkurrenten (wie PubMed, GoPubMed, Google, Google Scholar und Scirus) im Bereich des medizinischen Information-Retrievals eingesetzt wurde.

Einen weiteren Schwerpunkt des Projekts bildete die Untersuchung von geeigneten Maßnahmen zur Verbesserung der Gebrauchstauglichkeit (Usability) des MEDPILOT-Portals. Um empirisch fundierte Empfehlungen für einen zukünftigen MEDPILOT-Relaunch ableiten zu können, wurde daher ein szenariobasierter Usability-Test mit Personen aus der Zielgruppe des Web-Angebots (Wissenschaftler, Ärzte und Studenten) entwickelt und durchgeführt.

Das Buch wendet sich in erster Linie an alle, die über eine webbasierte Suchumgebung den Zugang zur Recherche nach medizinischer Literatur und Information anbieten und jene, die die Leistungsfähigkeit ihres Suchmaschinenangebots verbessern möchten. Darüber hinaus können auch alle Betreiber „Virtueller Fachbibliotheken“ und andere Anbieter von

[15] Zum Evaluationsteam der ZB MED gehörten Dr. Dipl.-Psych. Waldemar Dzeyk, Dipl.-Biol. und Germanistin Anu Maarit Stoor sowie Dipl.-Heilpädagogin Stefanie Paschke. Unterstützt wurde das Team temporär durch die Studentinnen der Bibliothekswissenschaften Natascha Dahmen und Katja Köhl.

[16] Logfiles sind Daten, die auf dem Server gespeichert werden und Informationen über die Nutzung eines Web-Angebots enthalten, indem der Server alle Aktionen auf der Website protokolliert.

Fachinformationen von seinem Inhalt profitieren, da hier neue vielversprechende Ansätze in der wissenschaftlichen Informationssuche aufgezeigt werden. Durch eine entsprechende Anpassung von fachspezifischen Thesauri an andere Inhaltsdomänen lässt sich diese Technik mit Gewinn für eigene Projekte einsetzen. Durch die detaillierte Beschreibung der Evaluationsmethoden erhofft sich der Autor, Anregungen für eine weiterführende wissenschaftliche Diskussion um adäquate Bewertungsmethoden zur Einschätzung der Trefferqualität wissenschaftlicher Suchmaschinen zu geben. Dazu gehört nicht zuletzt auch die Diskussion um die optimale Gestaltung der Gebrauchstauglichkeit eines webbasierten Information-Retrieval-Systems.

Kapitel 2 führt in den *theoretischen Hintergrund* des Forschungsprojekts ein. Zunächst werden die wichtigsten allgemeinen Einflussgrößen für den Erfolg eines webbasierten Retrieval-Systems anhand eines Rahmenmodells vorgestellt (Kap. 2.1 und Kap. 2.2), bevor die Suchmaschine MEDPILOT und die Probleme des medizinischen Information-Retrievals näher beschrieben werden (Kap. 2.3). In Kapitel 2.4 wird der spezielle Ansatz der MorphoSaurus-Technologie erläutert. Kapitel 2.5 beschäftigt sich mit den Erkenntnissen zum Suchverhalten der Suchmaschinennutzer sowie mit den speziellen Aspekten der Usability von webbasierten Suchmaschinenumgebungen.

In Kapitel 3 werden die *Projektziele und die konkreten Fragestellungen* skizziert. In Kapitel 4 werden die *im Projekt eingesetzten Methoden* vorgestellt. Hier geht es insbesondere um die Anwendung von Inhaltsanalysen in Verbindung mit der Auswertung von Logfiles. Des Weiteren werden die angewandten Methoden zur Evaluierung der Trefferqualität sowie die Vorgehensweise bei der Konkurrenzanalyse und der Durchführung der Usability-Untersuchung erörtert.

Die *Ergebnisse* der Retrieval-Tests sowie der Usability-Untersuchung werden in Kapitel 5 vorgestellt. Kapitel 6 enthält die *Diskussion der Ergebnisse* und Kapitel 7 das *Fazit der Untersuchung*. Hier wird auch die Frage aufgegriffen, inwiefern andere Inhaltsdomänen ebenfalls von der MorphoSaurus-Technik und den im Projekt erarbeiteten Forschungsergebnissen profitieren können. Abschließend werden *Desiderata* für mögliche Folgeprojekte formuliert.

2 Theoretischer Hintergrund

Im MorphoSaurus-Projekt ging es primär um die *Optimierung der Systemeigenschaften* der MEDPILOT-Suchmaschine, die *Evaluation der Retrieval-Leistung* nach Verbesserungsmaßnahmen sowie den *Vergleich mit konkurrierenden Suchmaschinen*. Wichtig war aber auch die *Entwicklung einer verbesserten Suchmaschinenschnittstelle mit einem hohen Grad an Benutzerfreundlichkeit*. Für eine theoretische Verortung des MorphoSaurus-Projekts in den Gesamtzusammenhang des medizinischen Information-Retrievals werden in den folgenden Abschnitten die wichtigsten Einflussfaktoren für ein erfolgreiches webbasiertes Information-Retrieval näher beleuchtet und anhand eines Rahmenmodells diskutiert.

2.1 Kennwerte zur Messung des Retrieval-Erfolgs

Die Evaluation von Information-Retrieval-Systemen hat in den Informationswissenschaften eine lange Tradition (vgl. Lewandowski, 2006; 2007a). Zumeist geht es hier um die Frage, *wie viele relevante Treffer* ein Nutzer bei Eingabe einer bestimmten Suchanfrage erhält. Je relevanter die Treffer sind, desto zufriedener sind die Nutzer eines Systems (vgl. Huffman & Hochster (2007). Auf der methodischen Ebene sind hier im Kern zwei Bewertungsmaße zu unterscheiden, die auch für das von uns betrachtete Web-Information-Retrieval zutreffen: der Precision- und der Recall-Wert.

Der Precision-Wert steht für das Verhältnis zwischen gefundenen relevanten Dokumenten und der Gesamtheit der gefundenen Dokumente in einer Recherche und ist somit das Maß für den ‚qualitativen Erfolg' einer Recherche bzw. die Genauigkeit.

$$\text{Relevanz (precision)} = \frac{\text{Zahl der gefundenen relevanten Dokumente}}{\text{Zahl der gefundenen Dokumente}}$$

Im *Web-Retrieval* wird dieser Wert in der Regel bis zu einer bestimmten Grenze bestimmt, da es ökonomisch nicht vertretbar wäre, sämtliche Treffer in Bezug auf ihre Relevanz einzuschätzen. Zumeist wird hier der Cut-Off-Wert von 20 gewählt (d.h., man zählt aus, wie viele unter den ersten 20 Dokumenten relevant für die jeweilige Anfrage sind).

Der Recall-Wert (Trefferquote) gibt Auskunft über das Verhältnis zwischen dem Anteil der relevanten ausgegebenen Treffer (Dokumente) und der Gesamtheit aller vorhanden

relevanten Dokumente einer Datenbank. Er ist somit ein Maß für den ‚quantitativen Erfolg' der Recherche.

$$\text{Vollständigkeit (recall)} = \frac{\text{Zahl der gefundenen relevanten Dokumente}}{\text{Zahl der relevanten Dokumente in der Datenbank}}$$

Daneben existieren noch die Maße „Fallout", definiert als der Anteil der ausgegebenen nicht relevanten Treffer an der Gesamtzahl der nicht relevanten Treffer der Datenbank sowie das Maß der „Generallity". Dabei handelt es sich um den Anteil der relevanten Dokumente im zugrunde liegenden Datenbestand (vgl. Lewandowski, 2006).

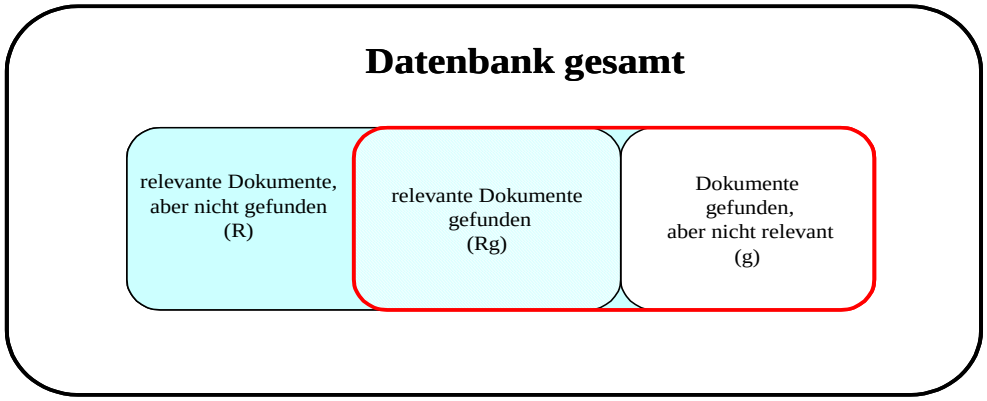

Abbildung 1. Verhältnis von Precision und Relevanz
[Relevanz = Rg / (Rg + g); Vollständigkeit= Rg / (Rg +R)].

Im MorphoSaurus-Projekt ging es primär darum, die Retrieval-Effektivität der Testsuchmaschine anhand der Relevanz bzw. Precision zu bestimmen, da die Ermittlung der anderen Maße mit einem nicht zu vertretenden Aufwand verbunden gewesen wäre. Wünschenswert wäre die Erhebung des Recall-Maßes auf jeden Fall. Dies würde aber voraussetzen, dass man alle für eine bestimmte Anfrage relevanten Treffer kennt. Bei einer Testdatenbasis von ca. 17 Mio. Einträgen war dieser Anspruch verständlicherweise nicht zu erfüllen. Darüber hinaus wurde im Verlauf des Projekts deutlich, dass die Integration einer neuen (semantischen) Technologie in die MEDPILOT-Suchmaschine nicht unter der Aussparung der Nutzer geschehen darf. Deshalb lag der zweite Schwerpunkt der Evaluation auf der Entwicklung und Überprüfung verschiedener Suchmaschinenfunktionen, die dem Rechercheverhalten der Suchmaschinennutzer entgegenkommen und durch eine benutzerfreundliche Gestaltung dazu beitragen, die Akzeptanz der Suchmaschinenoberfläche zu erhöhen.

2.2 Erfolgsfaktoren der webbasierten Literatursuche

Für ein besseres Verständnis des MorphoSaurus-Projekts werden im Folgenden die wichtigsten technischen und theoretischen Rahmenbedingungen erläutert. Die Einbettung des Projekts in den theoretischen Kontext erfolgt anhand eines Rahmenmodells, das die wesentlichen Einflussfaktoren für den Erfolg eines webbasierten Information-Retrieval-Systems zur Literatursuche beschreibt. Dabei erhebt dieses Rahmenmodell nicht den Anspruch auf Allgemeingültigkeit. Dennoch wird damit der Versuch unternommen, einen Ordnungsrahmen zu schaffen, um aus einer logisch-analytischen Perspektive mehr Klarheit in den Zusammenhang der verschiedenen Einflussfaktoren zu bringen.

Nach Jelitto (2007) lässt sich ein Web-Auftritt bzw. dessen Erfolg grundsätzlich nach den Kriterien des Nutzens, der Güte sowie der Zugänglichkeit bewerten. Der *Nutzen*, der allgemein als das wichtigste Bewertungskriterium eines Web-Auftritts gilt, bemisst sich daran „…inwieweit er einerseits die Ansprüche der Auftraggebenden und Mitwirkenden und andererseits die Ansprüche aller definierten Zielgruppen erfüllt“ (Jelitto, 2007, S. 31). Die *Güte* leitet sich ab von der technischen und inhaltlichen Qualität des Web-Auftritts und hängt eng mit der Gebrauchstauglichkeit (bzw. Usability) des Angebots zusammen. Die Zugänglichkeit eines Web-Angebots sagt etwas darüber aus, inwiefern es auch von Personen mit körperlichen Beeinträchtigungen ohne Probleme genutzt werden kann.

Das hier vorgeschlagene Modell nimmt die von Jelitto postulierten Kriterien zur Evaluation von Web-Auftritten auf und ergänzt diese durch die Aufnahme des Nutzerverhaltens als weitere wichtige Einflussgröße. Während die Usability als eine produktseitige Eigenschaft anzusehen ist, ist das Nutzerverhalten jedoch eine rezipientenseitige Dimension, deren Verständnis ganz wesentlich zum Erfolg eines Web- bzw. Suchdienstangebots beiträgt. Die Bewertungskriterien von Jelitto werden im vorgeschlagenen Modell stärker differenziert und teilweise unterschiedlich verortet (vgl. Abbildung 2).

Abbildung 2. Erfolgsfaktoren der webbasierten Literatursuche.

Ein weiterer Vorschlag zur Bestimmung der wichtigsten Qualitätsfaktoren – speziell bei der Bewertung von Suchmaschinenangeboten – stammt von Lewandowski (2006; 2007a) sowie Lewandowski und Hochstötter (2008). Nach Ansicht der Autoren bestimmt sich die Qualität von Web-Retrieval-Systemen über vier grundlegende Bewertungsdimensionen:

- **Qualität des Index**
  - – Größe des Datenbestands, Abdeckung des Web (bzw. der Domäne)
  - – Abdeckung bestimmter Bereiche (Sprachräume, Länder)
  - – Überschneidungen der Indices
  - – Aktualität des Datenbestands
- **Qualität der Suchresultate**
  - – Retrieval-Effektivität
  - – Zufriedenheit der Nutzer
  - – Überschneidungen der (Top-)Ergebnisse
- **Qualität der Suchfunktionen**
- **Nutzerfreundlichkeit und Benutzerführung**

Für die Messung der Retrieval-Effektivität schlägt Lewandowski (2007a) neben der Ermittlung des Precision-Wertes auch andere Maßzahlen vor, die sich aber vorerst noch in einem eher experimentellen Stadium ihrer Tauglichkeitsprüfung befinden. Daher wird hier nicht weiter auf sie eingegangen. Wie Jelitto verzichten Lewandowski und Hochstötter auf die Trennung von medien- und rezipientenseitigen Einflussgrößen und Prozessen. Die Faktoren Nutzerfreundlichkeit und Benutzerführung werden von den Autoren zwar explizit erwähnt, doch wird das *Nutzerverhalten* als solches auch hier nicht als eigener Erfolgs- bzw. Qualitätsfaktor betrachtet. Die hier in Anschlag gebrachten Bewertungsdimensionen sind im Gegensatz zu den Evaluationskriterien von Jelitto speziell für die Einordnung der Qualität eines webbasierten Suchsystems aufgestellt worden und besitzen deshalb eine größere Relevanz für das MorphoSaurus-Projekt.

Bisher hat sich die wissenschaftliche Auseinandersetzung zur Identifikation von Erfolgskriterien bei der Informationssuche im Netz vorwiegend auf informationswissenschaftliche Indikatoren bzw. Maßzahlen wie Precision oder Recall konzentriert (vgl. z. B. Fourie, 2006). Für eine umfassende wissenschaftliche Evaluation der Qualität und des Potenzials web-

basierter Retrieval-Systeme sind nach Meinung des Autors sämtliche der in Abbildung 2 genannten Faktoren zu berücksichtigen. Die vorliegenden Theorieansätze machen deutlich, dass eine valide Bewertung von Suchmaschinen nach einem *multimethodalen* Ansatz verlangt. Das bedeutet, dass aufgrund der Komplexität der Fragestellung mehrere unterschiedliche Methoden zum Einsatz kommen müssen, um der Vielfalt der zu erfassenden Phänomene gerecht zu werden. Zum Kern dieser Überzeugung gehört, dass neben den Systemeigenschaften stets auch die rezipientenseitigen Einflussgrößen wie das Nutzerverhalten mit zu berücksichtigen sind (Fourie, 2006).

Andere Forscher mögen hier noch weitere Faktoren geltend machen: So spielt auch der Bekanntheitsgrad eines Web-Auftritts bzw. eines Suchdienstes eine große Rolle für den Erfolg eines Web-Angebots. Die vorliegende Arbeit beschränkt sich jedoch auf die – nach Ansicht des Autors – wichtigsten Dimensionen und Einflussgrößen. Die Ausführungen gelten im Prinzip für alle webbasierten Suchsysteme, doch wird hier speziell immer wieder auf die Eigenheiten von medizinischen Suchsystemen Bezug genommen.

Für die Nutzer eines webbasierten Information-Retrieval-Systems bestehen die Herausforderungen bei der Literatur- bzw. Informationsrecherche vor allem im *Finden*, *Auswählen* und *Bewerten* der für sie relevanten Informationen. Dabei kommen im Wesentlichen vier Einflussfaktoren zum Tragen, die den Erfolg eines solchen Suchsystems begründen:

- die Systemeigenschaften der Suchmaschine (technische Qualität),
- der Inhalt (Umfang, Qualität, Aktualität),
- das Nutzerverhalten und die Nutzererwartungen,
- der Nutzen bzw. die Nützlichkeit.

Die *Systemeigenschaften* des Suchsystems definieren die technische Qualität und können zusammen mit den angebotenen *Inhalten* als *medienseitige bzw. produktseitige Einflussfaktoren* bezeichnet werden. Das *Nutzerverhalten und die Nutzererwartungen* sowie der *erlebte Nutzen bzw. die Nützlichkeit* des Web-Angebots können als *rezipientenseitige bzw. nutzerseitige Einflussfaktoren* für den Erfolg gelten. Wegen der großen Bedeutung dieser Einflussgrößen werden diese in den folgenden Abschnitten ausführlicher erläutert.

2.2.1 Die Systemeigenschaften der Suchmaschine

Als wichtige Systemeigenschaften, die die *technische Qualität* eines webbasierten Such-Systems für die elektronische Literaturrecherche bestimmen, lassen sich drei wichtige Faktoren identifizieren:

a. die eingesetzte *Suchtechnik*,
b. die Frage der *Accessibility* (Zugänglichkeit)
c. die *Usability* des Systems (Benutzerfreundlichkeit)

Zu a.: Eingesetzte Suchtechnik. Die Qualität der Informationsverarbeitung eines Suchsystems hängt wesentlich von der eingesetzten *Suchmaschinentechnologie* ab. Wie bereits weiter oben ausgeführt, bestimmt die Technik des eingesetzten Retrieval-Systems, wie schnell und wie gut die Suchanfragen der Nutzer analysiert und relevante Treffer zurückgemeldet werden (technische Qualität). Hier lassen sich grundsätzlich folgende Technologien bzw. Vorgehensweisen unterscheiden:

- die Katalogsuche (Sammlung redaktionell betreuter Inhalte / Links)
- die Metasuche (Zusammenführen verschiedener Quellen)
- die algorithmische Suche (ein gemeinsamer Index vorhanden, Relevanz-Ranking möglich, ebenso Einbezug von Thesauri oder Textmining-Methoden)
- die soziale Suche (Beispiel ‚social bookmarks', ‚tagging' von Treffern durch die User)
- die semantische Suche (Rückgriff auf Ontologien und Konzept-Relationen, Kombination mit algorithmischer Suchtechnologie möglich)

Um Anfragen der Nutzer zufriedenstellend beantworten zu können, werden die eingegebenen Suchterme durch die Algorithmen der Suchmaschine analysiert. Die obigen Ansätze unterscheiden sich dabei wesentlich in der Art der Verarbeitung der Nutzeranfragen. Dabei können sowohl formale als auch semantische Merkmale durch die Systemalgorithmen berücksichtigt werden. Je besser eine Suchmaschine mit sprachlich problematischen Phänomenen wie grammatikalischen Variationen, Synonymen, Homonymen, Komposita, Akronymen oder fremdsprachlichen Inhalten zurechtkommt, desto wahrscheinlicher werden auch relevante Treffer gefunden. Auf die Vor- und Nachteile der einzelnen Suchmaschinenansätze soll an dieser Stelle nicht weiter eingegangen werden. Hier wird auf die weiterführende Literatur verwiesen, etwa Maaß et al., 2009 oder Grießbaum et al., 2009. Für das MorphoSaurus-

Projekt wurde eine Kombination aus algorithmischer Suche und der von Averbis entwickelten Sprachverarbeitungstechnologie gewählt. In Kapitel 2.4 wird dieser neue Technologieansatz detailliert vorgestellt.

Zu b.: Accessibility. Ein weiterer Erfolgsfaktor eines Web-Auftritts ist die *Zugänglichkeit* oder ‚accessibility' des Systems. Dieser Begriff beschreibt die Gesamtheit bestehender *technischer Zugangshürden* zu einem Web-Auftritt.

Technische Hürden. Ein wichtiges Stichwort in diesem Zusammenhang ist der Grad der *Barrierefreiheit*. Damit wird die Fähigkeit eines Systems beschrieben, Nutzern auch bei unterschiedlichen Graden von vorhandenen körperlichen Einschränkungen (wie z. B. Sehbehinderungen) Zugang zum System und seinen Funktionen zu geben. Hierbei geht es eher um die Umsetzung von technischen Aspekten wie sie z. B. in den verschiedenen Richtlinien und Normen zur Barrierefreiheit beschrieben werden (z. B. BITV[17]). Zur Barrierefreiheit gehört aber auch die Unabhängigkeit von einer bestimmten technischen Plattform (Desktop-PC, Handy etc.), ebenso wie die Unabhängigkeit des Web-Angebots von einem bestimmten Betriebssystem oder einer speziellen (Browser-)Software. Eine gute Accessibility ist auch deshalb wichtig, weil Internetsuchmaschinen bzw. Crawler durch die technischen Zugangsbarrieren in ihrer Such- und Sammeltätigkeit gefördert oder behindert werden können. Ist ein Web-Angebot den Suchmaschinen nicht oder nur schlecht zugänglich, schmälert dies die Sichtbarkeit und letztlich den Erfolg einer Website.

Bei einem weitergefassten Barrierebegriff können Zugangshürden zum System aber auch finanzieller, rechtlicher oder inhaltlicher Art sein.

Finanzielle Hürden. Die Benutzung einiger Datenbanken im Bereich der Medizin ist zwar kostenlos (wie z. B. MEDLINE), für die Nutzung anderer Quellen fallen hingegen Gebühren an, deren Höhe sich an dem rechtlichen Status einer bestimmten Nutzergruppe orientiert.

Rechtliche Hürden. Abgesehen von den finanziellen Hürden ist der Zugang zu manchen Datenbanken auf einen bestimmten Nutzerkreis beschränkt (wie z. B. Mitglieder einer Universität, Angehörige eines Landes oder Mitglieder eines ausgewählten IP-Bereichs).

[17] Die Barrierefreie Informationstechnik-Verordnung (BITV) ist eine Ergänzung des Behindertengleichstellungsgesetzes (BGG) vom 27. April 2002.

Inhaltliche Hürden. In der Regel wird wenig beachtet, dass der Erfolg einer Website auch durch inhaltliche Zugangshürden beeinflusst wird. Durch das gewählte Textniveau (formal und inhaltlich) fällt die Nutzung des Web-Angebots bestimmten Zielgruppen leichter (z. B. Ärzten, Forschern), anderen dagegen schwerer (z. B. Laien, Fachnovizen etc.). Durch ein bestimmtes sprachliches Niveau der Darstellung kann der Anbieter die Ausrichtung des Angebots auf eine bestimmte Zielgruppe steuern (vgl. z. B. Hellbusch & Mayer, 2006 oder Weist, 2004).

Spezialsuchmaschinen richten sich häufig an Experten, die über die gebräuchliche Fachterminologie verfügen. Die Nutzung einer solchen Suchmaschine durch Laien und Novizen des Fachgebiets kann durch deren mangelndes terminologisches Fachwissen zu unbefriedigenden Suchergebnissen führen. Dies gilt es bei der Entwicklung eines Web-Angebots mit zu berücksichtigen (vgl. Kap. 2.5.2).

Zu c.: Usability. Hierbei geht es um eine optimierte Informationsarchitektur. Das heißt, der User wird – über die grundlegende Funktionalität einer Suchmaschine hinaus – in der Navigation, im Textverständnis und in den Rückmeldeprozessen so unterstützt, dass er die Suchmaschine als effektives und effizientes Werkzeug zur Beschaffung von Informationen erlebt. Darüber hinaus muss die Interfaceoberfläche so gestaltet sein, dass die User neben der Akzeptanz auch Freude an der Benutzung entwickeln können. Für den Erfolg bei der Informationssuche spielt dabei eine große Rolle, ob das System leicht zu erlernen und zu bedienen ist. Mit diesem Punkt ist der Bereich der *Benutzerfreundlichkeit* bzw. *Usability* angesprochen. Die *Benutzbarkeit* eines Systems hat einen entscheidenden Einfluss auf die Akzeptanz durch die Nutzer (vgl. Kap. 2.5.2 u. Kap. 2.5.3). Neben der Orientierung und Navigation spielt sie insbesondere bei der Zielerreichung eine wichtige Rolle. Ein barrierearmes System kann dabei die Benutzerfreundlichkeit wesentlich erhöhen (s.o.). Der Begriff der Usability definiert sich *im Kern* aus den drei folgenden Aspekten (DIN ISO 9241-11):

- **Effektivität**. Sie gibt darüber Auskunft, wie gut bzw. vollständig eine Aufgabe von dem Benutzer ausgeführt werden kann.
- **Effizienz**. Diese Größe beschreibt den Wirkungsgrad bzw. den Aufwand, den ein Benutzer einsetzen muss, um eine Aufgabe zu lösen.
- **Akzeptanz**. Hierbei handelt es sich eine subjektive Größe, um die Zufriedenheit der User im Umgang mit einem System zu beschreiben.

Darüber hinaus zählt die *Selbstbeschreibungsfähigkeit* eines Systems zu den wesentlichen Faktoren der Benutzerfreundlichkeit. Dies berührt die Frage, inwieweit die Funktionselemente einer Website oder Suchmaschine ohne weiterführende Erklärungen nachvollziehbar und verständlich sind, sodass die Nutzer das System ohne Umwege bedienen können. Wenn die Selbstbeschreibungsfähigkeit sich auf einem niedrigen Niveau befindet, kann ein System weder besonders effektiv noch effizient sein. Zudem wird auch die Akzeptanz des Gesamtsystems negativ beeinflusst.

Sehr wichtig in diesem Zusammenhang sind die Aufbereitung der Trefferliste sowie gegebenenfalls erklärende Rückmeldungen bei Null-Treffer-Suchen. Eine detaillierte Betrachtung von Usability-Problemen – speziell von Suchmaschinen – erfolgt in Kapitel 2.5.3. All diese vorgenannten Eigenschaften können als *Systemeigenschaften* einer Suchmaschine bezeichnet werden, da sie mehr oder weniger von technischen Parametern bestimmt werden.

2.2.2 Der Inhalt

Auf der Medienseite ist neben den Systemeigenschaften *der Inhalt des Angebots* der zweite große Erfolgsfaktor eines Web-Retrieval-Angebots. Hier ist entscheidend, *welche Art von Inhalten* (‚scope') angeboten wird und ob diese Inhalte die fragliche Domäne möglichst vollständig *abdecken* (‚coverage'). Zudem spielt es eine Rolle, wie *aktuell* die Informationen sind (‚timeliness'), die angeboten werden.

Der Inhalt des Web-Angebots zählt zu den wichtigsten Bewertungsdimensionen aus Sicht der Nutzer. Während es für die durchschnittlichen Nutzer eher schwierig ist, den Inhalt eines Angebots zu bewerten, können sich Experten oder fortgeschrittene Novizen eines Fachgebiets sehr viel eher einen schnellen Überblick verschaffen. In der Regel tun sie dies anhand:

a. der *Qualität*,
b. des *Umfangs* und
c. der *Aktualität* der angebotenen Informationen.

Zu a.: Qualität. Der Begriff der Qualität ist zunächst ein subjektives Konzept. Im Zusammenhang mit dem Angebot einer Spezialsuchmaschine bemisst sich die Qualität zum einen am Niveau der abrufbaren Informationen und zum anderen hängt sie ab von der subjektiven Einschätzung der Nutzer. Um eine genügend hohe Akzeptanz durch die anzusprechende Zielgruppe (z. B. Wissenschaftler und Studenten) zu erreichen, sollte das Datenbank-Angebot

alle wesentlichen Quellen abdecken, die von der entsprechenden Fachcommunity für wichtig und unerlässlich gehalten werden (z. B. Zeitschriften mit hohem Impactfaktor). Das Einverständnis darüber, welche Quellen für eine Domäne als wichtig angesehen werden, ist sicherlich dem zeitlichen Wandel unterlegen und dementsprechend regelmäßig anhand der Nutzerwünsche zu überprüfen. Aus Sicht der Nutzer ist es wichtig, dass die Trefferrückmeldungen nach Suchanfragen möglichst viele relevante Treffer unter den ersten 10 bis 20 Treffern aufweisen. Eine Suchmaschine, die zwar viele Quellen von hohem Niveau in ihrem Datenbestand hat, aber keine geeignete Technologie einsetzt, um daraus bei Suchanfragen eine hochwertige Trefferliste zu erzeugen, verfehlt ihr Ziel und verschenkt ihr Potenzial. Deshalb hängt die *wahrgenommene Qualität* von Suchanfragen-Ergebnissen auch von der Qualität der Algorithmen zur Analyse der Suchterme und ihrer Fähigkeit zur Bildung eines Relevanzrankings ab.

Zu b.: Quantität (Umfang) bzw. Abdeckung. Erst wenn der Datenbestand einer Suchmaschine eine gewisse kritische Masse überschreitet, ist davon auszugehen, dass auch Nutzer mit unterschiedlichen Informationsbedürfnissen zufrieden gestellt werden können. Letztendlich ist es häufig nicht so sehr die Quantität, die den Erfolg bestimmt, sondern vielmehr die Qualität der Daten und Quellen. Allerdings zeigt das Beispiel Google, dass mit der Masse auch die Wahrscheinlichkeit steigt, passende Treffer zu finden.

Zu c.: Aktualität. Gerade in den Lebenswissenschaften kommt es häufig auf den aktuellen wissenschaftlichen Kenntnisstand an. Kann eine Suchmaschine hier keine zeitnahen Aktualisierungen bieten, wird sie schnell unattraktiv für Fachwissenschaftler. Andere Suchmaschinen, wie z. B. im Bereich der historischen Wissenschaften müssen mit ihren Quellen auch weit zurückliegende Zeiträume gut abdecken können, um als kompetent wahrgenommen zu werden.

2.2.3 Das Nutzerverhalten und die Nutzererwartungen

Auf der Nutzerseite sind es vor allem die Aspekte der Mensch-Computer-Interaktion (HCI), die darüber entscheiden, ob eine Suchmaschine von den Usern angenommen wird oder nicht.

Der Nutzer reagiert nicht einfach nur auf das Angebot, das ihm über das Web-Interface präsentiert wird, sondern er *interagiert* mit dem System. Der Nutzer bedient eine grafische Oberfläche mit technischen Merkmalen. Diese muss er zunächst verstehen. Dabei bietet ihm

das System mehr oder weniger gute Hilfestellungen an. Je besser die Benutzerfreundlichkeit eines Systems gestaltet wurde, desto eher gelingt es dem Nutzer seine Ziele zu erreichen. Wichtige Einflussvariablen, die das Nutzerverhalten beschreiben, lassen sich zwei Dimensionen zuordnen:

a. der *Medienkompetenz* der Nutzer sowie
b. dem *habitualisierten Nutzungsverhalten* gegenüber bestimmten Internet-Angeboten oder -Genres.

Zu a. Medienkompetenz. Sie gilt als Schlüsselqualifikation in der Informationsgesellschaft (Glotz, 2001). Nach Groeben (2004) handelt es sich um eine kritische Analysefähigkeit, die als praktische Anwendung des jeweiligen medienspezifischen Strukturwissens aufgefasst werden kann. Menschen müssen zur adäquaten Bewertung von Informationsangeboten auf ihr medienspezifisches Wissen über die inhaltlichen und formalen Aspekte des Mediums zurückgreifen (vgl. Schreier & Appel, 2002). Neben *strategischem Wissen*, wie z.B. die Anwendung effizienter Suchstrategien (vgl. Navarro-Prieto, Scaife & Rogers, 1999), ist für eine kompetente Nutzung auch ein *Metawissen* über Qualitätskriterien zur Beurteilung von Informationen erforderlich.

Für die Literaturrecherche im Internet bedeutet *strategisches Wissen* vor allem, dass die Nutzer sich im Ausmaß ihrer *Interneterfahrung* bzw. *Informationskompetenz* (‚information literacy') unterscheiden, d.h. im Grad von Routine- und Handlungswissen bei der Auswahl und Bedienung von Suchmaschinen und Datenbanken sowie im Ausmaß ihrer Rechercheкompetenz. Damit ist gemeint, dass die Nutzer z.B. über die Fähigkeit verfügen müssen, eine Suche mit Synonymen durchzuführen oder die Suchanfrage angemessen zu reformulieren oder spezielle Suchfunktionen wie Boolesche Operatoren oder Feldsuchen zu gebrauchen. Für die Zufriedenheit mit einem System spielt es zudem eine große Rolle, vor welchem Hintergrundwissen bzw. Domänenwissen die Ergebnisse von Suchanfragen bewertet werden. So unterscheiden sich Laien, Studenten und Wissenschaftler im Ausmaß ihrer Medizinkenntnisse. Mit *Metawissen* ist eine *spezifische Bewertungskompetenz* gemeint. Dabei handelt es sich um ein übergeordnetes Wissen über Zusammenhänge, Strukturen, Standards und Normen, welches ergänzt wird durch prozesshafte Variablen wie die Fähigkeit zu kritischem Denken (‚critical thinking'). Im Bereich der Bewertung medizinischer Informationen steht hier die Fähigkeit im Vordergrund, Inhalte und Methoden zu verstehen und Ergebnisse von

Forschungsarbeiten kritisch bewerten und bezüglich ihrer Qualität einordnen zu können. Das Ausmaß der Bewertungskompetenz hängt dabei eng mit dem Umfang des medizinischen Wissens zusammen. Je größer das medizinische Wissen ist, über das ein Nutzer verfügt, desto differenzierter kann er die Qualität von Datenbankinhalten einschätzen.

Eine etwas anders akzentuierte Fassung des Konstrukts der Medienkompetenz im Zusammenhang mit der Nutzung von Computern und Internetdiensten bietet der Begriff der „Digital Literacy" (vgl. Gilster, 1997). Nach Eshet-Alkali und Amichai-Hamburger (2004) sind es fünf Hauptfaktoren bzw. Fertigkeiten, die digitale Literarizität konstituieren: 1. Fotovisuelle Fertigkeiten (z. B. das „Lesen" von grafischen Displays), 2. Reproduktive Fertigkeiten (Herstellung von neuen und sinnvollen Dingen unter Benutzung vorhandener digitaler Vorlagen), 3. Fertigkeiten zur Entnahme von Wissen aus hypertextuellen nonlinearen Strukturen, 4. Informationelle Fertigkeiten (Bewertung der Qualität und Stichhaltigkeit von Informationen) und 5. Sozio-emotionale Fertigkeiten (das Verstehen und Beherrschen der Regeln in der computervermittelten Kommunikation).

Zu b. Habitualisiertes Nutzungsverhalten. Die Nutzung von Medien besitzt häufig einen habitualisierten Charakter. Das bedeutet, dass sich im Laufe der Zeit bestimmte Nutzungsgewohnheiten herausbilden, die den Umgang mit dem Medium mitbestimmen. Auch wenn es sich im Falle des Internets um ein noch recht junges Medium handelt, lassen sich bestimmte Nutzungsgewohnheiten beobachten. Ein Beispiel dafür ist der Umgang mit Suchmaschinen. Für die Deutschen ist die Suche im Internet fast synonym mit der Nutzung von Google. Im Oktober 2009 vertrauten ca. 90% der deutschen Nutzer bei der Internetsuche auf Google (webhits.de, Oktober 2009). Haben sich die Nutzer erst an gewisse Formen der Interfacegestaltung und Benutzerführung gewöhnt, erwarten sie diese auch bei anderen ähnlichen Web-Angeboten. Durch die Gewöhnung an Google war lange Zeit der Trend zu beobachten, dass andere Suchmaschinenanbieter ihre Oberfläche ebenfalls sehr schlicht und einfach gestalteten: Dem User werden auf der Einstiegsseite oft nur sehr wenige Optionen zur Einschränkung der Suchergebnisse angeboten, so wie sie es von Google gewohnt sind. Inzwischen rückt aber sogar Google von dieser Art Purismus ab, um den Usern das Auffinden besserer bzw. relevanterer Treffer zu ermöglichen. Da Nutzerverhalten und Nutzererwartungen eine so große Rolle für die Akzeptanz von Suchmaschinenangeboten spielen, werden die hierzu vorliegenden wissenschaftlichen Erkenntnisse in Kap. 2.5 im Überblick referiert.

2.2.4 Der erlebte Nutzen

Der erlebte *Nutzen* eines Web-Auftritts ist das Ergebnis eines komplexen Prozesses, der auf dem Zusammenspiel verschiedener Einflussgrößen beruht. Er gehört zu den wichtigsten Faktoren, die User dazu bewegen, immer wieder zu einem Web-Angebot zurückzukehren. Andere Gründe sind z. B. die *Vertrauenswürdigkeit* und *Kompetenz des Anbieters* oder auch schlicht die *Freude*, die die Benutzung des Angebots auslöst. Dieser „Joy of Use" ist als motivierender Faktor nicht zu unterschätzen, spielt aber bei informationsorientierten Angeboten nicht die wichtigste Rolle.

Der erlebte Nutzen hängt unmittelbar davon ab, ob die Nutzer ihre *Ziele* auf der Website erreichen. Werden die User in ihren Erwartungen hinsichtlich des Nutzens bestätigt, steigt die Chance für die Akzeptanz des Web-Angebots. Das Vertrauen in das System nimmt zu und gleichzeitig festigt sich das Image des Anbieters als kompetente Informationsquelle. Der Nutzen, den die User durch den Umgang mit dem System erfahren, führt schließlich dazu, dass ein Web-Angebot erneut aufgesucht wird. Ist ein System aber z. B. nicht barrierearm und benutzerfreundlich, können die User ihre Ziele nicht oder nur sehr umständlich erreichen. Die Folge ist mangelnde Akzeptanz durch die User sowie ein als gering eingestufter Nutzen. Darüber hinaus spielt auch die Zufriedenheit mit den Trefferrückmeldungen eine sehr wichtige Rolle. Das bedeutet: Je mehr relevante Treffer sich unter den ersten von der Suchmaschine zurückgemeldeten Treffern befinden, desto eher werden die User das System als nützlich erleben.

Der Anbieter eines Web-Retrieval-Systems kann anhand unterschiedlicher Parameter feststellen, ob die User sein System annehmen. Das Controlling kann hier zum einen über Logfile-Analysen erfolgen, wobei hier zu den wichtigsten Kennzahlen die Veränderung der Klickraten zählt (vgl. Heindl, 2003; Jansen, 2006). Aber allein auf die Logfiledaten sollte man sich nicht verlassen, da sie nur indirekt Maßzahlen zur Bewertung des Angebots durch die Nutzer liefern (vgl. Grimes, Tang & Russell, 2007). Dennoch lohnt es sich mittels des Logfiles folgende Fragen kontinuierlich zu verfolgen:

- Welche ist die beliebteste Einstiegsseite (bzw. ‚landing page') und welche ist die häufigste Seite, von der aus das Angebot wieder verlassen wird?
- Welche Inhalte sind besonders beliebt und welche werden kaum nachgefragt?
- Wie lange verweilen die User auf den Seiten des Angebots?

- Wie viele Artikel/Datenbankeinträge werden pro Besuch angesehen bzw. abgerufen oder heruntergeladen?
- Wie hoch ist die Konversionsrate? Darunter versteht man den Quotienten aus der Besucherzahl und der Anzahl derjenigen, die tatsächlich etwas bestellt, gekauft oder heruntergeladen haben. Allgemeiner ausgedrückt: Wie viele der Besucher einer Website führen eine gewünschte Aktion aus?

Daneben bieten Nutzerumfragen und Usability-Tests mit Personen aus der Zielgruppe die Möglichkeit, den Grad der Zufriedenheit mit dem Web-Angebot oder der Trefferliste zu erfassen (vgl. Huffman & Hochster, 2007). Zudem lassen sich durch diese Erhebungsmethoden Veränderungswünsche der User feststellen.

Neben den eher inhaltlichen Erfolgsfaktoren auf der Medien- und der Rezipientenseite lassen sich natürlich noch weitere Einflussfaktoren identifizieren, wie z.B. die Sichtbarkeit bzw. die Bekanntheit (des Web-Angebots/der Suchmaschine) als Voraussetzung für dessen Nutzung. In den großen Suchmaschinen im Ranking weit oben gefunden zu werden, ist heutzutage essenziell für die erfolgreiche Weiterentwicklung eines Web-Angebots. Inzwischen beschäftigt sich ein ganzer Industriezweig damit, Websites so zu optimieren, dass diese bei Eingabe bestimmter Suchbegriffe möglichst weit oben in den Trefferlisten der großen Suchmaschinen erscheinen. Unter dem Stichwort „Suchmaschinenoptimierung“ („**S**earch **E**ngine **O**ptimization“ bzw. „SEO“) wird hier auf die entsprechende Literatur verwiesen (z.B. Erlhofer, 2008; Fischer, 2006).

2.3 MEDPILOT

Mit MEDPILOT[18] besitzt die ZB MED eine der profiliertesten Suchmaschinen für den Bereich der medizinischen Literaturrecherche. Unter www.medpilot.de bietet sie ihren Nutzern ein frei zugängliches Fachportal mit Zugriff auf eine Vielzahl kostenloser medizinischer Datenbanken, Kataloge und Informationsquellen. Tabelle 1 gibt einen Überblick über die Quellen und Datenbanken, die mit MEDPILOT durchsucht werden können.

[18] MEDPILOT basiert auf einer Kooperation der Deutschen Zentralbibliothek für Medizin (ZB MED) mit dem Deutschen Institut für Medizinische Dokumentation und Information (DIMDI).

Tabelle 1. Das Datenbank-Angebot in MEDPILOT (Stand: Mai, 2009).

Datenbanken in MEDPILOT	
• **Fachübergreifende Datenbanken**	• **Bestandskataloge**
- MEDLINE	- ZB MED Medizin (OPAC)
- AWMF-Leitlinien	- NLM (National Library of Medicine)
- CC MED (deutsche u. in Deutschland erschienene Zeitschriftenartikel)	- Deutsche Zahnärztebibliothek
- Cochrane Database of Systematic Reviews (CDSR)	- ZB MED Ernährung, Umwelt, Agrarwissenschaften (GREENPILOT)
- Cochrane Database of Abstracts of Reviews of Effectiveness (DARE)	- Elektronische Zeitschriftenbibliothek Regensburg
- Deutsches Ärzteblatt	- Lehmanns Online Bookshop
- Hogrefe Verlag	• **Web- und Multimedia-Datenbanken**
- Karger-Verlagsdatenbank	- Link-Datenbank der ZB MED
- Kluwer-Verlagsdatenbank	- Virtuelle Videothek für die Medizin (VVFM)
- Krause und Pachernegg Publikations-Datenbank	• **13 gebührenpflichtige Datenbanken**
- Springer-Verlagsdatenbank	- Excerpta Medica Database (EMBASE)
- Springer Pre-Print	- PSYNDEX
- Thieme-Verlagsdatenbank	- Allied and Complementary Medicine (AMED)
• **Spezialdatenbanken**	- BIOSIS Previews
- Animal Testing Alternative Methods (AnimAlt-Zebet)	- CAB Abstracts
- Dokumentations- und Informationssystem Qualitätssicherung (DIQ) - Literatur-Datenbank	- EMBASE Alert
- Ethik in der Medizin (ETHMED)	- International Pharmaceutical Abstracts (IPA)
- Datenbank klinischer Studien aus Hämato-Onkologie	- ISTPB + ISTP/ISSHP
- Health Care Literature Information Network (HECLINET)	- PsycINFO
- Sozialmedizin (SOMED)	- SCISEARCH
- XTOXLINE (Toxilogie u. Pharmakologie)	- SOCIAL SCISEARCH
- Health Technology Assessment (HTA) Database	- Cochrane Central Register of Controlled Trials
	- NHS Economic Evaluation Database (NHSEED)

Als virtuelle Fachbibliothek wendet sich MEDPILOT primär an Forscher, Ärzte, Medizinstudenten und andere Berufsgruppen des Gesundheitswesens (vgl. El-Menouar, 2002, 2004). Eine medizinische Suchmaschine ist besonders dann attraktiv für Ärzte, wenn die Daten *kontinuierlich aktualisiert* werden und die Mediziner erkennen können, aus *welchen Quellen* die Informationen stammen (vgl. Schneider, 2004; Reng, Friedrich, Timmer & Schölmerich, 2003). Zusätzliche Kriterien für die Beurteilung von medizinischen Informationsquellen sind vor allem die *Wissenschaftlichkeit*, die *schnelle Verfügbarkeit* sowie die *Benutzerfreundlichkeit* (vgl. Butzlaff, Telzerow, Lange & Krüger, 2001). Heinold und Spiller (2007) haben im Rahmen des vascoda-Projekts[19] die Nutzer „Virtueller Fachbibliotheken“ (ViFas) zu ihren

[19] Bei vascoda handelt es sich um den Zusammenschluss von ca. 40 Partnern, die unter vascoda.de den Zugang zu verschiedenen Fachinformationen über ein Suchmaschinen-Interface bieten. Es ist der bisher umfassendste Ansatz zur Vereinigung Virtueller Fachbibliotheken unter einer Adresse in Deutschland.

Erwartungen gegenüber fachspezifischen Informationsangeboten befragt und einen Überblick über die Wünsche von Wissenschaftlern als Nutzer dieser Informationsplattformen zusammengestellt (vgl. Tabelle 2).

Tabelle 2. Erwartungen der Nutzer Virtueller Fachbibliotheken (aus Heinold & Spiller, 2007).

Bedarf	Begründung
Komplette bibliographische Angaben inkl. Abstract	Ohne diese Angabe ist kaum eine sinnvolle Entscheidung darüber möglich, ob ein Treffer relevant ist
Vollständigkeit / Abdeckungsrate	Neben einer fachspezifischen Suche sollte auch interdisziplinär gesucht werden können
Umfangreiche Suchoptionen	Die Nutzer benötigen eine einfache und eine erweiterte Suchfunktion
Umfassende Funktionalität der Trefferliste	Die Trefferliste sollte sortierbar und eingrenzbar sein, sowie in verschiedene Formate exportierbar
Effizienz der Suche	Wichtig sind schnelle Ladezeiten und Erhalt von Suchbegriffen
Kostenlos zugängliche Inhalte	Für die Wissenschaftler ist es wichtig, innerhalb eines Systems zu arbeiten, das ihnen Volltexte liefert
Volltextsuche	Nicht nur Abstract und Titel, sondern der gesamte Inhalt sollte durchsuchbar sein
Ohne Registrierung zugänglich	Registrierungen werden allgemein abgelehnt, zumindest das Abstract muss ohne zugänglich sein
Nutzerführung / Navigation	Die Suche sollte intuitiv bedienbar sein, Datenbanken und Operatoren sollten erläutert werden
Layout / optische Gestaltung	Die Seiten sollten klar strukturiert sein, die Suchfunktionalität sollte im Vordergrund stehen

Die Aufzählung in Tabelle 2 kann auch als Übersicht über den *erwarteten Nutzen* eines webbasierten Literatursuchsystems verstanden werden (vgl. auch Kap. 2.2, Abbildung 2). Die Ergebnisse dieser Befragung zeigen deutlich, wo die Schwerpunkte eines wissenschaftsorientierten Information-Retrieval-Systems aus Usersicht liegen sollten.

2.3.1 Technologie der bisherigen MEDPILOT-Suche

Der technische Hintergrund im Bereich Suche bestand bei MEDPILOT bis Ende 2008 ausschließlich aus *der Zusammenführung medizinischer Quellen* mithilfe einer *Metasuche*. Dabei können mit einer Suchanfrage eine große Anzahl kostenloser und auch kostenpflichtiger Datenbanken recherchiert werden (vgl. Schneider, 2004). Anschließend werden die Suchergebnisse in einer Treffer-Ergebnisseite zusammengeführt. Bei einem solchen System hängt die Zeit bis zur vollständigen Ergebnisdarstellung u.a. davon ab, wie viele Quellen durchsucht werden müssen und ob diese zum Zeitpunkt der Anfrage auch problemlos

erreichbar sind. Deshalb dauert die Suche mit Metasuchmaschinen in der Regel länger als bei Suchsystemen, deren Datenquellen bereits in Form eines gemeinsamen Index vorliegen. Ein weiterer Nachteil von Metasuchmaschinen ist das Fehlen eines einheitlichen Relevanzrankings bezüglich des Inhalts der recherchierten Quellen.

Als Konsequenz einer früheren MEDPILOT-Evaluation in 2004 wurde – den Bedürfnissen der User folgend – eine einfache, einheitliche Bedienschnittstelle für den Sucheinstieg geschaffen (El-Menouar, 2004). Mit dieser Neuerung wurde 2006 erstmalig ein googleähnlicher Sucheinstieg geboten, der den Nutzererwartungen entsprach (Abbildung 3).

Abbildung 3. Googleähnlicher Sucheinstieg bei MEDPILOT (Stand: 2008).

Jede Suchanfrage startet mit einer voreingestellten Zahl von für Mediziner wichtigen und kostenlosen Datenbanken, die übergreifend mit einfachen Stichworten abgefragt werden können. Durch integrierte erweiterte Suchfunktionen können zusätzlich komplexe Suchanfragen generiert werden.

Die Ergebnisse einer Trafficanalyse[20] im Rahmen der ViFa-SYS-Studie – auf der Grundlage einer Analyse der Logfiles verschiedener Virtueller Fachbibliotheken – bestätigte im

[20] Eine Trafficanalyse beschreibt die „Wanderungsbewegungen" der Nutzer einer Website. Z. B. kann ermittelt werden, wie viele der User, die die Homepage der Website aufrufen, sich tatsächlich bis hin zu einer bestimmten Unterseite der Website bewegen und wie viele davon vorher die Website verlassen.

Nachhinein die Entscheidung für ein einfach strukturiertes Design (vgl. Heinold und Spiller, 2007). Bei acht der 42 untersuchten Virtuellen Fachbibliotheken wurde eine solche Analyse durchgeführt. Für MEDPILOT zeigte sich die positivste Bilanz in der Nutzerstatistik. Aufgrund der Fokussierung auf einen Suchschlitz als dem wesentlichen Bestandteil des Zugangs zur virtuellen Fachbibliothek Medizin sprangen im Vergleich zu den anderen ViFAs die wenigsten User bereits auf der Eingangsseite ab.

Obwohl sich die Performance und die Benutzeroberfläche von MEDPILOT mit dem Relaunch in 2006 deutlich verbesserten, war die Suchmaschine bei vielen sprachlich problematischen Suchanfragen noch ebenso wenig effektiv, wie viele ihrer Konkurrenten (z. B. PubMed). Für MEDPILOT blieb die Verarbeitung der folgenden Sprachaspekte weiterhin schwierig: ein intelligenter und toleranter Umgang mit Rechtschreibfehlern und lexikalischen Varianten, das automatische Erkennen von synonymen Suchbegriffen oder Komposita, das Erkennen von relevanten fremdsprachlichen Treffern sowie eine Auflösung von Akronymen.

2.3.2 MEDPILOT und die Probleme des medizinischen Information-Retrievals

Die Variationen der natürlichen Sprache, insbesondere der medizinischen Fachsprache, stellen enorme Herausforderungen für gängige Text-Retrieval-Systeme dar. Bisher mangelt es den meisten Retrieval-Systemen an morphologischer und lexikalisch-semantischer Funktionalität sowie an der Fähigkeit zur Analyse großer mehrsprachiger Dokumentenbestände.

Zu den sprachlichen Variationen werden morphologische, syntaktische und lexiko-semantische Variationen gezählt. Die *morphologischen Variationen* beinhalten Phänomene wie Flexion, Derivation und Komposition. *Syntaktische Variationen* sind Variationen auf der Ebene von Mehrwortausdrücken oder Sätzen. *Lexiko-semantische Variationen* beziehen sich auf die Bedeutung von Wörtern und Mehrwortausdrücken. Hierzu zählt die Bedeutungsgleichheit verschiedener Wörter (Synonymie) gleichermaßen wie die verschiedenen Bedeutungen eines Wortes (Homonymie).

Da die bisherige MEDPILOT-Suche technologisch auf einem exakten (bzw. bei der Trunkierung auf dem partiellen) Abgleich zwischen dem Suchterm und den Wörtern in den Zieldokumenten basiert, konnten sprachliche Variationen bisher gar nicht oder nur sehr unzureichend mit einer Anfrage gefunden werden (z. B. ‚Carcinom' – ‚Karzinom', ‚Statistik' – ‚Statistiken'). Ein Suchterm wie ‚Karzinom' führt exakt dort zu einem Treffer, wo dieses

Wort buchstabengetreu in einem Dokument vorkommt. Bei Verwendung eines Wildcard-operators (z.B. *), werden mit dem Suchterm ‚Karzinom*' hingegen alle Dokumente gefunden, in denen ein Textwort mit der Zeichensequenz ‚Karzinom' beginnt und sich daran eine beliebige Buchstabensequenz anschließt (etwa ‚Karzinomverdacht'). Als Alternative wird oftmals das Stemming als Standardmethode zur morphologischen Analyse im Information-Retrieval eingesetzt. Mittels Ersetzungsregeln werden morphologische Varianten eines Lexems auf einen gemeinsamen Pseudostamm zurückgeführt, wobei jedoch gerade dieser Ansatz Schwächen bei der Behandlung kontextsensitiver Effekte der deutschen Morphologie (etwa der Umlautung im Falle ‚Krampf' und ‚Krämpfe') aufweist.

Die Recherche mit einem Kompositum als Suchterm (z.B. ‚Todesursachenstatistik') findet fast ausschließlich nur Treffer, die genau diesen Begriff enthalten. Treffer, in denen die Einzelbestandteile des Kompositums (‚Todesursache' ‚Statistik') vorkommen, werden nicht oder nur sehr eingeschränkt gefunden. Bedeutungsgleiche Wörter (Synonyme) können nur über den Umweg eines eingebundenen Thesaurus bzw. Schlagwortkatalogs gefunden werden.

Neben den generellen sprachlichen Variationen bietet die Domäne der Medizin eine spezielle Problematik: Viele Krankheiten, Diagnose- und Therapieverfahren etc. besitzen mehrere unterschiedliche fachspezifische Bezeichnungen und werden darüber hinaus auch mit laiensprachlichen Begriffen belegt. In den herkömmlichen Retrieval-Systemen kommt es z.B. bei der Eingabe des Suchterms ‚Nasennebenhöhlenentzündung' einerseits und ‚Sinusitis' andererseits zu völlig unterschiedlichen Trefferrückmeldungen, obwohl beide Begriffe das Gleiche bedeuten. Hinzu kommt, dass in der Medizin eine Unzahl von Akronymen existiert. Teilweise besitzen Akronyme mehrere verschiedene Bedeutungen, sodass es für herkömmliche Suchsysteme sehr schwierig ist, diese korrekt aufzulösen (‚i.v.' kann z.B. bedeuten: ‚in vivo' oder ‚in vitro'; aber auch: ‚intravenös' oder ‚intravaskulär').

Ein weiteres großes Problem der bisherigen MEDPILOT-Suche ist die fremdsprachliche Inflexibilität. Es werden nur Suchergebnisse in der jeweiligen Anfragesprache gefunden. Eine mehrsprachige Suche ist bisher ebenso wenig möglich. Darüber hinaus lässt sich mit dem bisherigen System, die für das Deutsche typische Bildung von Komposita, nicht adäquat abdecken (z.B. ‚Herzmuskelentzündung' vs. ‚Entzündung des Herzmuskels').

2.4 Die MorphoSaurus-Technologie

Für die Lösung der oben beschriebenen Probleme entwickelte die Firma Averbis die so-genannte Averbis Core Engine (CE)[21]. Es handelt sich um einen innovativen Ansatz zur maschinellen Sprachverarbeitung, welcher darauf beruht, Wörter nicht mehr als Ganzes zu erfassen (vgl. Daumke, 2007; Markó, 2008; Daumke, Schulz, Müller, Dzeyk, Pacheco, Cancian, Nohama & Markó, 2009). Stattdessen werden Wörter in ihre kleinsten Bestandteile (Subwörter bzw. Morpheme im weiteren Sinne) zerlegt. Dabei handelt es sich sprachlich um semantisch kleinste Einheiten. Subwörter gleicher Bedeutung werden sprachübergreifend in Basiskonzepte gruppiert. Diese einfache und innovative Grundidee lässt sich am besten anhand eines Beispiels erklären (vgl. Abbildung 4):

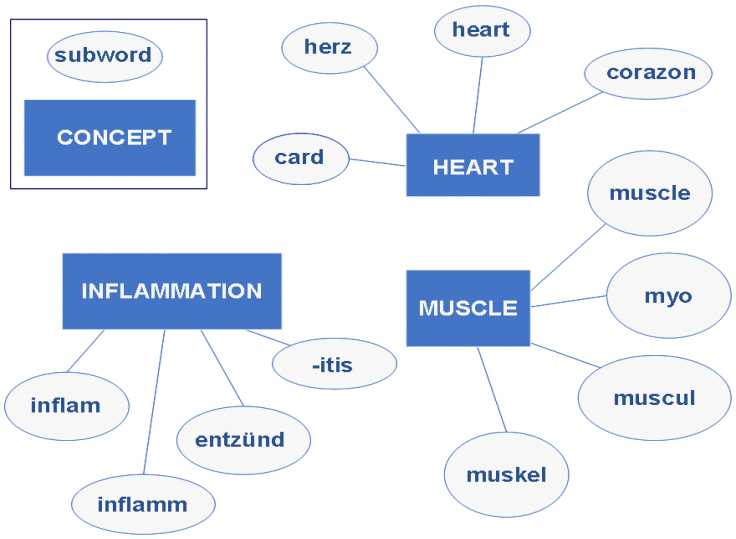

Abbildung 4. Relation von Subwörtern und Konzepten (aus Dzeyk & Markó, 2008, S. 16).

Das Wort ‚Herzmuskelentzündung' wird durch die Averbis CE zunächst in die Subwörter ‚Herz', ‚Muskel' und ‚Entzünd(ung)' zerlegt und anschließend auf die zugehörigen Basis-konzepte HEART, MUSCLE und INFLAMMATION abgebildet. Ebenso werden auch das deutsche Wort ‚Myokarditis', das englische ‚myocarditis' oder die Phrase „inflammation of the heart muscle" durch die Averbis CE in die Konzepte HEART, MUSCLE und INFLAMMATION zerlegt.

[21] Wenn im Folgenden von der „Averbis Core Engine" die Rede ist, so ist dies gleichbedeutend mit der „MorphoSaurus-Technik".

Diese Funktionalität gründet sich also auf eine Art Wörterbuch bzw. Thesaurus, der Äquivalenzklassen zu Subworten enthält. Zu jedem Subwort werden also zusätzliche Informationen abgespeichert, wie die Art des Subwortes (Präfix, Suffix, Wortstamm etc.), die Zugehörigkeit zu einer bestimmten Sprache oder einer Äquivalenzklasse. Auf diese Weise lassen sich auch Äquivalenzklassen miteinander verbinden, die die gleiche semantische Bedeutung haben. Somit ist es unerheblich, in welcher Variation eine Anfrage formuliert und in welcher Sprache sie gestellt wird – durch die intelligente Sprachverarbeitung wird in der Regel die gleiche Anzahl an relevanten Treffern gefunden.

Nach Markó (2008, S. 30) besteht der Prozess, mit dem Dokumente in eine mehrsprachige Darstellungsform überführt werden, aus folgenden drei Schritten (vgl. Abbildung 5):

1. *Orthografische Normalisierung*. In einem ersten Schritt erfolgt die Normalisierung von Wörtern: Großbuchstaben werden hierbei in Kleinbuchstaben gewandelt und sogenannte „diakritische Zeichen“ entfernt (z. B. Punkte, Striche, Häkchen oder kleine Kreise, die eine besondere Aussprache oder Betonung markieren). Zudem werden länderspezifische Zeichen, wie im Deutschen z. B. das ‚ß‘, in eine (sprachübergreifende) Normalform überführt (‚ss‘).

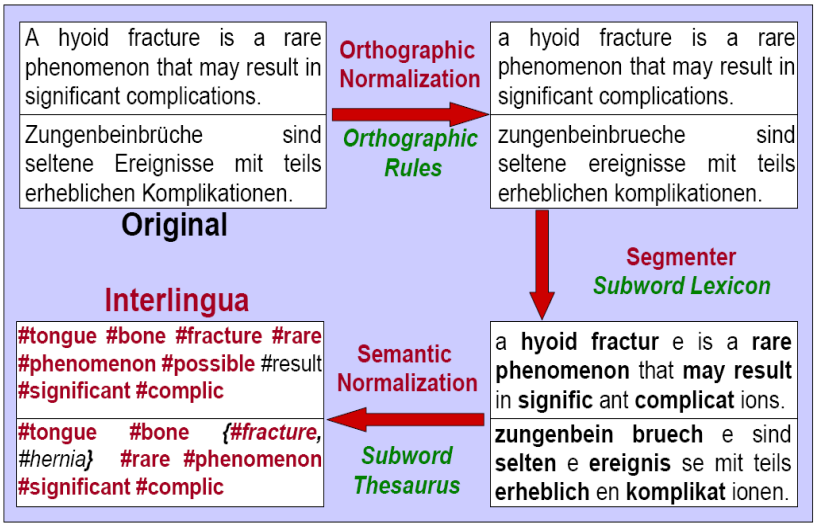

Abbildung 5. Überführung in eine mehrsprachige Darstellungsform (aus Markó, 2008, S.31).

2. *Subworterstellung*. Ein morphosyntaktischer Parser zergliedert Wörter in Subwortketten (Segmenter) und überführt diese in ein Subwortlexikon.

3. *Semantische Normalisierung*: Alle bedeutungstragenden Subworte werden durch einen sprachunabhängigen, synonymen semantischen Bezeichner („MID“) ersetzt. Die korrekte Identifikation doppeldeutiger Wörter und deren Zuordnung zu den entsprechenden Be-

zeichnern erfolgt mithilfe eines stochastischen Modells, welches durch die Analyse des (Kon-)Textes die wahrscheinlichste Variante auswählt.

Die Betrachtung von Basiskonzepten als grundlegende Analyseeinheiten eröffnet die Möglichkeit, selbst riesige Begriffswelten wie die der Medizin mit einem überschaubaren Inventar an lexikalischen Einheiten abzudecken. Im Vergleich zu Vollwort-Lexika wird die Größe der lexikalischen Ressourcen um das ca. Zehnfache reduziert. Zugleich verringert sich der Aufwand für die Erstellung, Integration und Pflege neuer Begrifflichkeiten auf ein Minimum.

Unscharfe Anfragen: Eine zweite zentrale Herausforderung an Suchmaschinen besteht in einer sinnvollen Treffereinschränkung der häufig kurzen und ungenauen Suchanfragen. Eine Nutzeranalyse der MEDPILOT-Suchmaschine (bzw. des Logfiles) hat ergeben, dass die Mehrzahl der Suchenden möglichst wenig kognitiven Aufwand in die Recherche investiert – mehr als 60 % der Nutzer stellen nur Ein- und Zwei-Wort-Anfragen. Darüber hinaus werden die erweiterten Suchmöglichkeiten (Phrasensuche, Operatoren, Wildcards, Feldsuche) nur wenig genutzt. Herkömmliche Suchmaschinen haben große Probleme, diejenigen Nutzer zufrieden zu stellen, die lediglich Ein-Wort-Anfragen formulieren.

In diesen Fällen kann eine sogenannte *Navigationssuche* (‚faceted search') weiterhelfen. Mit definierten Begriffen aus passenden Klassifikationen können Anwender durch das Auswählen einzelner Kategorien bequem durch eine sonst unerschließbare Menge von Informationen geführt werden. Die Averbis GmbH hat im Zusammenhang mit der MorphoSaurus-Technologie eine auf die medizinische Fachsprache spezialisierte *Facettensuche* entwickelt. Bei Eingabe eines Suchbegriffes werden verwandte Suchbegriffe automatisch nach Krankheiten, Medikamenten, Operationen etc. kategorisiert angeboten. Die Nutzer werden somit durch die Möglichkeit des *Browsings* in ihrem Rechercheprozess unterstützt. Für die Navigationssuche werden Terme aus dem MeSH-Katalog[22] verwendet. Bei diesen sogenannten Medical Subject Headings handelt es sich um fachspezifische Stichwörter, mit denen Fachliteratur in der größten medizinischen Datenbank MEDLINE verschlagwortet wird.

[22] MeSH = „Medical Subject Headings". Hierbei handelt es sich um einen kontrollierten Satz (Vokabular) von ca. 16.000 Schlagworten, die hierarchisch in 15 Baumstrukturen wie z. B. Anatomie, Krankheiten oder Medikamente aufgegliedert sind. Zusätzlich gibt es einen Satz von 76 Unterpunkten („subheadings"), wie z. B. Therapie oder Diagnose, die die Spezifität der eigentlichen Schlagwörter erhöhen.

Neben der Möglichkeit der Navigationssuche können dynamische Vorschlagsfunktionen für Suchanfragen eine wichtige Hilfe für Benutzer sein (vgl. Shneiderman, 1994). Diese erlauben es, Ein- oder Zwei-Wort-Eingaben so zu spezifizieren und zu ergänzen, dass den Nutzern Treffer mit höherer Relevanz präsentiert werden können.

2.5 Suchverhalten und Usability von Suchmaschinen

Web-Information-Retrieval-Systeme müssen prinzipiell eine noch größere Benutzerfreundlichkeit aufweisen als andere Web-Angebote, da hier *die unterschiedlichsten Nutzergruppen* auf Anhieb verstehen müssen, wie das System funktioniert. Nach einem kurzen Überblick über die verwendeten Datenerhebungsmethoden werden in den folgenden Abschnitten wesentliche Ergebnisse der Suchmaschinen-Nutzerforschung sowie Erkenntnisse im Bereich der Usability von Suchmaschinen beschrieben.

2.5.1 Datenerhebungsmethoden

Für die Untersuchung des Nutzerverhaltens werden in der Forschung folgende Erhebungsmethoden eingesetzt: Befragungen (Fragebogen, Interviews, Fokusgruppen), Beobachtungsstudien, Laborexperimente sowie Logfile-Analysen. Je nach Anlage der Untersuchung können Usability-Tests und Eyetracking-Studien als Beobachtungsstudien oder Laborexperimente angelegt sein. Zu den eingesetzten Methoden informiert beispielsweise Lewandowski (2005). Am häufigsten werden szenariobasierte Usability-Tests eingesetzt.

Andere alternative Methoden sind z. B. das Card Sorting, die Kurzanalyse durch Experten, die Heuristische Evaluation, der Cognitive Walkthrough, die Benutzerbefragung, Online-Usability-Tests sowie Online-Befragungen über Online Panels. Nach Meinung von Experten ist der Erkenntnisgewinn bei Usability-Tests – im Verhältnis zum Ressourceneinsatz anderer Methoden – jedoch am größten (vgl. Jacobsen, 2005). Zu den speziellen Methoden, Hintergründen und Ergebnissen des Usability-Testings informieren ausführlich beispielsweise Krug (2006) sowie Nielsen & Loranger (2006), die auch über neuere Ergebnisse der Usability-Forschung im Bereich der Suchmaschinennutzung berichten. Einen guten Überblick über die psychologischen Hintergründe liefert etwa Wirth (2004). Über die Möglichkeiten der Optimierung der Web-Usability informiert z. B. Kalbach (2008).

2.5.2 Suchverhalten von Suchmaschinen-Nutzern

In der Sozialpsychologie existiert der Begriff des „kognitiven Geizkragens“ (‚cognitive miser‘) zur Beschreibung der menschlichen Neigung, sich nicht unbedingt mehr anzustrengen als nötig, um eine Aufgabe zu erledigen oder eine Situation einzuschätzen. Bei diesem Prinzip der kognitiven Ökonomie geht es darum, mit möglichst geringem Einsatz – unter Verwendung von (kognitiven) Heuristiken – einen möglichst großen Effekt zu erzielen (vgl. z. B. Tversky & Kahneman, 1974 sowie Wirth & Schweiger, 1999). Die Ergebnisse, die bisher zum Suchverhalten in webbasierten Information-Retrieval-Systemen vorliegen, scheinen diese These zu bestätigen. So lässt sich damit etwa erklären, warum die meisten Suchanfragen nur aus Ein- und Zwei-Wort-Termen bestehen. Analysen zum Userverhalten, die aus Logfiles, Inhaltsanalysen und Nutzerbeobachtungen gewonnen wurden, haben z. B. Schmidt-Mänz (2007), Jansen (2006), Schmidt-Mänz und Bomhardt (2005) oder Beitzel, Jensen, Chowdhury, Grossman und Frieder (2004) vorgelegt. Die bisher umfangreichste *Umfrage* zum Verhalten von Nutzern im Umgang mit Suchmaschinen – in Deutschland – hat Machhill (2003) publiziert. Die folgenden Abschnitte beschreiben zusammenfassend wichtige Erkenntnisse jüngerer Forschungsarbeiten zum *allgemeinen Suchverhalten* von Suchmaschinennutzern:

1. *Es werden nur wenige Suchworte benutzt*. Die Mehrheit der Suchphrasen besteht aus Ein- bis Drei-Wort-Anfragen. Im Durchschnitt umfasst eine Suchanfrage 2,6 Suchworte. Diese Kurzanfragen machen ca. 2/3 der gesamten Suchanfragen aus (vgl. Spink & Jansen, 2004), wobei Ein-Wort-Suchen am häufigsten zu beobachten sind. Untersuchungen neueren Datums konstatieren jedoch eine Rückläufigkeit der Ein-Wort-Suchen. So waren es in der Studie von OneStat (2007) nur noch ca. 15% der Suchanfragen, die lediglich aus einem Wort bestanden. Die meisten User verwendeten – laut dieser Studie – zwei (31,9%) oder drei Worte (27%) für die Formulierung ihrer Suche. Die Tendenz zur Benutzung von mehreren Suchworten wird mit der wachsenden Informationskompetenz der User erklärt.

2. *User machen Rechtschreibfehler*. Ob aus Zeitdruck, Unaufmerksamkeit oder Unkenntnis: Die Nutzer begehen bei der Eingabe ihrer Suchterme regelmäßig orthografische Fehler. Zu den häufigsten Fehlern zählen dabei Buchstabendreher, doppelt angeschlagene Tasten und Auslassungen. Nielsen und Loranger (2006, S. 152) fanden in 7,5% der Suchanfragen Rechtschreibfehler.

3. *Usern fehlen oft Informationen*. Um eine gute Suchanfrage zu formulieren, fehlt es den Usern häufig an Hintergrundwissen und Informationen, um die Suche zu reformulieren oder zu spezifizieren.

4. *Logische Operatoren werden nur selten genutzt*. Boolesche Operatoren finden nur in etwa jeder zehnten Suchanfrage Verwendung (Wolfram, Spink, Jansen & Saracevic, 2001; Spink & Jansen, 2004). Zudem fanden Jansen, Spink und Saracevic (2000) Belege dafür, dass ca. die Hälfte der Anfragen mit Booleschen Operatoren Fehler enthalten.

5. *Vor allem die ersten Treffer werden genutzt*. Die meisten User nutzen bei einer Recherche allerdings nur die ersten Links einer Trefferliste (vgl. hierzu z. B. Machhill 2003 oder Jansen, Spink & Pedersen, 2005). In einer Studie von Jupiter Research (iProspect, 2006) zeigte sich, dass 62 % der Suchmaschinennutzer bereits auf ein Suchergebnis innerhalb der ersten Trefferseite klickten. 90 % der User klickten auf einen Treffer innerhalb der ersten drei Ergebnisseiten. Nielsen und Loranger (2006, S. 27) berichten, dass ihre Testpersonen in 93 % der Suchanfragen nur die erste Seite der Ergebnisliste aufriefen (die zehn Ergebnisse umfasste). Nur 47 % der Testpersonen waren dazu bereit, durch die Ergebnisliste zu den unteren Treffern zu scrollen. Auf den Treffer an Position eins entfielen mehr als die Hälfte der Klicks (51 %), auf den zweiten Treffer klickten noch 16 % und auf den dritten Treffer nur noch sechs Prozent der Testpersonen (vgl. Tabelle 3). Nur sieben Prozent der Klicks entfielen auf die zweite Ergebnisseite.

Tabelle 3. Wohin die Benutzer auf der Ergebnisseite klicken (aus Nielsen & Loranger, 2006, S. 38).

Position in den Suchmaschinenergebnissen	**Klicks auf Links in dieser Position**
#1	51 %
#2	16 %
#3	6 %
#4	6 %
#5	5 %
#6	4 %
#7	2 %
#8	1 %
#9	1 %
#10	2 %
#11	5 %

Hinweis: Die Zahlen ergeben aufgrund der Rundung keine vollen 100 %.

6. *Mehr Aufmerksamkeit und mehr Vertrauen gegenüber den ersten Treffern*. Ergebnisse von Studien, die unter Einsatz von Eyetracking-Systemen durchgeführt wurden, bestätigen die Beobachtung, dass die User besonders den ersten Treffern einer Liste Aufmerk-

samkeit schenken und diese vertrauenswürdiger finden, auch wenn diese Treffer weniger relevant sein mögen als nachgeordnete Treffer (vgl. Pan, Hembrooke, Joachims, Lorigo, Gay & Granka, 2007 sowie Lorigo, Haridasan, Brynjarsdóttir, Xia, Joachims, Gay & Granka, 2008).

7. *Nutzer neigen zu Reformulierungen und Verbesserung von Suchanfragen*. 41% der Suchmaschinen-Nutzer, die keine zufriedenstellenden Treffer auf der ersten Ergebnisseite fanden, wechselten entweder die Suchmaschine oder veränderten ihren Suchterm (iProspect, 2006). Spink, Jansen, Wolfram und Saracevic (2002) haben beobachtet, dass Suchmaschinennutzer ihre Anfragen in 40 bis 52 Prozent der Fälle reformulieren, um das Gesuchte zu finden. Laut Machhill und Welp (2003) gehören zu den wichtigsten Veränderungen von Suchanfragen folgende Kategorien:

   - „Streichung oder Ergänzung von zentralen Begriffen, die aus der Formulierung der Suchanfrage entnommen sind.
   - Veränderung der logischen Operatoren in der Suchanfrage (das logische UND bzw. ODER).
   - Verbesserung von Tippfehlern oder von unzweckmäßigen Anfragen in Form von Halbsätzen, Fragesätzen etc.." (S. 247, ebd.).

8. *Suchverhalten von Experten und Laien*. Neben Studien zum allgemeinen Nutzerverhalten wurden auch Untersuchungen zu bestimmten Nutzergruppen durchgeführt (für eine Übersicht vgl. Spink & Jansen 2004, S. 21ff.). So haben etwa Hölscher und Strube (2000) das Suchverhalten von Experten und Laien untersucht und festgestellt, dass Experten bzw. Informationsspezialisten durchaus komplexere Suchanfragen formulieren (im Durchschnitt 3,64 Suchterme) und darüber hinaus auch die Boolesche Syntax häufiger einsetzen als Laien. Am effektivsten bei der Suche sind jedoch Personen, so die Autoren, die eine hohe inhaltliche Expertise besitzen und zugleich über eine hohe Webkompetenz verfügen.

9. *Suchverhalten von Medizinern*. Über das spezielle Informations-Suchverhalten von Medizinern im Web ist bisher noch nicht viel bekannt. Das *allgemeine Informations-Suchverhalten* von Ärzten wird beispielsweise bei Davies (2007) beschrieben. Die Erwartungen von Medizinern an eine Literatursuchmaschine hat z.B. El-Menouar (2002; 2004) beschrieben. Anhand der Logfiles eines Tages analysierten Herskovic, Tanka, Hersh und Bernstam (2007) die Benutzung der Datenbank PubMed durch Mediziner. Dabei stellten

die Forscher fest, dass die meisten Anfragen ebenfalls aus nur einem Suchbegriff bestanden. Lediglich 11,3% der Suchanfragen enthielten einen Booleschen Operator. Insofern unterscheidet sich das Suchverhalten von Medizinern nicht wesentlich von dem Verhalten anderer User.

Obwohl die Nutzer über ganz unterschiedliche Kompetenzen in der Formulierung ihrer Informationsbedürfnisse verfügen, investiert die Mehrzahl der Suchenden also möglichst wenig Aufwand in eine Recherche. Diese Strategie ist durchaus sinnvoll und erweist sich unter Umständen als rational, erst recht, wenn Nutzer nicht genau wissen, nach was sie suchen. Die Suche mit wenigen und einfachen Suchworten sorgt dafür, dass die Ergebnisse zunächst weit streuen. In einem zweiten Suchvorgang können die Suchbedürfnisse dann näher spezifiziert werden. Aber auch das gegenteilige Vorgehen lässt sich beobachten: Von einer eher spezifischen Suche mit mehreren Suchbegriffen, die zunächst nur wenige Treffer bringt, wird durch Weglassen und / oder Austauschen von Suchtermen eine Verallgemeinerung der Suchanfrage angestrebt und damit auch die Treffermenge erhöht. Das Prinzip der kognitiven Sparsamkeit könnte auch dafür verantwortlich sein, dass die erweiterten Suchmöglichkeiten (Phrasensuche, Operatoren, Wildcards, Feldsuche) im Allgemeinen von Usern nur wenig genutzt werden. Untersuchungen zur Usability von Bibliothekskatalogen scheinen diese Beobachtungen zu belegen (vgl. Schulz, 2002 sowie White, Wright & Chawner, 2006).

Ursula Schulz (2001a) fasst die bisherigen Erkenntnisse zur Nutzerforschung pointiert zusammen:

- „Diese Besucher interessieren sich für ihre eigenen Probleme - nicht für unsere (Informationsexperten).
- Sie kennen ihre eigene Terminologie - nicht aber unsere.
- Sie glauben, dass wir ihre Zeit stehlen, wenn wir erwarten, dass sie sich in unsere Wissensorganisation einarbeiten oder Einführungen lesen.
- Sie sind in Eile und interessieren sich nicht für unsere Idee einer professionellen inhaltlichen oder formalen Erschließung. Wenn sie nach 10 Sekunden nicht verstehen, dass die Site 'ihnen etwas bringt', halten sie sich an bereits bekannte oder einfachere Alternativen.

Kurz: Besucher im Web gehen den Weg des geringsten (kognitiven) Aufwands.“ (S. 6-7, im Onlinedokument, ebd.)

Für den Bereich der Bibliotheken und der über sie angebotenen Suchmaschinen bedeutet dies: Ein Recherche-Interface für die Suche nach wissenschaftlicher Literatur sollte möglichst einfach, verständlich und benutzerfreundlich gestaltet werden.

2.5.3 Usability von Suchmaschinen

In den folgenden Ausführungen geht es um die Benutzerfreundlichkeit des Suchmaschinen-Interface (‚search engine usability'). Eine breite wissenschaftliche Auseinandersetzung mit den besonderen Usability-Problemen von Suchmaschinen hat bisher noch nicht stattgefunden, daher werden die bisherigen Erkenntnisse, Positionen und Einsichten von Autoren, die sich bereits mit diesem Thema auseinandergesetzt haben, hier kurz referiert.

Neben den allgemeinen Usability-Prinzipien der Effektivität, Effizienz und Akzeptanz (DIN ISO 9241-11, vgl. Kap. 2.2.1.1) sind bei Benutzungsschnittstellen von interaktiven Systemen besonders folgende Aspekte der *Softwareergonomie* bzw. *Dialoggestaltung* nach der internationalen ISO-Norm (EN ISO 9241, Teil 110, früher Teil 10) zu berücksichtigen:

a. *Aufgabenangemessenheit*
b. *Selbstbeschreibungsfähigkeit*
c. *Steuerbarkeit*
d. *Erwartungskonformität*
e. *Fehlertoleranz*
f. *Individualisierbarkeit*
g. *Lernförderlichkeit*

Die folgenden Abschnitte beschreiben die genannten Normen zur Gestaltung einer *Dialogschnittstelle*, wobei zunächst die jeweiligen Abschnitte aus der ISO-Norm zitiert werden. Anschließend erfolgt eine Übertragung auf den Bereich der Suchmaschinensysteme, sodass konkret nachvollziehbar wird, welche Bedeutung die an sich abstrakten Normen für die praktische Ausgestaltung der Suchmaschinen-Usability besitzen. Die Empfehlungen sind u.a. abgeleitet aus den Überlegungen von Schulz (2001a; 2001b; 2002).

Zu a.: Aufgabenangemessenheit. **„Ein Dialog ist aufgabenangemessen, wenn er den Benutzer unterstützt, seine Arbeitsaufgabe effektiv und effizient zu erledigen."**

Effektivität kann als Maß der Zielerreichung bezeichnet werden. Sie fragt danach, *wie gut* es den Nutzern gelingt, ihre Ziele zu erreichen. Übertragen auf den Bereich der Suchmaschinen bedeutet *Effektivität*, dass die Nutzer tatsächlich das finden, was sie zu suchen beabsichtigten (beispielsweise einen Artikel zum Thema Diabetes). Die *Effizienz* hingegen beschäftigt sich mit der Frage, welchen Aufwand Nutzer z. B. für eine Recherche betreiben müssen.

Auch wenn eine Fachsuchmaschine alle wesentlichen Texte, Abstracts usw. zu einem Fachgebiet enthalten sollte; eine Suchmaschine ist nicht aufgabenangemessen gestaltet, wenn die Recherche deshalb so lange dauert, weil die Oberfläche so umständlich zu bedienen ist, dass Rechercheziele nicht oder nur unter großem Zeitaufwand erreicht werden können. Auswirkungen auf die Effizienz hat auch der technische Hintergrund des Angebots (vgl. Kap. 2.2.1). Wird z. B. eine exotische Technologie verwendet, die die Nutzer dazu zwingt zusätzliche Plugins zu installieren, damit die Anwendung flüssig läuft oder überhaupt funktioniert, kann dies die Nutzer dazu veranlassen, diese Website generell zu meiden. Dies hängt jedoch vom Genre des Internetangebots ab: Bei Websites, bei denen eher die Unterhaltung im Vordergrund steht, spielen Verzögerungen durch Technik eine nicht ganz so große Rolle. Folgende Empfehlungen lassen sich für die Gestaltung eines Suchmaschinenangebots ableiten:

- *Moderne Suchtechnologie und intelligente Algorithmen*. Die Effektivität und Effizienz wird ganz wesentlich auch von den Möglichkeiten der Suchmaschine zur Analyse der Nutzereingaben bestimmt. Der Einsatz moderner Suchmaschinentechnologie und intelligenter Algorithmen zur Analyse von Suchtermen führt zu einer Steigerung der Rechercheleistung des Systems.
- *Einsatz optimierter Grafiken*. Grafiken sind so gestaltet, dass beim Seitenaufbau möglichst wenig Wartezeit entsteht.
- *Schnelle Suche*. Der eigentliche Suchvorgang sollte ebenfalls nur einige Sekunden in Anspruch nehmen.
- *Vorpositionieren des Cursors*. Der Cursor sollte gleich bei Aufruf des Angebots auf das Eingabefeld des Suchschlitzes gesetzt sein, damit kein Zeitverlust entsteht.
- *Bereitstellen von Unterstützungsfunktionen*. Vorschlagsfunktionen, die Nutzer bei der Auswahl von Suchworten unterstützen, können den Suchprozess effektiver und effizienter machen, ebenso der Einsatz von Filtern oder Möglichkeiten zur Einschränkung des Suchraums bzw. zur Verfeinerung der Suche.

Neben der Recherche können auf einer Suchmaschinen-Website auch noch andere Ziele verfolgt werden, wobei ebenfalls auf Effektivität und Effizienz geachtet werden sollte:

- Die Kontaktaufnahme mit dem Anbieter: Die Ansprechpartner müssen leicht zu finden sein, sodass ein Benutzer/Kunde sein Ziel, eine E-Mail persönlich zu adressieren, erreichen kann
- Die Bestellung und der Download von Artikeln, Datenbankeinträgen etc.
- Das Vornehmen von persönlichen Einstellungen (Personalisierung)
- Die Benutzung von weitergehenden Suchhilfen (erweiterte Such- oder Filterfunktionen)

Insgesamt sollte die grafische Gestaltung des Suchmaschinenangebots die Nutzer so unterstützen, dass sie ihre Ziele ohne große Probleme und Zeitverlust erreichen können.

Zu b.: Selbstbeschreibungsfähigkeit. **„Ein Dialog ist selbstbeschreibungsfähig, wenn jeder einzelne Dialogschritt durch Rückmeldung des Dialogsystems unmittelbar verständlich ist oder dem Benutzer auf Anfrage erklärt wird."**

Die Selbstbeschreibungsfähigkeit gehört zu den wichtigsten Einflussgrößen für ein rasches Verständnis des Inhaltsangebots und der Navigation eines Systems. Die Nutzer sollten die Bezeichnungen und grafischen Elemente für die verschiedenen Funktionen innerhalb kürzester Zeit verstehen können. Angesprochen sind hier einerseits die Textverständlichkeit der Website sowie die Benennung der Funktionen und andererseits die Verständlichkeit von Grafik- und Designelementen als Teil der Benutzerführung. Für die Textverständlichkeit sind neben ergonomischen Aspekten wie Schriftart, Schriftgröße und Zeilenabstand vor allem inhaltliche Aspekte z. B. eine zielgruppenspezifische Sprache von Bedeutung. Grafik- und Designelemente sollten primär kein Selbstzweck sein, sondern durch die Art ihrer Gestaltung das Verständnis für den Dialogfluss mit dem System fördern.

- *Was bietet die Suchmaschine an?* Die Eingangsseite der Suchmaschine muss schnell, einfach und verständlich vermitteln können, worum es bei dem Angebot geht. Die Alleinstellungsmerkmale und der Mehrwert des Angebots müssen unmittelbar begreiflich sein.
- *Wird durch die Wahl der Bezeichnung Eindeutigkeit geschaffen?* Wenn die Nutzer übermäßig lange Zeit für das Verständnis brauchen, kann es daran liegen, dass die Bedeutung oszilliert. Die Nutzer sind sich nicht sicher. Unsicherheit verzögert den Weg zum Ziel, kann zum Abbruch des Navigationspfades beitragen und verringert letztlich die Akzeptanz des Angebots.

- *Handelt es sich um eine allgemein verständliche Bezeichnung oder einen Terminus, den nur wenige Fachleute verstehen?* Wendet sich die Suchmaschinenoberfläche nicht vorwiegend an Spezialisten, sind Bezeichnungen zu bevorzugen, die auch Laien problemlos verstehen können. Es mag viele Fachleute einer Domäne geben, aber beim Verständnis von Suchfunktionen scheinen wir erst am Beginn der Herausbildung professioneller Informationskompetenz zu stehen. Viele Begriffe der Informationstechnologie, die zwar Experten geläufig sind, werden von vielen Nutzern nur zum Teil oder gar nicht verstanden (vgl. eResult-Studie, 2008). Deshalb sollten auch in der Bezeichnung von Funktionen die Prinzipien der Textverständlichkeit wie Einfachheit Gliederung, Ordnung, Kürze, Prägnanz (vgl. Langer, Schulz v. Thun & Tausch, 1993) sowie kognitive Strukturierung Beachtung finden (vgl. Christmann & Groeben, 1999). Hilfreiche Beschreibungen für mediengerechtes Texten im Internet finden sich z. B. bei Heijnk (2002) oder Schmider (2003).
- *Benennung und Gestaltung von Such- Sortier- und Filterfunktionen*. Wenn die Benennung bestimmter Navigations- oder Funktionsbuttons nicht zweifelsfrei verständlich ist, kann dies bei den Nutzern zu Missverständnissen führen. Die Nutzer gelangen letztendlich gar nicht oder erst viel später dorthin, wohin sie möchten bzw. hin sollen. Oder sie werden enttäuscht, wenn durch eine bestimmte Benennung falsche Erwartungen geweckt werden, die sich nach einem Klick auf einen Button oder Link nicht erfüllen. Für ein Verständnis des theoretischen Hintergrunds sind hier die Erkenntnisse der Textpsychologie und Verständlichkeitsforschung relevant (vgl. z. B. Groeben, 1982 sowie Langer, Schulz von Thun & Tausch, 1993). Fundierte Hinweise im Zusammenhang mit der Benennung von Funktionen und Buttons liefern beispielsweise Wirth (2004) oder Nielsen und Loranger (2006).

Zu c.: Steuerbarkeit. **„Ein Dialog ist steuerbar, wenn der Benutzer in der Lage ist, den Dialogablauf zu starten sowie seine Richtung und Geschwindigkeit zu beeinflussen, bis das Ziel erreicht ist."**

Haben die Nutzer das Gefühl, dass sie die Website bzw. das Suchmaschinenangebot bezüglich der Funktionen kontrollieren können, oder machen sie die Erfahrung, dass sie ohne Wahlmöglichkeit den Vorgaben des Systems folgen müssen? In der Regel möchten die Nutzer das Gefühl von Kontrolle haben. Für die Steuerbarkeit spielen neben den Navi-

gationsmöglichkeiten insbesondere die Filter- und Sortiermöglichkeiten von Suchmaschinen eine große Rolle. Mit ihrer Hilfe kann die Suchmenge rasch eingeschränkt werden.

Nicht immer aber ist eine umfassende Wahlfreiheit bei der Steuerbarkeit vom Anbieter gewünscht. So ist es manchmal sogar sinnvoll, gewisse Wege in der Benutzerführung vorzugeben, ohne dass darunter die Usability leiden muss. Dies tun Anbieter z. B., um eine Desorientierung der Nutzer zu verhindern oder um den effizientesten Weg zu einem Ziel aufzuzeigen. Allzu strenge Vorgaben können hier aber leicht einen negativen Effekt auslösen. Die wohlgemeinte Unterstützung schlägt dann in ihr Gegenteil um: In eine Bevormundung der User, welche die Einschränkungen mit dem Verlassen der Website quittieren.

Zu d.: Erwartungskonformität. **„Ein Dialog ist erwartungskonform, wenn er konsistent ist und den Merkmalen des Benutzers entspricht, z. B. den Kenntnissen aus dem Arbeitsgebiet, der Ausbildung und der Erfahrung des Benutzers sowie den allgemein anerkannten Konventionen."**

- *Verstehen die Nutzer auf Anhieb, was der Klick auf ein Element auslösen soll?* Hier ist zu fragen, ob es den Nutzern schnell gelingt, die richtigen Hypothesen darüber zu entwickeln, was passiert, wenn sie auf einen Button oder eine Funktion klicken bzw. einem Link folgen. Das Auslösen falscher Erwartungen sollte auf jeden Fall vermieden werden, da es zu Enttäuschungen und Zeitverlust führt.

Zu e.: Fehlertoleranz. **„Ein Dialog ist fehlertolerant, wenn das beabsichtigte Arbeitsergebnis trotz erkennbar fehlerhafter Eingaben entweder mit keinem oder mit minimalem Korrekturaufwand durch den Benutzer erreicht werden kann."**

Das Lösen von Aufgaben darf nicht an den fehlerhaften Eingaben von Nutzern scheitern. Wenn Nutzer nach Eingabe eines Suchterms keine Treffer erhalten, so muss dies von der Suchmaschine ausreichend kommuniziert werden, sonst besteht die Gefahr, dass sich die Nutzer enttäuscht abwenden.

- *Gestaltung der Systemantwort bei Null-Treffer-Meldungen*. Die Anlässe für Null-Treffer-Meldungen können sehr unterschiedlich sein. Ein wichtiger Grund liegt in den Rechtschreibfehlern der Useranfragen, die von der Suchmaschine nicht aufgelöst werden können (vgl. Willson & Given, 2008). Null-Treffer-Meldungen können auch auftreten, weil die von der Suchmaschine indexierten Dokumente tatsächlich nicht das gesuchte

Wort enthalten. Suchmaschinen unterscheiden sich sehr darin, wie sie mit *unverständlichen* Nutzeranfragen umgehen. Sehr verbreitet ist inzwischen die von Google favorisierte *‚Meinten Sie?*‘-Lösung. Dabei wird in der Regel ein Alternativvorschlag in Form eines Links eingeblendet, der als wahrscheinlichste Alternative ausgewählt wurde. Die sogenannten „Spell Checker“ operieren mit integrierten Thesauri und Algorithmen, die die statistischen Wahrscheinlichkeiten von Worten bzw. Wortteilen berücksichtigen. Google geht darüber hinaus und nutzt auch die Analyse der Suchterme seiner Suchmaschinennutzer und bezieht die Häufigkeiten von falschen Schreibweisen in den Algorithmus mit ein (vgl. Whitelaw, Hutchinson, Chung & Ellis, 2009).

Schulz (2001b) ist z. B. der Ansicht, dass eine aussagekräftige Rückmeldung an die User in jedem Fall die bessere Alternative darstellt, als die lapidare Meldung „Kein Treffer gefunden“. Rückmeldungen der Suchmaschine bedeuten in jedem Fall eine Kommunikation mit dem User. Die alleinige Meldung der Trefferanzahl reicht als Rückmeldung auf keinen Fall aus. User haben zwar ein Bedürfnis nach Einfachheit, aber auch nach *Transparenz*. Wenn z. B. mit einem falsch geschriebenen Wort gesucht wird und keine Rückmeldung darüber erfolgt, dass es sich möglicherweise um einen Rechtschreibfehler handelt, werden die User der Suchmaschine in diesem Punkt keine Kompetenz attestieren können.

Zu f.: Individualisierbarkeit. **„Ein Dialog ist individualisierbar, wenn das Dialogsystem Anpassungen an die Erfordernisse der Arbeitsaufgabe, individuelle Vorlieben des Benutzers und Benutzerfähigkeiten zulässt.“**

Gerade Spezialsuchmaschinen versuchen, ihre Nutzer durch bestimmte „Mehrwertdienste“ oder *Personalisierungsmöglichkeiten* an sich zu binden. Hierzu gehört beispielsweise das Anbieten einer *Merkfunktion* für die individuelle Auswahl bestimmter Datenbankquellen, das Aktivieren und Abspeichern bestimmter Funktionen im Hinblick auf einen erneuten Besuch der Website oder das *Markieren von Treffern für den anschließenden Export*. Andere Möglichkeiten zur Individualisierung von Suchmaschinenangeboten sind z. B. ein *Newsletter* mit Hinweisen auf neue Literatur zu vorher festgelegten Suchbegriffen. Ebenso können bestimmte *Navigations-, Filter- oder Sortierfunktionen als dauerhafte Voreinstellung* dafür sorgen, dass die Nutzer das System an ihre persönlichen Vorlieben anpassen können.

Zu g.: Lernförderlichkeit. **„Ein Dialog ist lernförderlich, wenn er den Benutzer beim Erlernen des Dialogsystems unterstützt und anleitet."**

Das Suchmaschinenangebot sollte so gestaltet sein, dass es den Nutzern das leichte Erlernen des Systems ermöglicht und sie dazu ermuntert, die Strukturen der Website zu durchschauen. Dazu sollten die Nutzer durch das System eine ausreichende Unterstützung erfahren. Das kann durch eine *Sitemap*, ein *Online-Tutorial* oder durch das Anbieten von didaktisch gut aufbereiteten *FAQ* (Fequently Asked Questions) geschehen. Hierbei handelt es sich um Antworten auf immer wieder auftretende Fragen und Probleme der Nutzer. Andererseits geht es bei Suchmaschinen auch um direkte Unterstützungshilfen wie *Ergänzung von Suchworten, Funktionen zum Einschränken des Suchraums, Filter- und Sortiermöglichkeiten*. Zur Lernförderlichkeit trägt auch eine optimale *Gestaltung der Antwort- bzw. Trefferpräsentation* bei.

- *Erweiterte Suche.* Zur Unterstützung der Nutzer bei der Suche bieten die meisten Suchmaschinen eine Funktion wie die „Erweiterte Suche" an. Diese beinhaltet zumeist die Möglichkeit über bestimmte Suchfelder die Menge der in Frage kommenden Dokumente einzuschränken, um so zu relevanteren Treffern zu gelangen. Darüber hinaus finden sich hier Möglichkeiten, Boolesche Operatoren einzusetzen sowie verschiedene Arten von Sortiermöglichkeiten (nach Jahr, Autor etc.). Sinnvoll wäre für Mediziner sicher auch die Möglichkeit, die Trefferliste nach Kategorien wie „Journals", „Patientengruppen" oder „evidenzbasierter Literatur" ordnen zu können.

- *Automatische Unterstützung bei Suchanfragen (Automatic Concept-Based Query Expansion).* Ca. zwei Drittel der Nutzer suchen mit Ein- oder Zwei-Wort-Suchtermen. Bei Ein-Wort-Anfragen werden häufig sehr viele Treffer zurückgemeldet, die für die Nutzer nicht relevant sind. Durch sinnvolle Vorschläge zur Ergänzung des Suchterms während der Eingabe können die Nutzer bei der Formulierung Erfolg versprechender Suchanfragen unterstützt werden. Zumeist geschieht dies durch eine Analyse des Suchterms, wobei für die Vorschläge berücksichtigt wird, wie eng das gesuchte Konzept mit anderen Konzepten assoziiert ist. Dabei wird die Häufigkeit des gemeinsamen Auftretens in den Dokumenten der Datenbank oder des Logfiles untersucht. Google und Yahoo bieten diese Funktion schon seit einiger Zeit an. Untersuchungen zeigen, dass die sogenannte Autosuggest-Funktion User erfolgreich bei der Informationssuche unterstützt. Durch das Anbieten dieser Suchwortergänzungen wird das Informationsbedürfnis spezifiziert. Zu

allgemeine Anfragen können so vermieden werden und damit steigt auch die Chance, mehr relevante Treffer zu finden (vgl. Aly, 2008). Durch die Präsentation von Alternativvorschlägen und passenden Suchtermen wird die kognitive Beanspruchung der Nutzer reduziert und daher das kognitive System weniger belastet. So ist aus der Kognitionspsychologie bekannt, dass es Menschen erheblich leichter fällt, etwas unter Zuhilfenahme eines Hinweisreizes wiederzuerkennen (‚recognition'), als sich ohne Hilfe an etwas aktiv zu erinnern (‚recall') (vgl. z. B. Anderson & Bower, 1972).

- *Filtereinsatz.* Eine andere Möglichkeit, den Suchraum nach bereits abgesendeter Suchanfrage einzuschränken und so schneller an relevante Treffer zu gelangen, bietet die Facettensuche moderner Suchmaschinen (‚faceted search'). Dabei werden durch sinnvolle Auswahloptionen Teilmengen der ursprünglichen Treffermenge erzeugt, die im günstigsten Fall übersichtliche Trefferlisten liefern. Dies wird auch als ‚Refinement' bzw. als ‚Eingrenzen' oder ‚Verfeinern' der Suche bezeichnet. Die Möglichkeit, mit wenigen Klicks die Menge der potenziellen Treffer zu reduzieren, stellt für die User eine effiziente Heuristik für den Suchprozess bereit.
- *Design der Antwortpräsentation.* Hier geht es darum, wie ein einzelner Treffer einer Trefferliste gestaltet werden muss, um die Nutzer schnell darüber zu informieren, was sie bei Anklicken eines Links erwartet. Ausführliche Informationen zu einem Treffer erwarten die User erst im zweiten Schritt der Recherche. Die Ergebnisse der Untersuchungen von Kaczmirek (2003) zeigen, dass der Beschreibungstext des einzelnen Treffers (innerhalb einer Trefferliste), der in der Regel auf ein dahinter liegendes Dokument verweist, am besten einen Umfang von ca. 400 Zeichen haben sollte, um die Entscheidung des Users in Bezug auf sein weiteres Auswahlverhalten (Klickverhalten) optimal zu unterstützen. Als Resümee seiner Untersuchung schlägt Kaczmirek vor, zur Beschreibung der Links in Suchmaschinen redaktionellen, zusammenhängenden Text zu verwenden. Da sich die Bedürfnisse der User jedoch unterscheiden, sollte die Textmenge auf jeden Fall variabel einstellbar sein. Leroy, Xu, Chung, Eggers und Chen (2007) haben untersucht, wie sich dynamische Tools zur Unterstützung der User bei der Benutzung dreier verschiedener Websites mit unterschiedlichen Unterstützungsangeboten auf die Effektivität und Effizienz von Aufgabenlösungen auswirken. Die Autoren stellten fest, dass sich die drei Seiten zwar nicht hinsichtlich der Effektivität unterschieden, sehr wohl

aber im Hinblick auf die Effizienz. Mit dynamischen Tools fiel es den Probanden leichter, ihre Aufgabe zu lösen und zudem waren sie bei der Aufgabenlösung schneller.

Neben den genannten Usability-Prinzipien sollten bei der Untersuchung eines Web-Angebots immer auch folgende Einflüsse berücksichtigt werden:

- *Die Expertise der User*. Für die Einschätzung der Gebrauchstauglichkeit spielt es u. U. eine große Rolle, ob ein Experte oder ein Novize das System benutzt (z. B. ein Thema recherchiert). Bestimmte Unterstützungshilfen (wie z. B. FAQs, Hilfeseiten), die ein Anfänger vielleicht noch braucht, werden von einem Experten möglicherweise ignoriert oder sogar als lästig empfunden.
- *Ziele der User*. Hier geht es um die Frage: Was wollen die User? Je nach motivationaler Lage sind ganz unterschiedliche Ziele mit der Nutzung des Recherche-Angebots verbunden. Während es den einen Usern vielleicht nur darum geht, an kostenlose Dokumente zu gelangen, spielt es für andere dagegen eine größere Rolle, ob sie bei ihrer Recherche vernünftig geleitet und unterstützt werden. Wiederum anderen Nutzern geht es in erster Linie um das Auffinden möglichst aktueller Publikationen. Hier ist der Anbieter gut beraten, eine Zielanalyse der Userwünsche vorzunehmen, sodass eine Priorisierung bei der Gestaltung des Angebots vorgenommen werden kann.
- *Kontext der Nutzung*. Es gibt verschiedene Situationskomponenten, die die Nutzung eines Systems beeinflussen. Dazu zählen beispielsweise: Stress, motivationale Variablen wie die emotionale Befindlichkeit und der Grad des inhaltlichen Interesses (Involvement), Zeitmangel, jegliche Art der Ablenkung (z. B. Umgebungslärm) oder die aktuell fehlende Bereitschaft Geld für einen Artikel oder Download zu bezahlen. Anbieter können zwar keinen Einfluss auf die Nutzungssituation nehmen; dennoch ist es für sie wichtig zu wissen, *unter welchen Umständen das eigene Angebot hauptsächlich genutzt wird*. Regelmäßige Userbefragungen der Zielgruppe sind hier ein Muss.

Zusammenfassend sind nach Schulz (2001b) folgende Aspekte und Kriterien bei der Gestaltung der „Search Engine Usability“ zu beachten:

Tabelle 4. Wichtige Aspekte der „Search Engine Usability“ (nach Schulz, 2001a, vollständige Liste s. ebd.).

Wichtige Aspekte der „Search Engine Usability“	
1. Zweck und Besonderheiten der Suchmaschine	- Der Kunde erhält allgemein verständliche, knappe Aussagen über Zweck und Umfang des Suchtools, ohne danach suchen oder scrollen zu müssen.
2. Die Benutzungs-oberfläche für die Eingabe	- Ein minimalistisches Interface gewährleistet einen geringen kognitiven Aufwand. - Die Default-Suche besteht aus einem einzigen Eingabefeld. - Es wird keine Verwendung von Operatoren und Optionen erwartet.
3. Die Verarbeitung der Eingabe	- Die Suchmaschine verwendet Algorithmen im Hintergrund, die Fehlertoleranz und eine intelligente Aufbereitung der Suchformulierung gewährleisten (Rechtschreibkorrektur, terminologische Hilfen, Stemming-Algorithmen, Relevanz-Feedback, gestuftes Matching-Verfahren).
4. Ausgabe der Ergebnisse	- Die Antworten der Suchmaschine entsprechen den Erwartungen der Kunden (Rückmeldung über Suchformulierung, Antwortzeit, Relevanzranking, Verständlichkeit und Inhalt der Ergebnispräsentation, zweckmäßige Sortierung).
5. Hilfestellungen und Rückmeldungen	- Ein transparentes Interaktionsdesign gibt dem Kunden die Gewissheit, Kontrolle über seine Aktionen zu haben. - Hilfestellungen sind kontextsensitiv, konstruktiv und allgemein verständlich. - Systemrückmeldungen verwenden die Sprache des Kunden (keinen Jargon). - Ein (menschlicher) Ansprechpartner ist schnell und einfach erreichbar und hilft bei Schwierigkeiten individuell weiter.
6. Grundsätzliche Usability-Kriterien	- Sprache, Einhaltung von Standards, Navigation und Ästhetik entsprechen grundsätzlichen Usability-Kriterien.
7. Accessibility	- Ermöglichung eines barrierefreien Zugangs.

3 Das MorphoSaurus-Projekt

Im folgenden Kapitel werden die Projektziele und Fragestellungen des MorphoSaurus-Projekts beschrieben.

3.1 Projektziele

Die Stärken der MEDPILOT-Suchmaschine liegen im Wesentlichen in den qualitativ hochwertigen Inhalten, der leichten Zugänglichkeit sowie in der relativ guten Auffindbarkeit der Inhalte durch die Nutzer. Wie in Kap. 2.3.2 bereits beschrieben, stieß das bisherige MEDPILOT-System jedoch dort an seine Grenzen, wo es um die intelligente Verarbeitung von sprachlich problematischen Nutzeranfragen ging. Durch die Kooperation von ZB MED und Averbis im MorphoSaurus-Projekt wurde deshalb eine innovative Problemlösung im medizinischen Information-Retrieval angestrebt.

Im Mittelpunkt des Projekts stand die Implementierung und Evaluation der Averbis-Technik anhand zweier ausgewählter Datenbanken (CC MED mit ca. 460.000 Titeln und MEDLINE mit ca. 16 Millionen Titeln, Stand Frühjahr 2008). MEDLINE wurde auch deshalb ausgewählt, weil Mediziner bei Recherchen vor allem auf diese Datenbank (via PubMed) zurückgreifen (vgl. auch Hahn, Wermter, DeLuca, Blasczyk, Poprat, Bajwa & Horn, 2007). Auf der Grundlage speziell zusammengestellter Suchterm-Testkollektionen wurde ermittelt, wie gut das bisherige MEDPILOT-System und die neue Averbis-Testsuchmaschine in der Lage sind, verschiedene sprachliche Aspekte zu verarbeiten. Zum einen wurde hier verglichen, wie viele relevante Treffer sich unter den jeweils ersten 20 Ergebnismeldungen (Hits) befanden. Zum anderen wurde zugleich auch die jeweils zurückgemeldete Gesamt-Treffermenge erfasst.

Neben der Implementierung einer Testsuchmaschine auf Basis moderner Suchmaschinentechnologie (Lucene) mit integrierter MorphoSaurus-Technik (Ziel 1) ging es um die Entwicklung einer geeigneten Evaluationsmethodik mit validen Indikatoren zur Messung des Retrieval-Erfolgs der Testsuchmaschine (Ziel 2). Zusätzlich sollte ein Vergleich zwischen dem Testsystem mit MorphoSaurus-Technik und dem bisherigen MEDPILOT-System sowie potenziellen Konkurrenten darüber Auskunft geben, wie die Leistungsfähigkeit des neuen Systems zu bewerten ist (Ziel 3). Das vierte Projektziel bestand in der Verbesserung der

Usability (Ziel 4): Durch eine verbesserte Suchmaschinen-Usability sollte der Zugang zu relevanten Dokumenten und deren Abruf wesentlich erleichtert werden.

Tabelle 5. Ziele des MorphoSaurus-Projekts der ZB MED.

Ziele des MorphoSaurus-Projekts	
1.	Implementierung moderner Suchmaschinentechnologie und der Averbis Morpho-Saurus-Technik (Averbis Core Engine) in eine Testsuchmaschine mit einem großen medizinischen Datenkorpus
2.	Entwicklung valider Evaluationsmethoden
3.	Evaluation der MorphoSaurus-Technologie
	• Vergleich zwischen dem bisherigen MEDPILOT-Metasuchsystem und dem Averbis-Testsystem mit MorphoSaurus-Technologie und Vergleich der Trefferrelevanz-Werte im Hinblick auf spezifische sprachliche Verarbeitungsprobleme
	• Vergleich der Trefferrelevanz zwischen der Averbis-Testsuchmaschine und potenziellen Konkurrenten
4.	Optimierung der Benutzerfreundlichkeit (Usability) der Testsuchmaschine

3.2 Fragestellungen

Die konkreten Fragestellungen des Projekts leiteten sich aus den zuvor genannten Zielen ab. Sie betreffen vornehmlich die in Tabelle 5 genannten Ziele drei und vier. Die Entwicklung und Anwendung angemessener Evaluationsmethoden (Ziel 2) wird in Kapitel 4 (Methoden) ausführlich beschrieben. Die Darstellung der Implementierung der Averbis Core Engine ist nicht Gegenstand dieser Arbeit. Für die programmiertechnischen Hintergründe wird auf die Monografien von Daumke (2007) und Markó (2008) verwiesen.

3.2.1 Verarbeitung sprachlich problematischer Suchanfragen

Eine der wichtigsten Fragestellungen des MorphoSaurus-Projekts war der Vergleich der Retrieval-Effektivität zwischen dem Averbis-Testsystem und dem bisherigen MEDPILOT-System. Die konkrete Fragestellung lautete hier: Welche Unterschiede lassen sich zwischen den beiden Systemen bezüglich der Verarbeitung sprachlich problematischer Suchanfragen feststellen? Dazu wurden folgende Sprachaspekte untersucht:

- *Rechtschreibfehler*. Der Umgang mit Rechtschreibfehlern ist eine Herausforderung für jede Suchmaschine. Kann der Einsatz einer fehlertoleranten Technik eine wesentliche Verbesserung der Trefferrelevanz bewirken?

- *Akronyme*. Abkürzungen sind gerade auf dem Gebiet der Medizin schwierig aufzulösen. So kann ein Akronym ganz verschiedene medizinische Begriffe bezeichnen. Durch ein implementiertes Modul zur Akronymerkennung sollte die Erkennungsleistung der Averbis-Testsuchmachine deutlich verbessert werden.
- *Synonyme*. Wenn die Bedeutung zweier Wörter gleich ist oder sich sehr ähnelt, bezeichnet man diese als „synonym" (Beispiel: ‚Orange' – ‚Apfelsine'). Suchmaschinen mit konventioneller Suchtechnik haben große Schwierigkeiten bei der Verarbeitung von Synonymen. Deshalb erfolgt die Verarbeitung von Synonymen häufig über den Umweg eines integrierten Thesaurus bzw. Schlagwortkatalogs. Die Averbis-Technologie verspricht hier wesentlich bessere Ergebnisse, da hier ein *Thesaurus auf Subwortebene* zum Einsatz kommt, der wesentlich leistungsfähiger ist.
- *Komposita*. Bei einem Kompositum handelt es sich um eine Zusammensetzung zweier selbstständig vorkommender Wörter. Herkömmlichen Suchmaschinen gelingt es nur sehr schlecht, Einzelworte aus den Komposita der Nutzeranfragen zu extrahieren und die Bedeutung korrekt zu analysieren (vgl. z.B. die Trefferanzahlen nach Eingabe von ‚Todesursachenstatistik' vs. ‚Tod Ursachen Statistik'). Durch den Einsatz des Subwort-Lexikons der Averbis Core Engine gelingt eine Extraktion der bedeutungshaltigen Wortteile von Komposita.
- *Übersetzung DEUTSCH – ENGLISCH, ENGLISCH – DEUTSCH*. Oft möchten Mediziner mit einer deutschen Anfrage auch die passenden fremdsprachlichen Treffer angezeigt bekommen. Bisher gelingt es Suchmaschinen kaum, dieses Problem zur Zufriedenheit der Nutzer zu lösen. Dadurch, dass das Averbis-System die Möglichkeit der Identifikation von Fremdwörtern auf der Subwortebene bereitstellt, steigt die Übersetzungsleistung stark an.
- *Laiensprache – Expertensprache*. In Abhängigkeit vom Umfang des Domänenwissens des Nutzers wird eine Suchmaschine eher mit fach- oder laiensprachlichen Suchtermen abgefragt. Trotzdem erwarten die Nutzer ähnlich qualifizierte Treffer. Auch in diesem Punkt wird durch den Einsatz des Averbis-Systems eine Steigerung der semantischen Identifikationsleistung erwartet, sodass es letztlich unerheblich ist, ob jemand mit dem Wort ‚Nasennebenhöhlenentzündung' oder mit dem Wort ‚Sinusitis' sucht.
- *Grammatikalische Variationen*. In der Regel führen z.B. Suchworte, die im Singular eingegeben werden, zu anderen Treffern und Trefferzahlen als Suchworte, die im Plural

formuliert werden. Hier sollte die MorphoSaurus-Technik Abhilfe schaffen, sodass unabhängig vom Numerus ähnlich relevante und ähnlich viele Treffer gefunden werden. Andere Variationen sind z.B. die Nominalisierung (‚überforderte Angehörige' vs. ‚Überforderung Angehöriger') oder die Bildung eines Genitivattributes (‚Lagerung Neugeborene' vs. ‚Lagerung Neugeborener').

Insgesamt wurde bei sämtlichen sprachlich problematischen Aspekten eine bessere Retrieval-Leistung der Averbis-Testsuchmaschine im Vergleich zum bisherigen MEDPILOT-System erwartet.

3.2.2 Konkurrenzanalyse und Benchmarking

Die zweite Fragestellung beschäftigte sich mit der Retrieval-Effektivität des Averbis-Testsystems im Vergleich zu den potenziellen Konkurrenten. Als Testkandidaten für das Benchmarking wurden folgende Anbieter ausgewählt: PubMed, Scirus, Google, Google Scholar sowie GoPubMed. Unter Einsatz einer repräsentativen Testkollektion von Suchtermen sollte die jeweilige Retrieval-Effektivität ermittelt werden, sodass folgende Fragen beantwortet werden können:

- Welche Suchmaschine weist die höchsten Werte in der Trefferrelevanz auf?
- Welche Suchmaschine weist die höchsten Trefferzahlen auf?
- Welche Suchmaschine weist die geringste Zahl an Null-Treffer-Meldungen auf?

Zur Einschätzung der Analysefähigkeiten der Konkurrenten wurde auch ein Vergleich der Menge der Null-Treffer-Meldungen herangezogen. Diese Meldungen geben u.a. darüber Aufschluss, ob die implementierten Algorithmen in der Lage sind, auch defizitäre und problematische Anfragen der Nutzer so zu analysieren, dass einerseits möglichst viele relevante Treffer zurückgemeldet werden und gleichzeitig die Zahl der Null-Treffer-Meldungen möglichst gering ausfällt.

3.2.3 Fragestellungen zur Usability der Averbis-Testsuchmaschine

Die Benutzerfreundlichkeit der Testsuchmaschine wurde in Anlehnung an die Erkenntnisse der Usability- und Nutzerforschung verbessert. Ob die Maßnahmen zur optimierten Nutzerführung auch tatsächlich greifen, sollte im Rahmen eines szenariobasierten Usability-Tests

untersucht werden sowie unter Einsatz eines Fragebogens. Hier ergaben sich folgende Fragestellungen:

1. *Wie effektiv und effizient gelingt die Lösung von Aufgaben*? Anhand von spezifischen Rechercheaufgaben sollte erstens herausgefunden werden, ob diese Aufgaben durch die Mediziner mithilfe des neuen Testsystems überhaupt gelöst werden können (Effektivität). Zweitens wurde untersucht, wie lange die Testpersonen zur Lösung dieser Aufgaben benötigten (Effizienz). Von Interesse war auch, ob sich hier Unterschiede zwischen inhaltlichen Experten (Ärzten) und Studierenden beobachten lassen. Ebenso sollte die Frage überprüft werden, ob interneterfahrene Nutzer die Lösung von Rechercheaufgaben besser und schneller bewältigen als unerfahrene Nutzer.

2. *Wie werden die realisierten Unterstützungshilfen angenommen?* Im Rahmen des MorphoSaurus-Projekts wurden verschiedene Unterstützungshilfen realisiert, deren Erfolg im Rahmen eines Usability-Test überprüft werden sollte.

- *Hilfen auf Suchschlitzebene.* Es wurde eine *Auto-Vorschlagsfunktion* realisiert, die passende MeSH- und ICD-10-Begriffe während der Eingabe des Suchterms durch die Nutzer einblendete.
- *Hilfen zur Einschränkung des Suchraums.* Als Mittel zur Eingrenzung des Suchraums wurde ein *Schieberegler* entwickelt. Ist er ein effektives Instrument, um relevantere Treffer zu erhalten? Wie viele Testpersonen bevorzugen den Schieberegler anstelle von herkömmlichen Optionsschaltern?
- *Facettensuche und Verfeinerung der Suche.* Durch die Einblendung von *verwandten Suchbegriffen* nach Abschicken des Suchterms wird den Nutzern die Möglichkeit gegeben, ihre Suche zu verfeinern. Wie werden diese Filtermöglichkeiten zur Navigation in der Ergebnismenge bewertet?
- *Präsentation der Treffer(liste)*. Wie zufrieden sind die Nutzer mit dem *Trefferformat* und dem *Highlighting* der Suchworte?
- *Selbstbeschreibungsfähigkeit*. Ist die Selbstbeschreibungsfähigkeit der Testsuchmaschine ausreichend hoch, sodass die Nutzer *auch beim ersten Besuch direkt erkennen, welche Inhalte die Suchmaschine anbietet und wer diesen Dienst anbietet*? Wie bewerten die Nutzer bestimmte *Funktionen und Funktionsbezeichnungen*? Hier ging es auch um die Bewertung von *Optionsschaltern und Sortiermöglichkeiten*.

4 Methoden

Im folgenden Kapitel werden die im MorphoSaurus-Projekt eingesetzten Methoden vorgestellt. Verschiedene Autoren haben schlüssig belegt, dass die Erforschung der Optimierung eines Suchmaschinenangebots eines *multimethodalen Ansatzes* bedarf (vgl. Kap. 2.2). Eine ausschließliche Analyse der Retrieval-Leistung ist nicht ausreichend. Daher wurde im MorphoSaurus-Projekt eine methodische Herangehensweise gewählt, die neben der Messung der Retrieval-Leistung auch die Usability der Testsuchmaschine sowie eine Analyse der MEDPILOT-Logfiles berücksichtigt.

Nachdem die Testsuchmaschine auf der Grundlage einer Lucene-Umgebung aufgesetzt wurde, hat Averbis die Datenbank MEDLINE und zwei der wichtigsten ZB MED-Datenbanken (CC MED, ZB MED OPAC) in die Testumgebung implementiert. Anschließend wurden die von Averbis entwickelten Algorithmen zur Analyse der Nutzeranfragen integriert.

Vor Beginn der Evaluation mussten aufseiten des Evaluationsteams jedoch noch einige Vorarbeiten geleistet werden. Dazu gehörte die Durchführung einer Analyse des MEDPILOT-Logfiles und darauf aufbauend die Inhaltsanalyse der Suchterme. Das Ziel war die Gewinnung von Erkenntnissen zum Nutzerverhalten (vgl. Kap. 2.5), die Analyse der von den MEDPILOT-Nutzern gesuchten medizinischen Inhalte sowie die Konstruktion valider Testkollektionen zur Überprüfung der Retrieval-Leistung.

4.1 Vorbereitende Untersuchungsschritte

Vor der eigentlichen Evaluation wurden einige vorbereitende Untersuchungen durchgeführt, um folgende Fragen zu klären:

- Welche Komplexität weisen die Suchanfragen auf?
- Welche Inhalte suchen die Nutzer?
  - Wie hoch ist der prozentuale Anteil an originären Chemieanfragen?
- Wie lassen sich valide Testkollektionen aus dem Logfile entwickeln?

4.1.1 Analyse des MEDPILOT-Logfiles und Inhaltsanalyse

Um Rückschlüsse darüber ziehen zu können, ob sich der Grad der Komplexität der Suchanfragen von MEDPILOT-Nutzern (als Angehörige von Gesundheitsberufen) von dem

Komplexitätsgrad der Anfragen anderer Nutzer unterscheidet, wurde zunächst das MEDPILOT-Logfile analysiert. Für den Zeitraum von über sieben Monaten wurden sämtliche Suchterme aus dem Logfile extrahiert. Insgesamt umfasste das Logfile für den Zeitraum von November 2006 bis Juni 2007 142.922 Suchanfragen bzw. Queries. Die Datenbasis für die Inhaltsanalyse bildete eine Zufallsstichprobe von 10.000 Suchtermen. Das sind ca. sieben Prozent aller Suchanfragen aus dem untersuchten Zeitraum. Anschließend wurde ein an die Fragestellung angepasstes Kategoriensystem erstellt, welches insgesamt 24 Inhaltskategorien umfasste. Mithilfe dieses Kategoriensystems wurde neben den medizinischen Inhaltsklassen auch das Vorkommen von z. B. Booleschen Operatoren, Feldsuchen oder Rechtschreibfehlern untersucht.

Im nächsten Schritt erfolgte dann die Zuordnung der Queries zu den jeweiligen Inhalts- bzw. Formalbereichen. Vorab wurde die Reliabilität des Kategoriensystems anhand eines 100 Suchterme umfassenden Teilsamples überprüft, welches ebenfalls durch Randomisierung zusammengestellt wurde. Für die Berechnung der Reliabilität wurde die intersubjektive Übereinstimmung der Einschätzungen zweier Experten bei der Zuweisung der Suchterme zu den Kategorien herangezogen (sog. Interrater-Übereinstimmung). Die Interrater-Reliabilität – als Maß für die Höhe der Übereinstimmung – lag bei 86%. D. h., die Experten waren sich (unabhängig voneinander) in 86 von 100 Fällen darin einig, welcher Inhaltskategorie ein Suchterm zuzuordnen war. Dies stellt einen akzeptablen Wert dar, sodass anschließend sämtliche 10.000 Suchanfragen nur noch von einer Person den Inhaltskategorien zugeordnet wurden. Dabei handelte es sich um eine erfahrene Biologin und Germanistin mit umfangreichem medizinischem Wissen. Einzelheiten zur Methode der Inhaltsanalyse finden sich etwa bei Rustemeyer (1992) oder Groeben und Rustemeyer (2001).

4.1.1.1 Welche Komplexität weisen die Suchanfragen auf?

Es wurde bereits darauf hingewiesen, dass Suchmaschinennutzer versuchen, ihre ‚kognitive Last' möglichst gering zu halten und lediglich ein oder zwei Suchterme verwenden (z. B. Jansen, Spink & Saracevic, 2000, S. 207ff). Deshalb wurde auch für die MEDPILOT-Anfragen angenommen, dass der Großteil der Nutzer lediglich kurze Suchen durchführt, die vornehmlich aus Ein- oder Zwei-Wort-Anfragen bestehen. Darüber hinaus weiß man, dass nur wenige Nutzer die Möglichkeiten von Booleschen Operatoren, Trunkierungen oder Feldsuchen nutzen (im Überblick Spink & Jansen, 2004). Ob sich die geringe Nutzung von

Operatoren auch bei einer wissenschaftlichen Suchmaschine wie MEDPILOT zeigt, sollte anhand von Analysen der Suchanfragen überprüft werden. Aufgrund der bisherigen Forschungslage (vgl. Kap. 2.5.2) war von der Annahme auszugehen, dass sich die Zielgruppe der Mediziner nicht grundlegend vom Verhalten der sonstigen Nutzer unterscheidet. Es wurde also ein geringer bis sehr geringer Gebrauch von Booleschen Operatoren und Feldsuchen erwartet.

4.1.1.2 Welche Inhalte suchen die Nutzer?

Die Kenntnis darüber, welche Inhalte von den Nutzern besonders nachgefragt werden, lässt sich für die Unterstützung der Nutzer während des Suchprozesses verwerten. Darüber hinaus stellt das Wissen um das Ausmaß der Nachfrage nach bestimmten Inhalten eine fundierte Grundlage zur Evaluation und Planung des Quellenangebots dar.

Zunächst interessierte das Evaluationsteam besonders die Frage, wie hoch der *Anteil an originären Chemieanfragen* im Logfile ist – gemessen an der Gesamtheit der Suchanfragen. Die Beantwortung dieser Frage war sehr wichtig, da die MorphoSaurus-Technologie mit einem Morphemthesaurus arbeitet, der speziell für den Bereich der Medizin entwickelt wurde. Für den Fall, dass die Nutzer in MEDPILOT verhältnismäßig häufig originäre Chemieanfragen stellen, wurde in Erwägung gezogen, einen eigenen Morphemthesaurus für diesen Bereich zu entwickeln. Ab einem Anteil von ca. 20% (an allen Queries) würde sich die Entwicklung eines solchen, auf den Bereich der Chemie und Handelsnamen spezialisierten Thesaurus, lohnen.

4.1.2 Erstellung der Testkollektionen

Bevor die konkrete Evaluation der Retrieval-Effektivität vorgenommen werden konnte, mussten eine Reihe von speziellen Testkollektionen entwickelt werden, mit denen die Leistungsfähigkeit der Testsuchmaschine hinsichtlich der Verarbeitung sprachlich problematischer Suchanfragen überprüft werden konnte. Um an valide Suchterme für die Überprüfung der Retrieval-Leistung zu gelangen, hat sich das Evaluationsteam für die Konstruktion der Testkollektionen an den Inhalten des Logfiles bzw. an den Inhalten der tatsächlich genutzten MEDPILOT-Suchterme orientiert.

Da die Suchterme schon während der Inhaltsanalyse des Logfiles über die Vergabe klassifizierende Tags (Schlagworte) bestimmten sprachlichen Problemklassen zugeordnet

wurden, gelang es relativ einfach, die Testkollektionen zu erstellen. Mithilfe von sogenannten *regulären Ausdrücken* (‚regular expressions')[23] wurden aus dem getaggten Gesamtlogfile diejenigen Suchterme herausgefiltert, die für eine bestimmte Klasse von sprachlich problematischen Anfragen stehen. Der nächste Schritt war eine Zufallsauswahl von jeweils 50 Suchanfragen aus den jeweiligen Problembereichen der Sprachverarbeitung. Zudem wurde für das Benchmarking eine *repräsentative Testkollektion* erzeugt, die ebenfalls auf einer Zufallsstichprobe gründete, aber dennoch sämtliche sprachliche Problemaspekte abbildete. Insgesamt wurden folgende Testkollektionen erstellt:

Tabelle 6. Verwendete Testkollektionen zur Evaluation der Verarbeitung problematischer Sprachaspekte.

Verwendete Testkollektionen	
Rechtschreibung	- 50 originale Suchterme mit Rechtschreibfehlern - keine modifizierten Suchterme
Akronyme	- 50 originale Suchterme mit Akronymen - keine modifizierten Suchterme
Synonyme	- 50 originale Suchterme (Synonyme) + - 50 modifizierte Suchterme
Komposita	- 50 originale Suchterme mit Komposita + - 50 modifizierte Suchterme (Komposita zerlegt)
Übersetzung	- 50 deutsche Suchterme (Übersetzung: de – en) (original) + - 50 englische Suchterme (Übersetzung: en – de) (modifiziert)
Sprachniveau	- 50 originale Suchterme mit Laiensuchworten + - 50 modifizierte Suchterme mit analogen Expertensuchworten
Grammatik	- 50 originale Suchterme + - 50 modifizierte Suchterme (z. B. Singular vs. Plural)
Repräsentative Testkollektion	- 50 originale Suchterme zufällig aus Logfile + - 50 modifizierte Suchterme

Zu einigen Testkollektionen wurden modifizierte Varianten erzeugt. Dies geschah z. B. im Rahmen der Überprüfung der Fähigkeit zur Zerlegung von Komposita. Zu der Komposita-Testkollektion – bestehend aus 50 Original-Suchtermen des Logfiles – wurde zusätzlich eine Alternativ-Testkollektion mit ebenfalls 50 Suchtermen erzeugt. Dafür wurden die Komposita der Original-Anfragen entsprechend zerlegt und modifiziert (z. B. original: ‚Mutterschutz-

[23] Hierbei handelt es sich um Suchformulierungen auf der Basis einer Programmiersprache mithilfe deren Syntax eine Textdatei nach bestimmten Zeichen- bzw. Buchstabenkombinationen durchsucht und ausgezählt werden kann (vgl. Friedel, 2007).

gesetz‘ vs. modifiziert: ‚Mutter Schutz Gesetz‘). Die Überprüfung der Trefferrelevanz wurde anschließend mit den Suchanfragen der Testkollektionen am Beispiel zweier Datenbanken durchgeführt: MEDLINE und CC MED (Current Contents of Medicine). Bei CC MED handelt es sich um eine ZB MED-Datenbank, die über ca 450.000 Einträge aus dem Bereich der Medizin verfügt. Gesammelt wird hier schwerpunktmäßig Literatur (Zeitschriftenbeiträge) aus dem deutschen Sprachraum. Die Datenbank MEDLINE wird herausgegeben von der NLM (National Library of Medicine) in den USA. Diese Sammlung enthält Nachweise der internationalen Literatur aus der Medizin (einschließlich der Zahn- und Veterinärmedizin), der Psychologie und dem öffentlichen Gesundheitswesen. Sie umfasste 2008 etwa 16 Mio. Einträge. Zu den Quellen gehören ca. 4.800 internationale Zeitschriften. Die Nachweise sind vorwiegend in englischer Sprache abgefasst. Durch einen Vergleich der Retrieval-Ergebnisse für die spezifischen Sprachaspekte konnte ermittelt werden, wie leistungsfähig die Averbis-Algorithmen (der Testsuchmaschine) gegenüber dem Ansatz der bisherigen MEDPILOT-Suche sind.

4.2 Evaluation der Retrieval-Effektivität

4.2.1 Trefferrelevanz

Nach Abschicken einer Suchanfrage einer bestimmten Testkollektion wurde die Gesamtzahl der zurückgemeldeten Treffer notiert sowie für jeden Treffer unter den ersten 20 vermerkt, ob dieser für die spezielle Suchanfrage als relevant einzustufen war oder nicht. Diese Einschätzung wurde von Experten vorgenommen. Bei Unsicherheiten in der Zuordnung wurde versucht, eine intersubjektive Übereinstimmung zwischen mindestens zwei Experten herbeizuführen. Bei Uneinigkeit wurde eine dritte Person für die Einschätzung hinzugezogen. Im Anschluss an die Relevanzbeurteilung wurden die Precision-Werte für die Cut-Off-Grenzen von 5, 10, 15 und 20 (Treffer) berechnet. Waren z. B. 10 der ersten 20 Treffer relevant, bedeutet dies ein Precision-Wert von 0,5 (bzw. 50 %) für den Cut-Off-Wert von 20. Bei einem Cut-Off-Wert von 10 bedeuten 10 Treffer, dass sämtliche 10 Treffer relevant sind (Precision = 1,0 bzw. 100 % Trefferrelevanz).

Durch diese Vorgehensweise lässt sich natürlich nicht der tatsächliche Precision-Wert berechnen. Dieser ist aufgrund der großen Trefferzahlen kaum oder gar nicht zu bestimmen. Hierzu wäre es nötig zu überprüfen, wie viele der jeweils von der Suchmaschine zurück-

gemeldeten Treffer – jenseits der ersten 20 Treffer – tatsächlich relevant sind. Darüber hinaus wäre es für eine genaue Berechnung des Recall außerdem nötig zu wissen, wie sich der Anteil der gefundenen relevanten Dokumente zur Gesamtzahl der in der Datenbank enthaltenen relevanten Dokumente (zur jeweiligen Suchanfrage) verhält (vgl. dazu Kap.2.1).

Ein eindeutiger Vorteil der hier vorgenommenen vereinfachten Berechnung liegt aber in der hohen ökologischen Validität des Vorgehens. Die Nutzer von Suchmaschinen schauen sich tendenziell nur die erste Trefferseite an. Da bei den meisten Suchmaschinen lediglich 10 Treffer für die Ergebnisrückmeldung voreingestellt sind, spiegelt die Beschränkung auf die Analyse der ersten 20 Treffer (als Cut-Off-Wert) eine ausreichend valide Berücksichtigung des Userverhaltens wider (vgl. z. B. Lewandowski, 2006).

Anschließend wurde tabellarisch festgehalten, wie viele *relevante* Treffer sich unter den ersten 5, 10, 15 und 20 Treffern befanden, die als Antwort auf die Suchanfrage zurückgemeldet wurden. Dabei war für eine spätere Auswertung Folgendes zu beachten: Drei Treffer bei einem Cut-Off von p5 (also bei der Betrachtung von höchstens fünf Treffern) bedeuten eine Precision von 0,6. D. h., 60 % dieser fünf Treffer sind relevant. Jedoch bedeuten drei Treffer bei einem Cut-Off-Wert von 20, dass hier lediglich ein p-Wert von 0,15 erreicht wird. Das heißt, nur 15 % von 20 Treffern sind relevant. Bei der Einschätzung der Aussagekraft von Precision-Werten ist also stets darauf zu achten, für welchen Cut-Off-Wert diese Angabe gemacht wird. Die folgende Tabelle verdeutlicht diesen Zusammenhang:

Tabelle 7. Trefferzahlen und Precision bei verschiedenen Cut-Off-Werten (p5, p10, p15 und p20).

p5			p10			p15			p20		
Treffer	p	%	Treffer	p	%	Treffer	p	%	Treffer	p	%
1	0,2	20	1	0,1	10	1	0,07	6,7	1	0,05	5
2	0,4	40	2	0,2	20	2	0,13	13,3	2	0,1	10
3	0,6	60	3	0,3	30	3	0,20	20,0	3	0,15	15
4	0,8	80	4	0,4	40	4	0,27	26,7	4	0,2	20
5	1	100	5	0,5	50	5	0,33	33,3	5	0,25	25
			6	0,6	60	6	0,40	40,0	6	0,3	30
			7	0,7	70	7	0,47	46,7	7	0,35	35
			8	0,8	80	8	0,53	53,3	8	0,4	40
			9	0,9	90	9	0,60	60,0	9	0,45	45
			10	1	100	10	0,67	66,7	10	0,5	50
						11	0,73	73,3	11	0,55	55
						12	0,80	80,0	12	0,6	60
						13	0,87	86,7	13	0,65	65
						14	0,93	93,3	14	0,7	70
						15	1,00	100	15	0,75	75
									16	0,8	80
									17	0,85	85
									18	0,9	90
									19	0,95	95
									20	1	100

p = Precision

Für den Vergleich zwischen der Averbis-Testsuchmachine und dem bisherigen MED-PILOT-Systems sowie für die Konkurrenzanalyse wurden die Trefferrelevanzwerte der Testsuchanfragen jeweils gemittelt.

4.2.2 Durchschnittliche Trefferzahl

Während die Erfassung der Precision bzw. der Trefferrelevanz ein eher qualitatives Urteil zur Bewertung der Retrieval-Effektivität erlaubt, kann die Erhebung der mit einer Suchanfrage erzielten Trefferanzahl Auskunft über die quantitative Stärke einer Suchmaschine geben. Ob aber bei einer Suche 30, 100 oder 10.000 Treffer gefunden werden, interessiert den Nutzer aber nicht wirklich. Aus den Ergebnissen zum Suchverhalten bei Suchmaschinenusern (vgl. Kap. 2.5.2) wissen wir, dass nur die ersten 10 Treffer genutzt werden und unter diesen insbesondere die ersten drei bis fünf. Insofern sagen Trefferzahlen eher etwas darüber aus, wie gut die Suchmaschinenalgorithmen in der Lage sind mit bestimmten Suchtermen adäquate Dokumente aus dem Index zu filtern.

Dabei können manche Suchterme sehr große Trefferzahlen produzieren wie z. B. das Einzelwort ‚Aspirin'. Andere Suchterme, die z. B. Phrasen enthalten, seltene Terme, viele Worte oder falsch geschriebene Worte etc., erzielen eher weniger Treffer (z. B. ‚wirelsäule' statt ‚wirbelsäule'). Mit ebenfalls wenigen Treffern ist zu rechnen, wenn Akronyme nicht aufgelöst werden können (z. B. ‚tha' vs. ‚total hip arthroplasty').

Die unterschiedlichen Testkollektionen zu den verschiedenen Problembereichen der Sprachverarbeitung sind auch deshalb entwickelt worden, um hier eine methodische Vergleichbarkeit hinsichtlich der Trefferzahlen zu erreichen.

Für die angestrebten Vergleiche auf Grundlage *der repräsentativen Testkollektion* musste ein modifiziertes Vorgehen gewählt werden: Da die 50 bzw. 100 Suchterme sämtliche der zu untersuchenden Sprachphänomene abdecken sollten, war damit zu rechnen, dass die Trefferanzahl sehr unterschiedlich ausfallen würde. Um die Aussagekraft des Mittelwerts als Maß der zentralen Tendenz nicht zu entwerten, wurden Ausreißer (sog. Extremwerte) in den Trefferzahlen beseitigt. In der statistischen Literatur werden verschiedene Verfahren zum Umgang mit Ausreißerdaten beschrieben (vgl. Barnett & Lewis, 1978). Das hier angewendete Verfahren ist die „Windsorized-Technik", bei der alle Fälle von der weiteren Untersuchung

ausgeschlossen werden, deren Ergebnisse mehr als zwei Standardabweichungen vom Mittelwert entfernt liegen.

4.2.3 Null-Treffer-Meldungen

Darüber hinaus ist berechnet worden, wie viele der Suchanfragen zu Null-Treffer-Meldungen führten. Dies ist insbesondere für eine Modifikation und Verbesserung der Suchmaschinenparameter von Interesse. Suchmaschinennutzer wollen in der Regel keine Null-Treffer-Meldungen. Zum einen kann eine solche Meldung ein Hinweis auf Schwierigkeiten der Suchmaschine mit der Auflösung des Suchterms bedeuten und zum anderen kann es sein, dass die Suchanfrage fehlerhaft ist. Schließlich besteht noch die Möglichkeit, dass die Suchmaschine tatsächlich keinen Treffer zu der Suchanfrage liefern kann, da z. B. keine Literatur zu einem bestimmten Stichwort (bzw. zu der Kombination von Stichwörtern) im Index existiert.

4.2.4 Konkurrenzanalyse und Benchmarking

Eine weitere wichtige Fragestellung für die Evaluation war der Vergleich bzw. das Benchmarking der Retrieval-Leistung der Averbis-Testsuchmaschine (mit und ohne MorphoSaurus-Technologie) mit potenziellen Konkurrenten des MEDPILOT-Angebots.

In diesen Vergleich gingen die Ergebnisse von jeweils 100 Suchanfragen einer repräsentativen Testkollektion für den Bereich Medizin ein. Dabei wurden zwei Testszenarien durchgeführt:

A. Um einen validen Test der Retrieval-Leistung der verschiedenen Suchmaschinen zu ermöglichen, war es nötig, die Retrieval-Tests auf der Grundlage einer vergleichbaren Datenbasis durchzuführen. Deshalb beschränkte das Evaluationsteam den Vergleich in Testszenario A auf die Datenbank MEDLINE.
B. Für die Leistungsfähigkeit einer Suchmaschine spielt die Indexgröße eine wesentliche Rolle. Um eine Abschätzung des Einflusses der Indexgröße auf die Ergebnisse Trefferrelevanztests zu ermöglichen, wurde ein zweiter Vergleich durchgeführt. Hierbei wurden die Einstellungen der Suchmaschinen so gewählt, dass sie auf den maximalen Umfang ihres Index zurückgreifen konnten: also auf MEDLINE *und* alle anderen verfügbaren Quellen.

Tabelle 8 gibt eine Übersicht über die durchgeführten Testszenarien. Untersucht wurden folgende Suchmaschinen: Averbis-Testsuchmaschine (ohne und mit MorphoSaurus-Technik), PubMed, Scirus, Google, Google Scholar sowie GoPubMed.

Tabelle 8. Testszenarien im Benchmarking: Vergleich zwischen der Averbis-Testsuchmaschine mit potenziellen Konkurrenten. Durchführung der Tests: 2008.

	Testszenario A	Testszenario B
Suchmaschinen	**Vergleichsbasis: nur MEDLINE** Ermittlung der Trefferrelevanz mit je 100 Suchanfragen	**Vergleichsbasis: alle verfügbaren Daten(banken) der jeweiligen Suchmaschine (in Klammern: ungefähre Indexgröße)** Ermittlung der Trefferrelevanz mit je 100 Suchanfragen
(1) Averbis-Testsuchmaschine (ohne MorphoSaurus-Technik)	Nur MEDLINE (ca. 16 Mio.)	MEDLINE, CC MED, OPAC (ca. 17 Mio.)
	• 100 Suchanfragen	• 100 Suchanfragen
(2) Averbis-Testsuchmaschine (mit MorphoSaurus-Technik)	Nur MEDLINE (ca. 16 Mio.)	MEDLINE, CC MED, OPAC (ca. 17 Mio.)
	• 100 Suchanfragen	• 100 Suchanfragen
(3) MEDPILOT	Nur MEDLINE (ca. 16 Mio.)	
	• 100 Suchanfragen	
(4) PubMed	Nur MEDLINE (ca. 16 Mio.)	MEDLINE (ca. 16 Mio.)
	• 100 Suchanfragen	• 100 Suchanfragen
(5) Scirus	Nur MEDLINE (ca. 16 Mio.)	Alle Daten (ca. 480 Mio.)
	• 100 Suchanfragen	• 100 Suchanfragen
(6) Google	Nur MEDLINE (ca. 16 Mio.)	Alle Daten (ca. eine Billionen)
	• 100 Suchanfragen	• 100 Suchanfragen
(7) Google Scholar	-	Alle Daten (ca. zwei Mrd.)
	-	• 100 Suchanfragen
(8) GoPubMed	Nur MEDLINE (ca. 16 Mio.)	MEDLINE (ca. 16 Mio.)
	• 100 Suchanfragen	• 100 Suchanfragen

PubMed[24]. Bei PubMed handelt es sich um einen direkten Konkurrenten von MEDPILOT, der von vielen Medizinern als primäre Quelle für medizinische Informationsrecherchen in der Datenbank MEDLINE benutzt wird. Sie ist die umfangreichste medizinische Datenbank weltweit mit einem Bestand von ca. 16 Mio. Literaturnachweisen aus rund 4800 Zeitschriften (Stand 2008). Die Datenbank wird seit 1949 über die National Library of Medicine (NLM, USA) angeboten. Inzwischen gibt es eine Reihe von kostenlosen Zugängen zu dieser

[24] http://www.ncbi.nlm.nih.gov/pubmed/

Datenbank über verschiedene Anbieter, die mit mehr oder minder ausgefeilten Benutzeroberflächen und modifizierten Such- und Abfragemöglichkeiten werben.

Scirus[25]. Die Suchmaschine Scirus ist ein Angebot des Elsevier-Verlags, das seit 2001 die Suche nach wissenschaftlichen, technischen und medizinischen Publikationen ermöglicht. Auch hier lässt sich die Suche auf die MEDLINE-Quellen einschränken, wodurch eine Vergleichbarkeit mit PubMed und der Averbis-Testsuchmaschine erreicht wird. Der gesamte Index umfasst laut Angaben der Anbieter ca. 480 Mio. Dokumente.

Google[26]. Google wurde ausgewählt, weil sich viele Mediziner für ihre Recherchen lediglich auf diese Suchmaschine verlassen. Für viele Nutzer ist die Verwendung von Google geradezu synonym mit dem Vorgang der Internetrecherche. Durch die Möglichkeit, die Suchabfrage auf bestimmte Server zu beschränken, können gezielt MEDLINE-Daten recherchiert werden. Als globale Suchmaschine, die sich an keine spezifische Nutzergruppe wendet, besitzt sie den größten Index unter den großen Suchmaschinen. Es wird geschätzt, dass der Index ca. eine Billion URLs umfasst (Alpert & Hajaj, 2008). Genaue Zahlen sind auch hier schwer zu ermitteln. Eine Billion URLs bedeuten aber nicht unbedingt eine Billion verschiedener Internet-Seiten, da viele Seiten z. B. mehrere URLs mit dem exakt identischen Inhalt besitzen.

Google Scholar[27]. Dieses spezialisierte Angebot von Google im Wissenschaftsbereich wurde deshalb ausgewählt, weil es dezidiert auch die von MEDPILOT und PubMed angebotenen MEDLINE-Quellen enthält und sich so für eine Vergleichbarkeit sehr gut eignet, obwohl es sich nicht um ein speziell medizinisches Suchmaschinenangebot handelt. Der Dienst startete im Dezember 2004 und hat sich seitdem zu einem sehr gefragten Instrument der Recherche nach wissenschaftlicher Literatur entwickelt. Google Scholar bietet Literatur aus dem gesamten Spektrum der Wissenschaften an, auch aus dem Bereich der Medizin. Analysen und Tests haben inzwischen gezeigt, dass Google Scholar in den letzten fünf Jahren sein Angebot erheblich erweitert hat. Es wird geschätzt, dass der Index ca. zwei Mrd. Datensätze enthält. Neben den Vorteilen einer schnellen und kostenlosen Recherche bringt die Nutzung von Google Scholar auch einige Nachteile mit sich. So suchen die von Google

[25] http://www.scirus.com/

[26] http://www.google.de/

[27] http://scholar.google.de/

Scholar eingesetzten Web-Crawler nur einen Teil der bekannten medizinischen Datenbanken ab. MEDLINE-Artikel werden z. B. nur auszugsweise erfasst. Nach unseren Prüfungen enthält Google Scholar nur ca. zwei Drittel des zugänglichen MEDLINE-Materials. Über die Gründe für dieses Defizit schweigt sich Google aus. Undurchsichtig ist auch der Aktualisierungsrhythmus, mit dem der Index in den einzelnen Wissenschaftsbereichen auf den neusten Stand gebracht wird. Zu den Nachteilen von Google Scholar in der wissenschaftlichen Literaturrecherche informiert z. B. Lewandowski (2007b). Für einen Überblick, insbesondere für den Vergleich von Recherchen in PubMed und Google Scholar, sei hier auf den Artikel von Shultz (2007) verwiesen. Grundsätzlich ist ein direkter Vergleich von PubMed vs. Google Scholar problematisch, da z. B. über PubMed in MEDLINE mit sehr differenzierten Funktionen nach medizinischer Literatur gesucht werden kann. In Google Scholar ist dies nur sehr begrenzt möglich.

Google Scholar bietet jedoch auch eine Reihe von Vorteilen, die über die Möglichkeiten der PubMed-Suche hinausgehen (vgl. Giustini & Barsky, 2005), wie

- eine relevanzbasierte Suche (ein Suchbegriff führt zu den Literaturstellen, die am häufigsten zitiert werden),
- die Suche nach Buchinhalten,
- die Suche nach Abstractsammlungen großer Kongresse sowie
- die Erfassung der Inhalte von Doktor- oder Diplomarbeiten.

GoPubMed[28]. Bei GoPubMed handelt es sich um ein Suchmaschinenangebot der Firma *Transinsight*. Über die Suchmaschine kann im Datenbestand von MEDLINE recherchiert werden. Also war auch hier von einem Datenbestand von ca. 16 Mio. indexierten Dokumenten auszugehen. Laut Angaben dieser Firma baut GoPubMed auf semantischen Suchtechnologien auf und verspricht – in Verbindung mit Methoden des Textminings – ähnliche Problemlösungen wie die Averbis-Technologie. Daher wurde die Leistungsfähigkeit von GoPubMed im Rahmen der Konkurrenzanalyse ebenfalls untersucht.

[28] http://gopubmed.org/

4.3 Optimierung der Usability

Wie bereits in Kapitel 2 ausführlich beschrieben, ist für die Akzeptanz eines Systems neben den Systemeigenschaften auch die Benutzerfreundlichkeit von großer Bedeutung. Die Unterstützung der Nutzer durch eine optimierte Usability kann wesentlich dazu beitragen, dass Recherchen erfolgreich verlaufen. Eine hohe Usability wird dann erlebt, wenn die Nutzer im Verlauf ihrer Recherche:

1. tatsächlich an ihr Ziel gelangen (passende bzw. relevante Literatur finden) und
2. dieses Ziel möglichst zügig erreichen.

Das setzt voraus, dass die Funktionselemente einer Website in hohem Maße selbsterklärend sind. Durch die Selbsterklärungsfähigkeit der Website werden vor allem zwei Dinge befördert: das Verständnis für die Navigationsstrukturen sowie das Verständnis für die Benennung von Funktionselementen.

Im MorphoSaurus-Projekt war es – neben der Verbesserung der Trefferergebnisse – deshalb von großer Wichtigkeit, die Benutzerführung so zu optimieren, dass die Nutzer die Tools und Hilfestellungen auch wirklich als Unterstützung in ihrem Rechercheprozess erleben konnten. Das Evaluationsteam wählte zur Erreichung dieses Ziels ein dreistufiges Vorgehen: 1. Entwicklung von Unterstützungsfunktionen, 2. Entwicklung und Durchführung eines Usability-Tests sowie 3. Entwicklung und Einsatz eines Fragebogens.

4.3.1 Entwicklung von Unterstützungsfunktionen

In Anlehnung an die Erkenntnisse der bisherigen Suchmaschinenforschung (vgl. Kap. 2.5) und als Ergebnis der Rückmeldungen durch das Evaluationsteam, wurden durch Averbis verschiedene Unterstützungsfunktionen für die Nutzer realisiert:

1. *Hilfen auf Suchschlitzebene. Die Autosuggest- bzw. Autocomplete-Funktion.* Durch den Einbau einer Drop-Down-Hilfe auf Suchschlitzebene konnten die Nutzer sich bei Auswahl eines Suchwortes unterstützen lassen. Dabei wurden nach Eingabe eines Suchwortes automatisch als passend erkannte MeSH-Terme in Englisch und Deutsch sowie ICD-10-Terme eingeblendet, die die Nutzer durch Anklicken zur Spezifikation ihres Suchwortes benutzen können (vgl. Abbildung 6). Die Forderung nach dem Einsatz dynamischer Tools zur Unterstützung der Nutzer bei der Suche wurde schon Mitte der 90er Jahre erhoben (vgl.

Shneiderman, 1994). Doch erst seitdem entsprechende Technologien, wie AJAX oder Java, verfügbar sind, etablieren sich die dynamischen Unterstützungslösungen im webbasierten Information-Retrieval. Ein kurzer Überblick zum theoretischen Hintergrund findet sich z. B. bei Aly (2008).

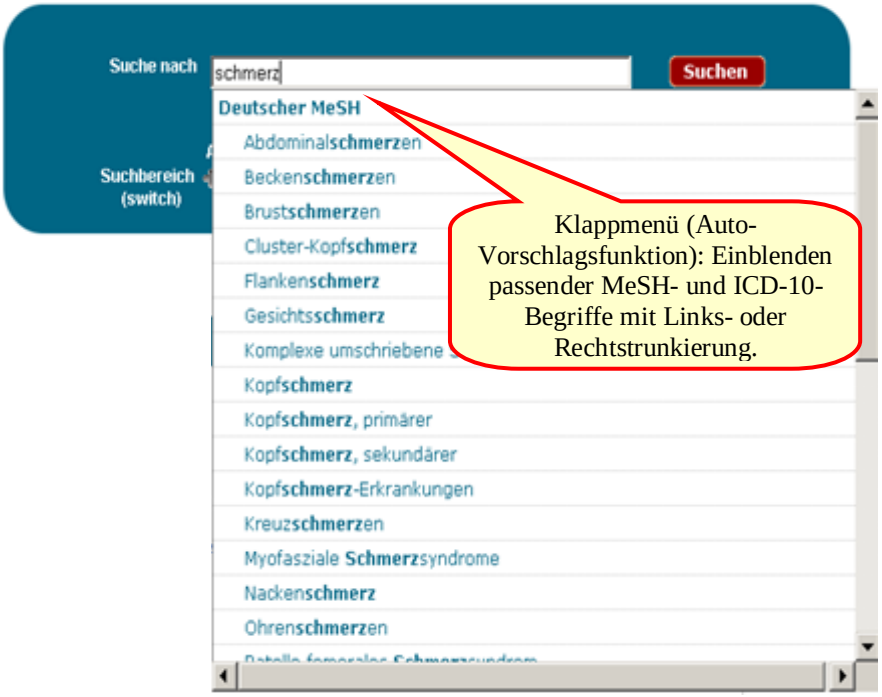

Abbildung 6. Autosuggest-Funktion. Dynamisches Tool zur Einblendung passender Suchvorschläge während der Eingabe des Suchterms.

Wie man aus den Forschungsergebnissen zur Analyse des Userverhaltens bei Suchmaschinen weiß, benutzen die meisten Personen nur Ein- bis Zwei-Wort-Anfragen für ihre Suche. Je weniger Suchworte benutzt werden, desto unspezifischer sind in der Regel aber auch die Treffer, die die Suchmaschine zurückmeldet. Dies liegt daran, dass bis heute kein Suchmaschinenalgorithmus in der Lage ist, mit minimalsten Vorgaben das Informationsbedürfnis (‚information need') genau zu erkennen. Aufgrund einer unterschiedlich hohen Suchkompetenz, Zeitmangels oder Bequemlichkeit sind viele Nutzer nicht in der Lage oder gewillt, lange über die Formulierung von adäquaten Suchtermen nachzudenken. Dies mag auch ein Ausdruck für die „Googleisierung" des allgemeinen Suchverhaltens sein. Google ist nicht zuletzt deshalb so erfolgreich, weil es den Suchprozess möglichst einfach hält – bei relativ hoher Ergebnisqualität. Hier hat in den letzten Jahren eine Habitualisierung der Mediennutzungsgewohnheiten stattgefunden – ein Gewöhnungsprozess der Nutzer, der sie

ihre Suchmaschinenerfahrung mit Google auf alle anderen Suchmaschinen übertragen lässt. Aus Sicht eines Suchmaschinenbetreibers mag dies bedauerlich sein. Als Anbieter muss man aber zur Kenntnis nehmen, mit welchen Erwartungen die Nutzer auf eine neue Suchmaschine zugehen. Angesichts dieser Sachlage scheint es deshalb strategisch sinnvoll zu sein, sowohl dem Prinzip der Einfachheit zu folgen als auch darüber nachzudenken, wie sich die Nutzer am besten darin unterstützen lassen, ihre Suche mit wenig Aufwand so zu modifizieren, dass Trefferergebnisse von hoher Relevanz gefunden werden.

2. *Hilfen zur Einschränkung des Suchraums.* Hier wurde ein *Schieberegler* entwickelt, der zur Einschränkung des Suchraums benutzt werden kann. Durch verschiedene Optionsstellungen kann der Suchraum auf bestimmte Felder im Index eingeschränkt und damit auch die Treffermenge gezielt beeinflusst werden (vgl. Abbildung 7).

Abbildung 7. Suchmaske mit Schieberegler zur Eingrenzung des Suchbereichs (grafischer Ansatz).

Um Präferenzen für den Schieberegler messen zu können, wurden den Testpersonen alternativ herkömmliche *Optionsfelder* zum Anklicken angeboten (vgl. Abbildung 8). Diese erfüllten die gleiche Funktion wie der Regler, nur dass hier mit konventionellen grafischen Elementen gearbeitet wurde.

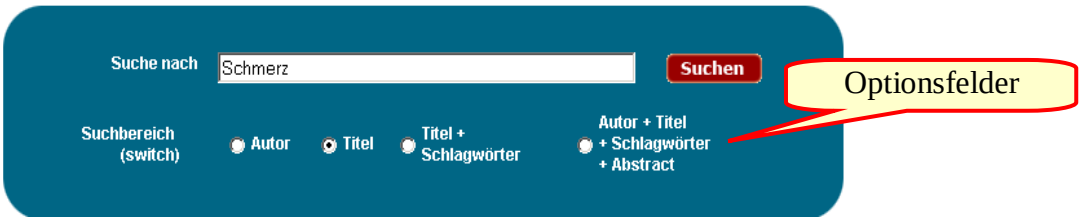

Abbildung 8. Suchmaske mit konventionellen Optionsfeldern zur Eingrenzung des Suchbereichs.

Einblendung Verwandter Suchbegriffe. Nach Absenden einer Suchanfrage wurden den Testpersonen dazu passende *Verwandte Suchbegriffe* aus dem Bereich der Medizin ange-

boten. Geordnet nach wichtigen medizinischen Oberkategorien wie Krankheit, Diagnostik, Untersuchungsmethode, Therapie, etc. wurden passende Begriffe eingeblendet und über eine nummerische und grafische Anzeige (Balken) auch die Anzahl der damit assoziierten Dokumente angezeigt (vgl. Abbildung 9).

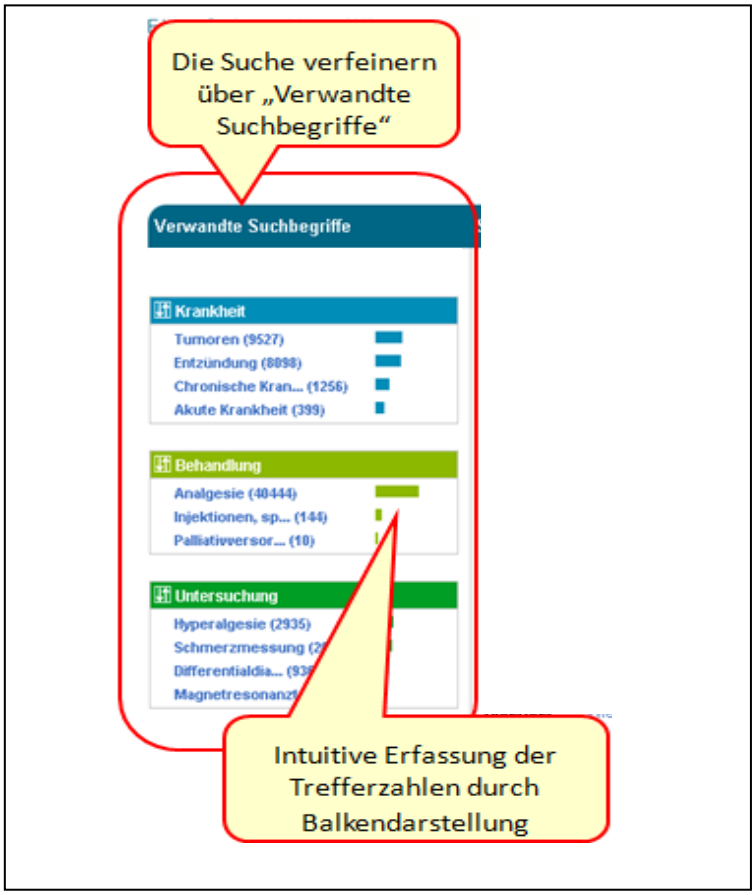

Abbildung 9. „Verwandte Suchbegriffe“. Funktion zum Verfeinern der Suche sowie intuitiv erfassbare Trefferzahlen durch Balkendarstellung.

Dabei wurde auf passende MeSH-Term-Kategorien zurückgegriffen, die mehr oder weniger eng assoziativ mit dem Suchterm verwandt sind. Durch die Einblendung dieser zusätzlichen Facetten erhielten die Testpersonen die Möglichkeit, die Suche weiter zu spezifizieren bzw. zu ‚verfeinern‘ – ohne lange über zu ergänzende Suchworte nachdenken zu müssen.

3. *Präsentation der Treffer(liste)*. In der Suchmaschinenforschung besteht keine Einigkeit darüber, was die optimale *Form der Präsentation* eines einzelnen Treffers ist. Das Entwicklerteam hat sich hier an Kaczmirek (2003) orientiert, der zu dem Schluss kommt, dass eine deutliche Erhöhung der sonst üblichen 50-200 Zeichen für die Darstellung eines Treffers

wesentlich dazu beiträgt, die Zufriedenheit mit der Trefferpräsentation zu erhöhen (vgl. auch Kap. 2.5.3). Wichtig war hier auch die Realisierung eines *Highlighting* der Suchworte, um mehr Orientierung zu schaffen und darüber hinaus durch eine vernünftige Strukturierung der Metainformation Übersichtlichkeit zu erzeugen.

4. *Zusätzliche Elemente*. Untersuchungsgegenstand war hier die Frage, wie die Nutzer die Möglichkeit der *Sortierung* und der *Optionen zur Auswahl der Treffer der einzelnen Datenbanken* sowie der *Vorabfilterung* des Suchraums durch Einschränkung auf eine bestimmte Datenbank bewerteten. Schließlich sollten die Probanden auch beschreiben, wie sie mit den *Systemrückmeldungen* zurechtkamen und ob sie diese als Hilfe und Unterstützung erlebten.

5. *Selbstbeschreibungsfähigkeit*. Die Selbstbeschreibungsfähigkeit der Averbis-Testsuchmaschine wurde *anhand der Lösung gestellter Aufgaben* sowie durch *Fragen zum Verständnis* der Funktionen im Verlauf eines fokussierten Interviews eingeschätzt.

4.3.2 Usability-Test und Fragebögen

Einen weiteren Schwerpunkt des Projekts bildete die Untersuchung von geeigneten Maßnahmen zur Verbesserung der Usability des MEDPILOT-Portals bzw. der Testsuchmaschine (vgl. Nielsen & Loranger, 2006). Um empirisch fundierte Empfehlungen für den MEDPILOT-Relaunch ableiten zu können, wurde ein szenariobasierter Usability-Test mit 24 Probanden aus der Zielgruppe des Web-Angebots durchgeführt. Die Gruppe der Testpersonen bestand aus 12 Ärzten und 12 Studenten der Medizin (nach dem Physikum) mit jeweils gleichen Anteilen an weiblichen und männlichen Testpersonen. Der Altersdurchschnitt lag bei 28,3 Jahren (SD = 5). Die Probanden wurden durch Aushänge und persönliche Ansprache an der Universitätsklinik Köln angeworben.

Im Rahmen der Usability-Untersuchung wurde auch ein fokussiertes Interview durchgeführt. Zudem wurden verschiedene Fragebögen zur Erhebung von demografischen Variablen und zur Einschätzung der verschiedenen Dimensionen der Benutzerfreundlichkeit eingesetzt sowie Fragen zur Erhebung des Images der Averbis-Testsuchmaschine. Das Ziel der Befragung war ein Vergleich mit Ergebnissen aus den Untersuchungen von El-Menouar (2004) und der vascoda-Usability-Studie 2007 (vascoda-Usability-Studie, eResult, 2007).

Im Folgenden werden die wichtigsten Variablen aus den verschiedenen Teilbereichen der Usability-Untersuchung vorgestellt[29]. Die verwendeten Fragebogeninstrumente finden sich im Anhang dieser Arbeit (vgl. Kap. 9):

I. Eingangsfragebogen (ca. 5 Min.)

Zu den wichtigsten erhobenen Variablen gehören:

a. die Erfassung der demografischen Variablen,

b. die Selbsteinschätzung der Internet- bzw. Literaturrecherchekompetenz sowie

c. die Frage, ob die Probanden MEDPILOT bereits kennen.

II. Szenariobasierter Usability-Test (ca. 20 Minuten)

Der Usability-Test begann mit dem Einstieg in die Homepage. Hier wurde untersucht, ob es den Testpersonen gelingt, drei wesentliche Basisinformationen nach einer Betrachtungsdauer von 30 Sekunden so zu perzipieren, dass im Anschluss drei Fragen richtig beantwortet werden können (vgl. Abbildung 10). 30 Sekunden ist die mittlere Zeit, die sich User nehmen, um eine für sie neue Website zu bewerten (Nielsen & Loranger, 2006). Danach wurden den Testpersonen drei Fragen in Form eines Gedächtnistests gestellt.

[29] Die vorliegende Untersuchung wurde mit Hilfe der Programmpakete *Grafstat* und Morae durchgeführt. Bei Grafstat handelt es sich um ein, für nichtkommerzielle Zwecke kostenloses, computerbasiertes Datenerhebungsprogramm von Uwe W. Diener (http://www.grafstat.de/) mit dem sich elektronische Fragebögen zur Datenerfassung erstellen lassen. Zusätzlich enthält es ein Auswertungsmodul für einfache statistische Kennwerte. Für komplexe statistische Operationen lassen sich die Daten sehr einfach in bekannte Formate exportieren.

Bei *Morae* handelt es sich um ein professionelles Programm zur Durchführung von Usability-Studien (http://www.techsmith.de/morae.asp). Dabei werden die Aktionen der User während des Usability-Tests aufgezeichnet. Zusätzlich können Audio- und Videosignale der Probanden per Netzwerk an einen zweiten Rechner übertragen werden, wo ein Protokollant, zusätzliche Beobachtungen vornehmen kann. Auch mit diesem Programm können Fragebögen für die Datenerhebung erstellt werden. Hinzu kommt ein Auswertungsmodul sowie die Möglichkeit, Highlightvideos aus dem Material der Untersuchung zusammenzustellen. Durch ein Tagging des Videos während oder nach der Untersuchung ist die Bearbeitung umfangreicher inhaltsanalytischer Fragestellungen möglich.

Abbildung 10. Werden die Basisinformationen wahrgenommen?

II.1 Fragen nach Einblendung der Eingangsseite (30 Sek.)

a. Wer betreibt die Seite? **Antwort**: „ZB MED“

b. Welche Art von Information können Sie hier recherchieren?
Antwort: „medizinische Informationen“

c. Was können Sie neben der Recherche noch tun?
Antwort(en): „Durchführen von Bestellungen“, „Aufruf von Volltexten“

II.2 Explorative Aufgabe (fünfminütige Recherche zu einer eigenen Frage)

Hier sollten die Testpersonen das Web-Angebot fünf Minuten nach eigenem Ermessen erkunden. Die einzige Auflage war, dass die Probanden Literatur zu einem Thema recherchieren sollten, das für sie von Interesse ist und bei dem sie sich gut auskennen. Damit wurde sichergestellt, dass die Testpersonen auch tatsächlich beurteilen konnten, ob die Ergebnisse ihrer Recherche auch relevant in Bezug auf ihr Informationsbedürfnis sind. Zusätzlich wurden die Probanden zum „lauten Denken“ aufgefordert. Dabei handelt es sich um eine in der Usability-Forschung verbreitete Methode zur Erhebung verbaler Daten (vgl. Ericsson & Simon, 1993; Nielsen, 1993).

II.3 Drei Rechercheaufträge (15 Minuten – Zeit pro Recherche: fünf Minuten)

Die Probanden hatten drei Rechercheaufträge mit unterschiedlichem Schwierigkeitsgrad auszuführen:

II.3.1 Leichte Aufgabe: Welches ist die neueste in der ZB MED vorhandene Auflage von Pschyrembel, Willibald: Klinisches Wörterbuch? Richtige Antwort: „261. Auflage, 2007".

II.3.2 Mittelschwere Aufgabe: Welche Erkrankung kann im Verbund mit Nasenpolypen und Aspirin-Intoleranz häufiger auftreten? Richtige Antwort: „Asthma" oder „Asthma bronchiale".

II.3.3 Schwere Aufgabe: Wie viele Gene codieren für die Geschmacksrezeptoren der Fruchtfliege (Drosophila)? Hier waren mehrere richtige Antworten möglich: Je nach dem, aus welchem Jahr der recherchierte Artikel stammt, unterscheidet sich der Wissensstand bezüglich der Anzahl der Gene. Optimal wäre natürlich das Heraussuchen eines Artikels aus der jüngsten Vergangenheit (2007 / 2008). Folgende Angaben wurden als korrekt gewertet:

- 56 Gene, Artikel aus 2003
- 60 Gene, Artikel aus 2004
- 68 Gene, Artikel aus 2007
- 68 Gene, Artikel aus 2008

Die Leistung der Probanden wurde über zwei Maße erfasst: 1. Es wurde festgehalten, ob es den Testpersonen innerhalb einer vorgegebenen Zeit von fünf Minuten gelang, die jeweilige Rechercheaufgabe mithilfe der Testsuchmaschine zu lösen (Effektivität) und 2. wurde gemessen, wie lange die Probanden tatsächlich zur Lösung der Aufgabe brauchten (Effizienz). Wenn nach fünf Minuten keine Lösung gefunden wurde, erfolgte der Abbruch durch den Versuchsleiter und die Aufgabe galt als nicht gelöst.

III. Fokussiertes Interview

Zur Erfassung der Bewertung der implementierten Unterstützungsfunktionen wurde ein fokussiertes Interview mit den Testpersonen durchgeführt (ca. 15–25 Minuten): Hier wurden die Probanden systematisch nach ihrem Verständnis und nach der

Bewertung der Funktionen, Optionen und Designelementen zur Verbesserung der Usability der Website befragt.

IV. Abschlussfragebogen

Zu den erhobenen Variablen gehörten hier:

- Ratingfragen zu klassischen Usability-Eigenschaften,
- die Frage nach der Präferenz bei der Sortierung von Treffern (Relevanz, Aktualität, Kombination aus Relevanz und Aktualität),
- die Bewertung der Benutzerfreundlichkeit der Testsuchmaschine mithilfe von Rating-Skalen (für den Vergleich mit einer früheren ZB MED-Untersuchung zur Bewertung von MEDPILOT (vgl. El-Menouar, 2004).
- Imagemessung mithilfe eines Polaritätsprofils: Dabei ging es um den Vergleich zwischen den Imagewerten für die wissenschaftliche Suchmaschine „vascoda“ (vascoda-Usability-Studie, eResult, 2007) und der Testsuchmaschine.

V. Aufklärung über den Hintergrund der Untersuchung (Debriefing) sowie Auszahlung der Aufwandsentschädigung

Aus forschungsethischen Gründen sollten die Teilnehmer einer wissenschaftlichen Untersuchung stets über die Hintergründe informiert werden (vgl. Dzeyk, 2001). In diesem Debriefing wurden die Testpersonen über den theoretischen Hintergrund aufgeklärt und es wurde ihnen Gelegenheit gegeben, Fragen zur Untersuchung zu stellen. Schließlich wurden die Mediziner und Studierende (der Medizin) mit 50,- € für ihre Teilnahme an der Untersuchung entlohnt. Der Incentive-Betrag wurde deshalb so hoch angesetzt, weil es sich in der Pretest-Phase der Untersuchung als sehr schwierig erwiesen hatte, Mediziner mit dicht gedrängtem Terminkalender zu einer Teilnahme zu bewegen.

5 Projektergebnisse

5.1 Inhaltsanalyse

5.1.1 Komplexität der MEDPILOT-Suchanfragen

Um die Komplexität der von den MEDPILOT-Nutzern eingesetzten Suchterme zu untersuchen, wurde mithilfe sogenannter „Regular Expressions“ (vgl. Kap. 4.1.1.3) analysiert, wie viele Suchworte pro Anfrage bzw. Suchterm benutzt wurden (N = 142.922 Suchanfragen). Tabelle 9 und Abbildung 11 fassen die Ergebnisse zusammen: In 35,8% der untersuchten Anfragen wurden Ein-Wort-Suchen ausgeführt. Weitere 30,1% der Suchterme bestanden aus zwei Suchworten. Letztlich wurden also zwei Drittel der Suchanfragen aus Ein- oder Zwei-Wort-Anfragen gebildet. Dies bedeutet: Die Mehrzahl der Anfragen besteht aus wenig komplexen Suchformulierungen. Das Ergebnis liegt also nahe an den aus der Literatur bekannten Zahlen (vgl. Kap. 2.5.2). Mediziner und Medizinstudenten unterscheiden sich offensichtlich im Durchschnitt nicht von der „normalen“ Internetpopulation.

WpQ	Anzahl	Prozent
1	51136	35,78
2	43316	30,07
3	22938	16,05
4	9461	6,62
5	5228	3,66
6	2794	1,95
7	1942	1,36
8	1439	1,01
9	1014	0,71
10	815	0,57
11	625	0,44
12	500	0,35
13	413	0,29
14	295	0,21
15	221	0,15
16	196	0,14
17	136	0,10
18	100	0,07
19	77	0,05
20	60	0,04
21-88	206	0,14

Tabelle 9. Formale Komplexität der Suchanfragen (WpQ = Worte pro Query bzw. Suchanfrage). Insgesamt wurde mit 372.138 Worten gesucht. Bei 142.922 Suchanfragen sind dies im Durchschnitt 2,6 Suchworte pro Query. 81,9% der Queries bestehen aus höchstens drei Suchworten.

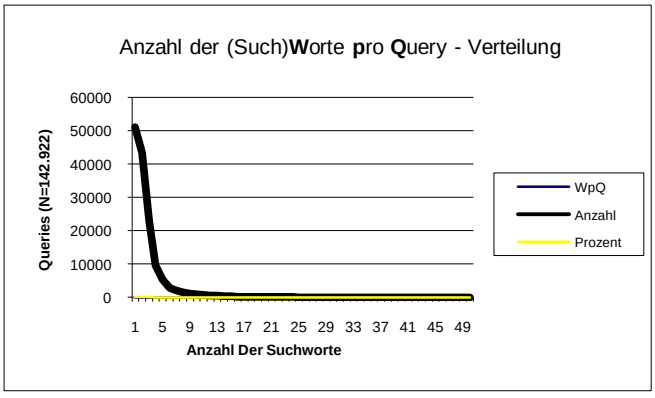

Abbildung 11. Suchworte pro Suchanfrage (N=142.922 Queries).

Der Anteil der Suchanfragen, die Feldbezeichner (z.B.: TI = Titel, AU = Autor) für eine nähre Spezifizierung der Suche beinhalteten, umfasste lediglich 2,4%.

Boolesche Operatoren wie AND, OR, NOT oder NOR wurden nur in ca. 10% der Suchen benutzt. Mediziner scheinen sich auch in diesem Punkt nicht wesentlich von anderen Nutzergruppen zu unterscheiden.

5.1.2 Welche Inhalte werden in MEDPILOT gesucht?

Die Ergebnisse der Inhaltsanalyse zeigten, dass in MEDPILOT vor allem nach Informationen aus den Bereichen *Krankheiten*, *Syndrome* und *Symptome* gesucht wird (30,85%). Auf dem zweiten Platz der meistgesuchten Themenbereiche fanden sich Fragen nach *Behandlungsmethoden* und *Therapieansätzen* sowie *diagnostische Fragestellungen* (28,45%). Eine dritte Gruppe von häufigen Suchinhalten betraf *sozialmedizinische Fragen*, *statistische Kennwerte* sowie *empirische Studien* und *epidemiologische Fragen* (15,58%). Erst auf dem vierten Platz mit 13,41% fanden sich Fragen, die aus den Bereichen *Chemie, Biochemie, Medikamentenwirkstoffe und Handelsnamen* stammten (vgl. Abbildung 12).

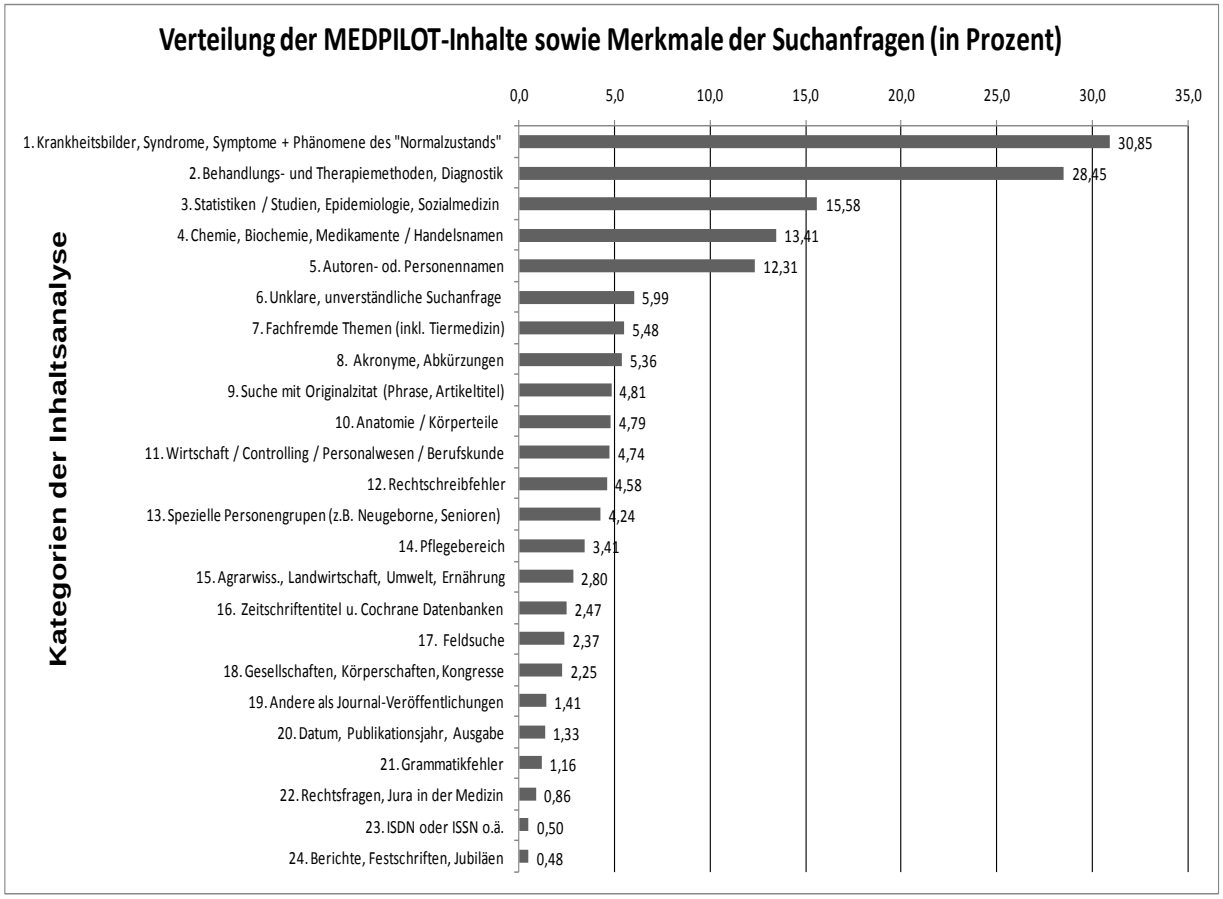

Abbildung 12. Gesuchte MEDPILOT-Inhalte und Merkmale der Suchanfrageen (Ergebnis der Inhaltsanalyse des MEDPILOT-Logfiles, n = 10.000 Suchanfragen).

Damit lag der Gesamtanteil dieser Art von Suchanfragen unter 20%, sodass der Entschluss gefasst wurde, vorerst keinen eigenen Chemiethesaurus auf der Basis von Morphemen bzw. Subworten zu entwickeln. Bei einer genaueren Analyse zeigte sich zudem, dass der Anteil von originären Chemieanfragen mit 2,1% noch weit unter dem Wert liegt, ab dem es sich lohnen würde, einen eigenen Thesaurus aufzubauen. Langfristig wäre es jedoch sinnvoll, eine Lösung zur Unterstützung der Suche nach biologischen und biochemischen Inhalten (zusammen ca. 6,2%) sowie Wirkstoffen und Medikamentennamen/Handelsnamen (zusammen ca. 5,1%) zu implementieren. Dies könnte über eine Erweiterung des Morphem- bzw. Subwortlexikons sowie über die Integration fachspezifischer Thesauri geschehen.

Die genaue Verteilung der 10.000 untersuchten Queries auf sämtliche 24 Kategorien bzw. Inhaltsbereiche kann Abbildung 12 entnommen werden. Das Ergebnis basiert auf einer Zufallsstichprobe von 10.000 Suchtermen aus der Gesamtmenge aller 142.922 Suchanfragen (der untersuchten 7 Monate).

5.2 Retrieval-Tests

Mithilfe systematischer Retrieval-Tests wurde ermittelt, wie hoch der Anteil relevanter Treffer bei Suchanfragen ist (zur Ermittlung der Precision, vgl. Kap. 2.1). Zunächst wurde ein direkter Vergleich zwischen der Averbis-Testsuchmaschine und der bisherigen MEDPILOT-Suche durchgeführt. Die Abschätzung der Leistungsfähigkeit für die Verarbeitung verschiedener Sprachphänomene erfolgte unter Einsatz jeweils spezifischer Testkollektionen (vgl. Kap. 4.1.1.3). In der folgenden Ergebnisbeschreibung werden vornehmlich die Resultate für den Cut-Off-Wert p5 berichtet (also bis zum fünften Treffer). Es werden jeweils die Mittelwerte für die Retrieval-Tests der beiden Datenbanken MEDPILOT und CC MED angegeben. Die Konzentration der Ergebnisdarstellung auf die Resultate für die ersten fünf Treffer leitet sich aus der Beobachtung ab, dass für die Nutzer insbesondere diese ersten Treffer von großer Wichtigkeit sind. Die Chance, dass diese Treffer angeklickt werden, ist bedeutend größer als bei den nachfolgenden Treffern (vgl. Kap. 2.5.2). Die Ergebnisse bzw. Precision-Werte für die anderen Cut-Off-Werte und die einzelnen datenbankbezogenen Vergleiche zwischen Averbis und MEDPILOT (MEDLINE und CC MED) können im Einzelnen Tabelle 10 und 11 entnommen werden.

5.2.1 Verarbeitung problematischer Sprachaspekte

5.2.1.1 Rechtschreibfehler

Die Averbis-Testsuchmaschine wurde mithilfe einer speziellen Testkollektion auf ihre Leistungsfähigkeit zur toleranten Verarbeitung von Rechtschreibfehlern mit insgesamt 50 Suchtermen getestet, die Rechtschreibfehler enthielten. Dieser Test wurde sowohl mit der Datenbank MEDLINE als auch mit der Datenbank CC MED durchgeführt. Bei einem Cut-Off-Wert von 5 wurde eine Trefferrelevanz von p = 0,255 ermittelt. Das bedeutet, 25,5% der unter den ersten fünf Treffern waren relevant (trotz Rechtschreibfehler in der Suchanfrage). Die gleiche Testserie wurde mit der MEDPILOT-Suchmaschine durchgeführt und zeigte dort im Durchschnitt eine Precision von 0,042 (4,2% relevante Hits unter den ersten 5 Treffern). Das heißt: MEDPILOT war – im Gegensatz zum Averbis-System – nicht bzw. nur in sehr geringem Maße dazu in der Lage, Suchphrasen mit Rechtschreibfehlern so aufzulösen, dass relevante Treffer gefunden werden konnten.

Tabelle 10. Anteil an relevanten Treffern (Precision*) für die Averbis-Testsuchmaschine und die bisherige MEDPILOT-Suchumgebung für die problematischen Sprachaspekte: *Rechtschreibfehler*, *Akronyme*, *Synonyme* und *Komposita*. Ergebnisse für die ersten fünf, 10, 15 und 20 Treffer (für die Datenbanken MEDLINE und CC MED). „Gesamt“ bedeutet: gemittelter Precision-Wert [in Prozent] aus den Resultaten für MEDLINE und CC MED.

Sprachaspekt		Vergleich der Trefferrelevanz (Precision-Werte in Prozent)					
		AVERBIS gesamt	MEDPILOT gesamt	AVERBIS MEDLINE	MEDPILOT MEDLINE	AVERBIS CCMED	MEDPILOT CCMED
		Mittelwert	Mittelwert	Mittelwert	Mittelwert	Mittelwert	Mittelwert
Rechtschreibfehler	p5 [%]	25,50	4,20	22,00	6,80	29,00	1,60
	p10 [%]	23,20	3,20	19,90	5,20	26,50	1,20
	p15 [%]	21,90	2,47	19,13	4,13	24,67	,80
	p20 [%]	21,10	2,10	18,45	3,60	23,75	,60
Akronyme	p5 [%]	35,20	33,20	42,00	36,80	28,40	29,60
	p10 [%]	31,60	29,30	39,40	35,20	23,80	23,40
	p15 [%]	30,07	26,87	38,93	34,00	21,20	19,73
	p20 [%]	29,00	25,60	38,60	33,70	19,40	17,50
Synonyme (Original)	p5 [%]	60,40	38,60	67,20	41,60	53,60	35,60
	p10 [%]	57,40	32,50	65,20	37,40	49,60	27,60
	p15 [%]	54,60	29,73	63,60	34,40	45,60	25,07
	p20 [%]	52,5	27,30	62,10	33,00	42,80	21,50
Synonyme (modifiziert)	p5 [%]	42,40	13,60	47,60	13,60	37,20	13,60
	p10 [%]	40,40	10,40	46,40	9,80	34,40	11,00
	p15 [%]	38,33	9,13	44,80	8,40	31,87	9,87
	p20 [%]	36,80	8,40	43,30	7,90	30,20	8,90
Komposita (Original)	p5 [%]	75,10	80,40	62,40	77,80	87,80	83,00
	p10 [%]	72,05	73,30	60,00	71,80	84,10	74,80
	p15 [%]	69,47	68,10	59,27	66,60	79,67	69,60
	p20 [%]	67,10	64,00	58,40	62,65	75,70	65,25
Komposita (zerlegt)	p5 [%]	81,40	34,00	75,60	37,20	87,20	30,80
	p10 [%]	78,40	28,40	73,40	31,60	83,40	25,20
	p15 [%]	77,33	24,73	73,33	27,33	81,33	22,13
	p20 [%]	76,20	22,60	72,10	24,90	80,20	20,20

* Die Precision-Werte (p5 bis p20) sind in Prozentangaben umgerechnet worden.

Tabelle 11. Anteil an relevanten Treffern (Precision*) für die Averbis-Testsuchmaschine und die bisherige MEDPILOT-Suchumgebung für die problematischen Sprachaspekte: A*utomatische Übersetzung*, *Laien-Experten-Sprache* und *grammatikalische Variationen*. Ergebnisse für die ersten fünf, 10, 15 und 20 Treffer (für die Datenbanken MEDLINE und CC MED). „Gesamt" bedeutet: gemittelter Precision-Wert [in Prozent] aus den Resultaten für MEDLINE und CC MED.

Sprachaspekt		Vergleich der Trefferrelevanz (Precision-Werte in Prozent)					
		AVERBIS gesamt	MEDPILOT gesamt	AVERBIS MEDLINE	MEDPILOT MEDLINE	AVERBIS CCMED	MEDPILOT CCMED
		Mittelwert	Mittelwert	Mittelwert	Mittelwert	Mittelwert	Mittelwert
Übersetzung de - en	p5 [%]	92,80	74,20	88,00	57,60	97,60	90,80
	p10 [%]	93,60	70,90	90,00	54,00	97,20	87,80
	p15 [%]	93,60	69,47	89,87	52,93	97,33	86,00
	p20 [%]	93,45	69,45	89,60	53,10	97,30	85,80
Übersetzung en - de	p5 [%]	91,80	67,00	91,20	40,40	92,40	93,60
	p10 [%]	91,00	65,20	90,60	39,40	91,40	91,00
	p15 [%]	90,93	64,20	91,20	39,60	90,67	88,80
	p20 [%]	90,70	62,70	91,10	39,60	90,30	85,80
Laiensprache	p5 [%]	92,40	78,80	90,00	70,40	94,80	87,20
	p10 [%]	91,20	71,00	88,80	61,60	93,60	80,40
	p15 [%]	90,27	65,80	89,07	55,73	91,47	75,87
	p20 [%]	88,80	62,85	88,90	52,80	88,70	72,90
Expertensprache	p5 [%]	93,00	77,20	93,60	70,00	92,40	84,40
	p10 [%]	92,00	72,90	92,40	66,80	91,60	79,00
	p15 [%]	90,73	70,20	92,27	63,33	89,20	77,07
	p20 [%]	90,00	68,95	92,40	63,40	87,6	74,5
Grammatik Variation (Original)	p5 [%]	48,80	20,00	54,40	20,40	43,20	19,60
	p10 [%]	43,30	14,70	51,40	15,20	35,20	14,20
	p15 [%]	39,87	11,73	49,87	12,27	29,87	11,20
	p20 [%]	36,70	9,65	47,00	9,90	26,4	9,4
Grammatik Variation (modifiziert)	p5 [%]	45,80	15,40	53,60	16,40	38,00	14,40
	p10 [%]	40,70	13,60	51,20	15,00	30,20	12,20
	p15 [%]	37,73	12,53	49,20	14,13	26,27	10,93
	p20 [%]	35,05	11,15	46,90	12,70	23,2	9,6

* Die Precision-Werte (p5 bis p20) sind in Prozentangaben umgerechnet worden.

5.2.1.2 Akronyme

Für die korrekte Auflösung von Akronymen und Abkürzungen (vgl. Tabelle 10) zeigte sich, dass die Averbis-Testsuchmaschine MEDPILOT insbesondere dann überlegen war, wenn es um die Analyse englischsprachiger Akronyme und Abkürzungen ging. Bei der Überprüfung anhand der MEDLINE-Daten schaffte es die MorphoSaurus-Technik 42% der Akronyme (unter den ersten fünf Treffern) richtig aufzulösen. MEDPILOT kam hier im direkten Vergleich auf einen Wert von 36,8%. Bei einem Vergleich anhand der vorwiegend deutschsprachigen CC MED-Daten zeigte sich kein bedeutender Unterschied zwischen den beiden Systemen (28,4% für Averbis vs. 29,6% für MEDPILOT). Im Gesamtvergleich – also über beide Datenbanken hinweg – gelang es Averbis mit 35,2% (unter den ersten fünf Treffern) etwas besser als MEDPILOT (p5 = 33,2%) Akronyme und Abkürzungen so aufzulösen, dass relevante Treffer gefunden wurden. Ein im Projektverlauf verbessertes Akronymmodul erreichte eine wesentlich bessere Leistung der Testsuchmaschine, doch stehen systematische Tests hier noch aus.

5.2.1.3 Synonyme

Hier wurde getestet, ob die beiden Systeme dazu in der Lage waren, mit *synonymen* Suchphrasen bzw. -begriffen eine ähnlich hohe oder gleiche Zahl an relevanten Treffern zu finden. Ein Beispiel für eine solche Suchanfrage aus der Testkollektion ist das Paar ‚Osteoporose' vs. ‚Knochenschwund'. Mit den beiden Suchworten sollten im besten Fall gleichviele relevante Treffer gefunden werden. Ein anderes Beispiel sind die Suchterme ‚Bluthochdruck Akupunktur' vs. ‚Hypertonie Akupunktur'.

Beim Rückgriff auf die MEDLINE-Datenbank fand die Averbis-Testsuchmaschine bei Eingabe von ‚Hypertonie Akupunktur' 186 Treffer, wobei 20 der ersten 20 Treffer relevant waren. Bei Eingabe von ‚Bluthochdruck Akupunktur' zeigte die Averbis-Suchmaschine 103 Treffer an, wobei sich unter den ersten 20 Treffern 19 relevante Treffer befanden. MEDPILOT konnte mit der Suchphrase ‚Hypertonie Akupunktur' nur neun Treffer finden, wovon sieben relevant waren. Bei Eingabe von ‚Bluthochdruck Akupunktur' wurden mit der MEDLINE-Abfrage kein einziger Treffer gefunden. Eine mögliche Erklärung für den Unterschied bei der Auflösung von Synonymen liegt darin, dass es sich bei MEDLINE um eine vorwiegend englischsprachige Datenbank handelt, sodass die MorphoSaurus-Technik mit ihrer Fähigkeit zur „Übersetzung von Fachbegriffen" hier eindeutig im Vorteil ist.

Für die Überprüfung der Trefferrelevanz in Bezug auf die Testkollektion *Synonyme* wurde folgendes Gesamtergebnis festgestellt: Mit den originalen Suchphrasen lieferte Averbis im Schnitt 60,4% relevante Treffer unter den ersten fünf Rückmeldungen. MEDPILOT erreichte hier lediglich eine Trefferrelevanz von 38,6%. Bei der Testung mit den modifizierten Suchphrasen (s.o.) zeigte sich bei Averbis eine Verringerung der Trefferrelevanz auf 42,4% (also knapp 30% weniger). Die Verringerung der Relevanzwerte bei MEDPILOT war dagegen viel deutlicher: Hier war im Mittel ein Rückgang um fast 65% zu verzeichnen: von 38,6% (Originalsuchphrasen) auf 13,60% (modifizierte Suchterme). Spezifische Vergleiche für die einzelnen Datenbanken können Tabelle 11 entnommen werden.

5.2.1.4 Komposita

Suchmaschinennutzer geben Suchworte häufig einzeln ein, anstatt ein Kompositum zu verwenden; wie z.B. ‚Augen Innendruck', statt ‚Augeninnendruck'. Eine Suchmaschine sollte in der Lage sein, zu erkennen, dass hier nach den gleichen oder ähnlichen Inhalten gesucht wird.

Für die Averbis-Testsuchmaschine wurde angenommen, dass sie eine bessere Auflösung von Komposita liefert. Diese These konnte durch die durchgeführten Tests bestätigt werden. Insgesamt betrug der Unterschied zwischen der Trefferrelevanz der getesteten Komposita-Suchphrasen (z. B. ‚Augeninnendruck') und den „zerlegten" Suchphrasen (z. B. ‚Augen Innendruck') bei Averbis im Mittel 6,3% (75,10% für zusammengesetzten Begriffe und 81,40% für die zerlegten Suchbegriffe). Der Wert für die Trefferrelevanz nahm hier also sogar noch zu! MEDPILOT lieferte wesentlich weniger relevante Treffer, wenn Komposita aufzulösen waren: Im Original waren es 80,4% und zerlegt lediglich nur noch 34% relevante Treffer unter den ersten fünf Rückmeldungen. Dies bedeutet einen Unterschied von fast 58%; ein eklatanter „Leistungsabfall", wenn es darum geht, Suchanfragen mit gleichem oder ähnlichem Inhalt korrekt aufzulösen. Hier zeigt sich der große Vorteil der MorphoSaurus-Technik: Durch den hinterlegten Morphemthesaurus gelingt es Averbis sehr gut, solche Unterschiede in Suchanfragen abzufangen und mit wenig Verlust aufzulösen. Ergebnisse für die Analyse der einzelnen Datenbanken sind Tabelle 11 zu entnehmen.

5.2.1.5 Übersetzungsleistung

Hier zeigte sich sehr deutlich die Stärke der Averbis-Testsuchmaschine gegenüber MEDPILOT. Im Gesamtergebnis waren bei der Testsuchmaschine unter den ersten fünf Treffern für die deutschen Suchterme 92,8% relevant. Für die ins Englische übersetzten Suchwörter verringerte sich dieser Anteil kaum: Hier waren es 91,8% relevante Treffer unter den ersten fünf Trefferrückmeldungen. Bei der Testung von MEDPILOT konnten für die deutschen Suchanfragen 74,2% relevante Treffer festgestellt werden; für die englischen Suchterme hingegen nur noch 67%. Wenn deutsche Suchanfragen gestellt wurden, erreichte die Testsuchmaschine eine höhere Precision und damit ca. 20% mehr relevante Treffer (unter den ersten fünf Hits) als MEDPILOT. Gemessen an Averbis wurden bei englischen Suchanfragen bei MEDPILOT sogar ein um 27,1% geringerer Wert festgestellt. Betrachtet man die Ergebnisse für die einzelnen Datenbanken, wird der Unterschied noch deutlicher: Bei deutschen Anfragen an die vorwiegend englischsprachige MEDLINE erreichte MEDPILOT eine Trefferrelevanz von 57,6% unter den ersten fünf Treffern; die Testsuchmaschine lag hier bei 88%. Dies ist eine Differenz von 30,4% zwischen den beiden Suchsystemen. Bei englischsprachigen Anfragen erreichte MEDPILOT im Mittel eine Trefferrelevanz von 40,4%. Bei den gleichen englischen Suchtermen kam die Testsuchmaschine hier auf einen Wert von

91,2%: ein Unterschied in der Precision von p = 0,51 (vgl. Tabelle 11). Diese Tests zeigen eindrucksvoll die Überlegenheit der MorphoSaurus-Technik bei sprachübergreifenden Suchanfragen.

Die Übersetzungsleistung durch die MorphoSaurus-Technik soll im Folgenden an einem Beispiel demonstriert werden. Verglichen werden die Ergebnisse einer deutschen Anfrage an MEDLINE über das Averbis-Testsystem und über die Suchmaschine Google. Gesucht wurden Publikationen zum Thema „Flüssigkeitsmangel und Austrocknungszustände im Alter“. Dazu wurde mit dem Suchterm ‚dehydration im alter‘ nach Literatur gesucht, die diese Worte im Titel enthalten sollten. Die Averbis-Testsuchmaschine lieferte unter den ersten 20 Treffern ausschließlich relevante Publikationen (mit Titeln wie z. B. „dehydration in the aged“ oder „dehydration in the elderly“.

Bei Eingabe des gleichen Suchterms bei Google (unter Beschränkung auf die MEDLINE-Suche) zeigte sich folgendes Bild: Wenn die Suche auf das Suchfeld „Titel“ eingeschränkt wird, findet Google keinen einzigen Treffer. Ohne diese Einschränkung findet Google zwar 304 Treffer, aber unter den ersten 20 Treffern befindet sich kein einziger Artikel, der das gesuchte Thema behandelt (vgl. Abbildung 13).

Abbildung 13. Beispiel für mangelhafte Übersetzungsleistung durch Google. Gesucht wurde mit dem Suchterm ‚dehydration im alter‘ (Juni, 2008).

Sucht man nicht nur ausschließlich in MEDLINE, werden natürlich viele relevante Treffer gefunden, wobei aber von den deutschen Treffern nur wenige aus wissenschaftlichen Quellen stammen. Die angezeigten englischsprachigen Treffer sind zumeist nicht relevant. Dies hat seinen Grund wiederum darin, dass Google keine ausreichende Übersetzungsleistung bietet.

5.2.1.6 Laien-Expertensprache

Ein weiteres Ziel des Projekts war die Verbesserung der gleichwertigen Behandlung von Laien- und Expertenanfragen. Ob ein Nutzer nun „Schluckstörung“ oder „Dysphagie“ als Suchbegriff in die Suchmaske eingibt, sollte keine Rolle spielen. Hier zeigte sich folgendes Ergebnis: Im Mittel fanden sich für die Averbis-Testsuchmaschine 92,4% relevante Treffer für die Suche mit laiensprachlichen medizinischen Begriffen unter den ersten fünf Trefferrückmeldungen. Die gleichen Suchinhalte als expertensprachliche Fachterme formuliert, erbrachten ein Ergebnis von 93%. MEDPILOT erreichte für die laiensprachlichen Sucheingaben einen Wert von 78,8% und für die expertensprachlichen Abfragen einen Wert von 77,20%. Insgesamt zeigte sich hier, dass bei Einsatz der MorphoSaurus-Technik nur mit wenig Verlust an relevanten Rückmeldungen – auf gleichbleibend hohem Niveau – zu rechnen ist, wenn alternativ zu einem Expertensuchwort mit einem Laienbegriff gesucht wird. Mit laiensprachlichen Suchtermen fand MEDPILOT 13,6% weniger relevante Treffer. Bei expertensprachlichen Suchen erreichte die Testsuchmaschine 15,8% mehr relevante Treffer unter den ersten fünf Hits (vgl. Tabelle 11).

5.2.1.7 Grammatikalische Variationen

Für die Bewertung der Leistungsfähigkeit bei der Verarbeitung grammatikalischer Variationen wurden die Ergebnisse von zwei Testkollektionen verglichen: Es wurden 50 Original-Suchphrasen getestet und 50 dazu passende grammatikalische Varianten. Zum Beispiel wurde der Suchphrase ‚blutdruckmessung bei kindern‘ die Singularbildung ‚blutdruckmessung beim kind‘ gegenübergestellt. Bei dieser Testreihe wurde ebenfalls eine Überlegenheit der MorphoSaurus-Technik erwartet.

Die Averbis-Testsuchmaschine lieferte für die Original-Anfragen unter den ersten fünf Hits im Mittel 48,8% relevante Treffer zurück. Für die grammatikalischen Variationen wurde eine Relevanzquote von durchschnittlich 45,8% festgestellt; also nur drei Prozent weniger. MEDPILOT erreichte für die Originalanfragen dagegen nur 20% Trefferrelevanz und für die

Variationen lediglich 15,4%. Insgesamt zeigte sich die Averbis-Technik bei der alternativen Verarbeitung von grammatikalischen Variationen tatsächlich sehr viel leistungsfähiger als MEDPILOT. Bei den Originalanfragen wurden 28,8% mehr Trefferrelevanz erzielt; bei den Variationen betrug der Unterschied sogar 30,4%.

5.2.1.8 Gesamtvergleich zwischen Testsuchmaschine und MEDPILOT

Für einen globalen Vergleich zwischen der Averbis-Testsuchmaschine und MEDPILOT wurden sämtliche Ergebnisse der Precision-Tests, die mit den verschiedenen Testkollektionen durchgeführt wurden, zusammengeführt und gemittelt.

Zunächst wurde ein Vergleich zwischen den einzelnen Suchmaschinen und Datenbanken vorgenommen (vgl. Tabelle 12), bevor dann ein Gesamtvergleich durchgeführt worden ist (vgl. Abbildung 14). Jeder Testkombination von Suchmaschine und Datenbank lagen also 600 Überprüfungen auf Trefferrelevanz zugrunde (12 Testkollektionen x 50 Suchanfragen = 600 Suchanfragen). Zusätzlich sind in Tabelle 12 auch die durchschnittlichen Trefferzahlen und die Anzahl der durchschnittlich gemeldeten Null-Treffer-Meldungen aufgeführt; jeweils getrennt für die Tests mit MEDLINE und CC MED.

Tabelle 12. Ergebnisse für den Anteil an relevanten Treffern (Precision) für die Averbis-Testsuchmaschine und die bisherige MEDPILOT-Suchumgebung in Abhängigkeit von der getesteten Datenbank (MEDLINE und CC MED) mit je N=600 Suchanfragen.

		Suchmaschine	
		Averbis-Testsuchmaschine (moderne Suchmaschinentechnologie + Morphosaurustechnik)	**MEDPILOT** (Metasuchmaschine)
Datenbank	**MEDLINE** (Umfang: ca. 16 Mio; vorwiegend englischsprachig)	Treffer (Mittelwert) = 154772 (n=600) Anzahl der Null-Treffer-Suchen = 16,8%	Treffer (Mittelwert) = 79066 (n=600) Anzahl der Null-Treffer-Suchen = 36,1%
		Anteil an relevanten Treffern bei verschiedenen Cut-Off-Werten: p5 = 65,63 %	Differenz 24,88%	p5 = 40,75 % p10 = 64,06 %	27,14%	p10 = 36,92 % p15 = 63,38 %	28,98%	p15 = 34,40 % p20 = 62,40 %	29,30%	p20 = 33,10 %
	CCMED (Umfang: ca. 450.000; vorwiegend deutschsprachig)	Treffer (Mittelwert) = 1659 (n=600) Anzahl der Null-Treffer-Suchen = 23,8 %	Treffer (Mittelwert) = 378 (n=600) Anzahl der Null-Treffer-Suchen = 44,1 %
		Anteil an relevanten Treffern bei verschiedenen Cut-Off-Werten: p5 = 65,13 %	Differenz 16,45%	p5 = 48,68 % p10 = 61,75 %	17,77%	p10 = 43,98 % p15 = 59,10 %	17,68%	p15 = 41,42 % p20 = 57,13 %	17,80%	p20 = 39,33 %

Suchmaschinen-Datenbank-Kombinationen. Insgesamt zeigte sich hier die Averbis-Testsuchmaschine dem MEDPILOT-Ansatz deutlich überlegen. Dies gilt sowohl für den Vergleich auf der Basis von MEDLINE als auch für den Vergleich auf der Grundlage der CC MED-Daten.

MEDLINE-Vergleich. Für den so wichtigen Bereich der ersten fünf Treffer (p5) übertraf Averbis die MEDPILOT-Suche hinsichtlich der *Trefferrelevanz* um 24,8% (65,63% vs. 40,75%). Betrachtet man die ersten 20 Treffer (p20) sind es sogar fast 30% (29,3%) mehr relevante Treffer, die mithilfe der MorphoSaurus-Technik gefunden wurden (62,4%% vs. 33,10%).

CC MED-Vergleich. Für den Cut-Off-Wert von p5 wurde für die Averbis-Testsuchmaschine ein Precision-Wert von 65,13% festgestellt. Für MEDPILOT lag die Leistungsfähigkeit bei der Trefferrelevanz bei 48,68%. Dies bedeutet: Im Durchschnitt fand MEDPILOT 16,45% weniger relevante Treffer unter den ersten fünf Hits. Für den Cut-Off-Wert von 20 (p20) wurde eine Differenz von 17,8% zugunsten der Averbis-Testsuchmaschine ermittelt (Averbis: 57,13% vs. MEDPILOT: 39,33%).

Gesamtvergleich (MEDLINE und CC MED). Neben den Einzelvergleichen für die beiden Datenbanken MEDLINE und CC MED wurde auch ein Gesamtvergleich auf Grundlage beider Datenquellen angestellt (vgl. Abbildung 14).

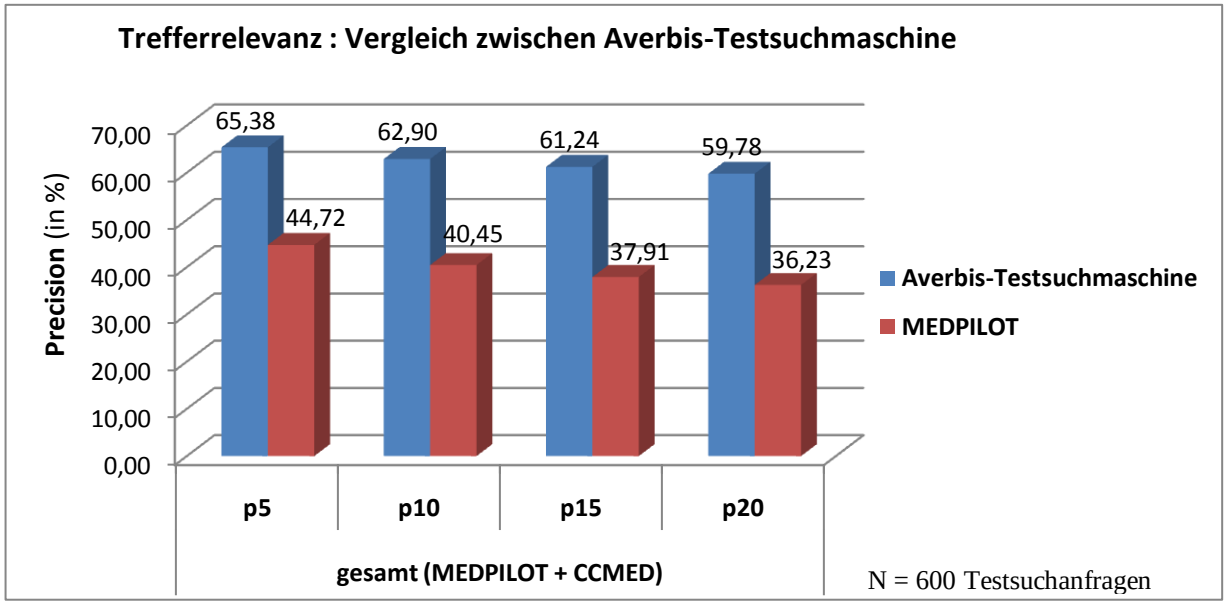

Abbildung 14. Vergleich der Trefferrelevanz zwischen der Averbis-Testsuchmaschine und der MEDPILOT-Suchumgebung (Precisionwerte in Prozent) für verschiedene Cut-Off-Werte. Die Ergebnisse wurden aus den Berechnungen für die Überprüfungen der Datenbanken MEDLINE und CC MED zusammengefasst (Berechnungsgrundlage waren also jeweils N = 600 Suchanfragen).

Es wurden folgende Precision-Werte festgestellt: Für den Cut-Off-Bereich p5 erreichte die Averbis-Testsuchmaschine einen Wert von 65,38%. Mit 44,72% lag MEDPILOT 20,66 Prozentpunkte darunter. Für den Cut-Off-Wert von p20 zeigte die Averbis-Testsuchmaschine einen Wert von 59,78%. Für MEDPILOT konnte ein Trefferrelevanzwert von 36,23% festgestellt werden. Dies entspricht einer Differenz von 23,55 Prozentpunkten zugunsten der Averbis-Testsuchmaschine.

Anzahl der Treffer (Vergleichsbasis: nur MEDLINE-Daten). Durchschnittlich wurden von der Averbis-Testsuchmaschine auf jede Suchanfrage 154772 Treffer zurückgemeldet (Median = 1789). Bei MEDPILOT waren es hingegen 79065 (Median = 7).

Anzahl der Treffer (Vergleichsbasis: nur CC MED-Daten). Für die Averbis-Testsuchmaschine wurde ein Mittelwert von 1659 Treffer (Median = 23) festgestellt und für MEDPILOT ein Mittelwert von 378 Treffer (Median = 1). Der Median von eins bedeutet hier, dass bei 100 Testsuchanfragen 50% der Trefferrückmeldungen höchstens einen Treffer enthielten (50 Suchanfragen).

Dabei variierte die Anzahl der Treffer sehr stark in Abhängigkeit von der jeweiligen Testkollektion (vgl. Tabelle 13). Indirekt ist dies ein Hinweis darauf, dass die Suchmaschinen in Bezug auf die Verarbeitung der verschiedenen Sprachaspekte unterschiedlich leistungsfähig sind. Wenn bei MEDPILOT bei Suchanfragen mit Rechtschreibfehlern kaum Treffer zurückgeliefert wurden, bedeutet dies nichts anderes, als dass die Fähigkeit zur toleranten Verarbeitung von Fehlern hier gar nicht oder nur in geringem Maß vorhanden war. Für MEDPILOT war es zudem sehr schwierig, Synonyme und Komposita sowie grammatische Variationen so aufzulösen, dass ausreichend viele Treffer gefunden werden konnten (unabhängig von der Einschätzung, ob diese auch relevant waren). Dies gelang bei der Testung auf der Grundlage der MEDLINE-Daten zwar etwas besser, doch tendenziell wurden auch hier erheblich weniger Treffer gefunden als durch die Testsuchmaschine.

Das Trefferverhältnis von Averbis und MEDPILOT für die durchschnittliche Anzahl der der Treffer bei der Datenbank CC MED betrug 4,4 : 1. Gegenüber MEDPILOT fand Averbis bei der Suche in der Datenbank CC MED also im Durchschnitt eine 4,4fach höhere Menge an Treffern. Für die Suche nach MEDLINE-Quellen betrug dieses Verhältnis 1,96 : 1. Die Averbis-Testsuchmaschine fand im Schnitt also fast doppelt so viele Treffer wie MEDPILOT.

Der *Anteil der Null-Treffer-Meldungen* (bei 600 Suchanfragen) lag – bezogen auf den MEDLINE-Vergleich – bei MEDPILOT mit 36,1 % mehr als doppelt so hoch wie bei Averbis (16,8 %). In über einem Drittel aller Suchanfragen wurden von MEDPILOT also keine Treffer zurückgemeldet. Bei den Tests mit der Datenbank CC MED wurden bei MEDPILOT in 44,1 % der Suchanfragen keine Treffer gemeldet. Bei der Averbis-Testsuchmaschine lag dieser Anteil lediglich bei 23,8 %.

Tabelle 13. Verteilung der durchschnittlichen Anzahl der Treffermeldungen bei Testung von MEDPILOT und der Averbis-Testsuchmaschine für die Datenbanken MEDLINE und CC MED mit je N=600 Suchanfragen (50 pro Sprachaspekt).

	CC MED		**MEDLINE**	
	Averbis	**MEDPILOT**	**Averbis**	**MEDPILOT**
Rechtschreibfehler	500,62	0,12	77977,04	2,61
Akronyme	20,46	17,78	1117,86	1193,54
Synonyme (Original)	27,92	6,16	12302,92	1609,96
Synonyme (modifiziert)	21,96	0,28	13304,12	251,76
Komposita (Original)	226,38	114,91	39494,08	15397,17
Komposita (zerlegt)	339,40	11,22	60943,02	7142,58
Übersetzung de - en	8609,00	3284,28	699086,68	81517,50
Übersetzung en - de	9417,82	731,18	873778,16	809650,06
Laiensprache	286,40	129,89	24410,52	2385,36
Expertensprache	434,08	233,04	34568,08	14859,04
Grammatik (Original)	9,48	3,02	12460,32	7220,02
Grammatik (Variation)	8,90	2,50	7826,38	7561,90
Durchschnitt (Mittelwert)	**1658,54**	**377,87**	**154772,43**	**79065,96**

Zusammenfassend kann Folgendes festgestellt werden: Durch die Anwendung der MorphoSaurus-Technologie ist eine im hohen Grad sprachunabhängige Verarbeitung fremdsprachlicher Inhalte möglich geworden. Die neue Technik zeigt zudem auch dort ihre Stärken, wo es um die Verarbeitung von Laien- und Expertensprache, die Analyse von Komposita, Synonymen und grammatikalischen Varianten geht. Bei der Verarbeitung von Rechtschreibfehlern und Akronymen zeigte sich in der ersten Version der Averbis-Testsuchmaschine allerdings noch Optimierungsbedarf. Mit der Integration eines Moduls zur Akronymerkennung und zur besseren Verarbeitung von Rechtschreibfehlern konnte die Leistungsfähigkeit zur Auflösung medizinischer Abkürzungen inzwischen jedoch erheblich gesteigert werden. Systematische Evaluationen für diese verbesserte Version der Testsuchmaschine in Bezug auf die genannten Sprachaspekte stehen allerdings noch aus.

5.2.2 Konkurrenzanalyse und Benchmarking

5.2.2.1 Vergleich der Trefferrelevanz auf der Grundlage von MEDLINE-Daten

Die folgenden Ergebnisse beruhen alle auf Tests, die mithilfe einer repräsentativen Testkollektion von 100 Suchtermen durchgeführt wurden. Bei einem direkten Vergleich mit konkurrierenden Suchportalen, bei dem die Suche auf die Datenquelle MEDLINE beschränkt wurde (ca. 16 Mio. Einträge), übertraf die MorphoSaurus-Technologie sogar die Leistung von Google (vgl. Abbildung 15).

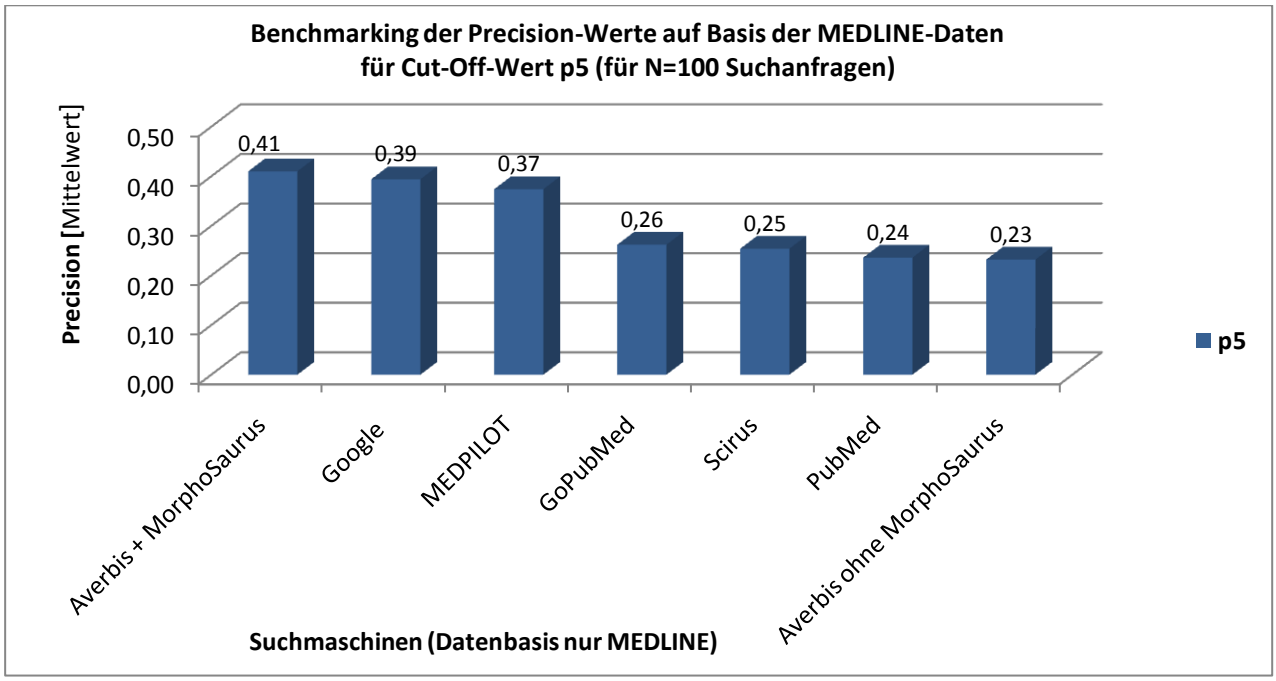

Abbildung 15. Ergebnisse des Benchmarkings für verschiedene Suchmaschinen auf Basis von MEDLINE-Daten. Vergleich der Precision-Werte (Mittelwert) für die ersten fünf Treffer (N = 100 Suchanfragen einer repräsentativen Testkollektion je Suchmaschinentest).

Um abschätzen zu können, welchen Beitrag die MorphoSaurus-Technik für eine verbesserte Verarbeitung der Suchterme leistet, wurden die Untersuchungen mit der Averbis-Testsuchmaschine sowohl mit, als auch ohne aktivierte MorphoSaurus-Technik durchgeführt. Unter den ersten fünf Treffern lag die Trefferrelevanz der Averbis-Testsuchmaschine *mit integrierter MorphoSaurus-Technik* bei 0,41 bzw. bei 41%. *Ohne Aktivierung der Sprachtechnologie* waren es lediglich 23% der ersten fünf Treffer, die als relevant gelten konnten. Dieses Niveau entspricht ungefähr den Werten, die die konkurrierenden Suchmaschinen PubMed (24%), Scirus (25%) und GoPubMed erreichten. Nur MEDPILOT (37%) und Google (39%) kamen auf höhere Werte. Der Einsatz der MorphoSaurus-Technik bewirkte im

Durchschnitt eine Steigerung der Trefferrelevanz-Werte in Höhe von 18%. Die Precision-Werte für sämtliche Cut-Off-Werte des Konkurrenzvergleichs auf Basis von MEDLINE können Tabelle 14 entnommen werden.

Tabelle 14. Benchmarking – Precision-Werte für verschiedene Suchmaschinen auf Basis einer repräsentativen Testkollektion (N=100 Suchterme) für die Cut-Off-Werte p5, p10, p15 und p20. Die alleinige Datenbasis der Suchmaschinentests war MEDLINE (ca. 16 Mio. Einträge).

	Precision / Trefferrelevanz			
Suchmaschine (Datenbasis nur MEDLINE)	**p5**	**p10**	**p15**	**p20**
Averbis + MorphoSaurus (Testsuchmaschine)	0,41	0,39	0,38	0,37
Google	0,39	0,37	0,34	0,33
MEDPILOT	0,37	0,35	0,33	0,32
GoPubMed	0,26	0,25	0,24	0,22
Scirus	0,25	0,23	0,22	0,21
PubMed	0,24	0,20	0,19	0,18
Averbis ohne MorphoSaurus	0,23	0,21	0,21	0,21

5.2.2.2 Vergleich der Trefferrelevanz auf der Grundlage sämtlicher Quellen

Die folgenden Abschnitte beschreiben die Ergebnisse der Tests, die unter Berücksichtigung sämtlicher zur Verfügung stehenden Datenquellen vorgenommen wurden. Mit der Vergrößerung des Index der Testsuchmaschine durch die Integration zusätzlicher Datenbanken erhöhte sich die Trefferrelevanz beträchtlich (vgl. Abbildung 16). Zusätzlich zu den MEDLINE-Daten wurden der ZB MED-OPAC sowie die CC MED-Daten in den Index aufgenommen. Dadurch erhöhte sich der Umfang von 16 auf 17 Mio. Einträge.

Im Durchschnitt verbesserte sich die Trefferrelevanz dadurch von 42% ohne die MorphoSaurus-Technologie auf 58% (bei p5) mit MorphoSaurus-Technologie. Dies bedeutet einen Zugewinn an Trefferrelevanz von 16% unter den ersten fünf Treffern. Dies galt allerdings nur, wenn die Suche im Feld „Titel“ durchgeführt wurde. Bei einer Suche über alle Felder wurde bei p5 eine Trefferrelevanz von 0,54 bzw. 54% erreicht. Offenbar wirkte sich die Einschränkung auf die Titelsuche positiv auf die Trefferrelevanz aus.

Bei leicht vergrößerter Datenbasis (von 16 auf 17 Mio.) wurde also nochmals ein beträchtlicher Sprung in der Trefferrelevanz erreicht. Die erneute Zunahme deutet auf einen

stabilen Effekt des MorphoSaurus-Algorithmus hin: Je größer die indexierte Datenbasis ist, auf die der Algorithmus für seine Analyse zurückgreifen kann, desto höher fällt die Trefferrelevanz aus. Eine systematische Prüfung in diesem Punkt steht allerdings noch aus.

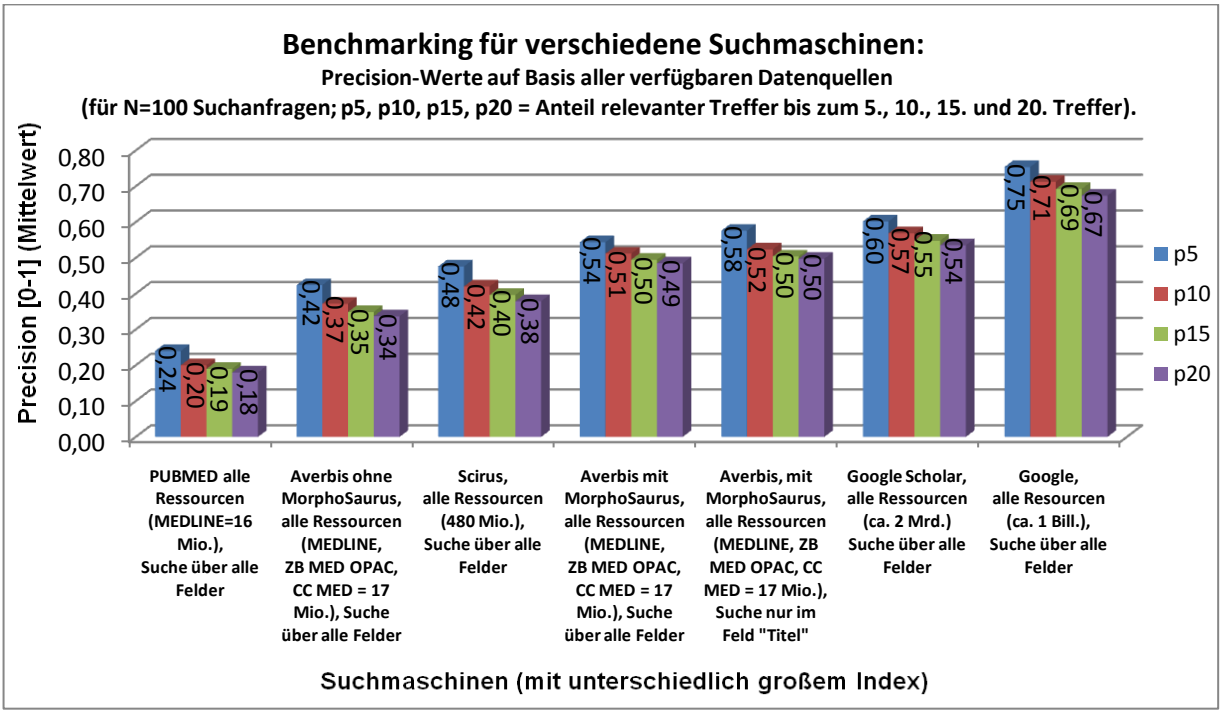

Abbildung 16. Vergleich der Precision-Werte (Mittelwert) für die ersten fünf, 10, 15 und 20 Treffer (N = 100 Suchanfragen) auf der Basis aller verfügbaren Datenquellen.

Damit ist das neue System schon jetzt fast so gut wie Google Scholar (p5 = 0,6). Im Vergleich zu Google Scholar, dessen Datenbasis auf ca. 2 Mrd. Dokumente geschätzt wird, besaß die Datenbasis der Testsuchmaschine lediglich ca. 17 Mio. Titel. Google erreichte die höchste Trefferrelevanz (p5 = 0,75). Es hat aber auch schätzungsweise bis zu einer Billion URLs indexiert (Alpert & Hajaj, 2008). Der Index von Scirus umfasst – nach eigener Angabe – ca. 480 Millionen Dokumente und erreichte damit einen Precision-Wert von p5 = 0,48. Zum Vergleich wurden in der obigen Grafik auch die PubMed-Daten aufgenommen. GoPubMed greift lediglich auf die bei PubMed angebotenen MEDLINE-Daten zurück (ca. 16 Mio.) und wurde deshalb in Abbildung 16 nicht gesondert aufgeführt.

5.2.2.3 Null-Treffer-Meldungen

Ein weiterer Indikator für die Leistungsfähigkeit eines Retrieval-Systems ist die Anzahl der Null-Treffer-Meldungen. Die Anzahl der potenziell relevanten Treffer hängt im Kern von zwei Einflussgrößen ab: erstens vom Umfang des Gesamtindex und zweitens von der

Mächtigkeit des Algorithmus für die Verarbeitung der Suchanfragen. Je größer der Gesamtindex ist, desto wahrscheinlicher ist es auch, dass relevante Treffer gefunden werden und je besser der Algorithmus für die Verarbeitung von Suchanfrage und Auswahl der Treffer funktioniert, desto weniger Null-Treffer-Meldungen werden als Ergebnis einer Suchanfrage angezeigt. Mit der Averbis-Technik gelingt es – mit einer im Vergleich zu Google relativ kleinen Datenbasis – die Quote der Null-Treffer-Meldungen relativ gering zu halten. Abbildung 17 zeigt einen Vergleich der Null-Treffer-Meldungen für die 100 Suchanfragen der repräsentativen Testkollektion. Hierbei wurden zwei Testsituationen betrachtet: 1. Vergleiche auf Basis der reinen MEDLINE-Daten (graue Balken) und 2. der Vergleiche unter Berücksichtigung aller verfügbaren Datenquellen (schwarze Balken).

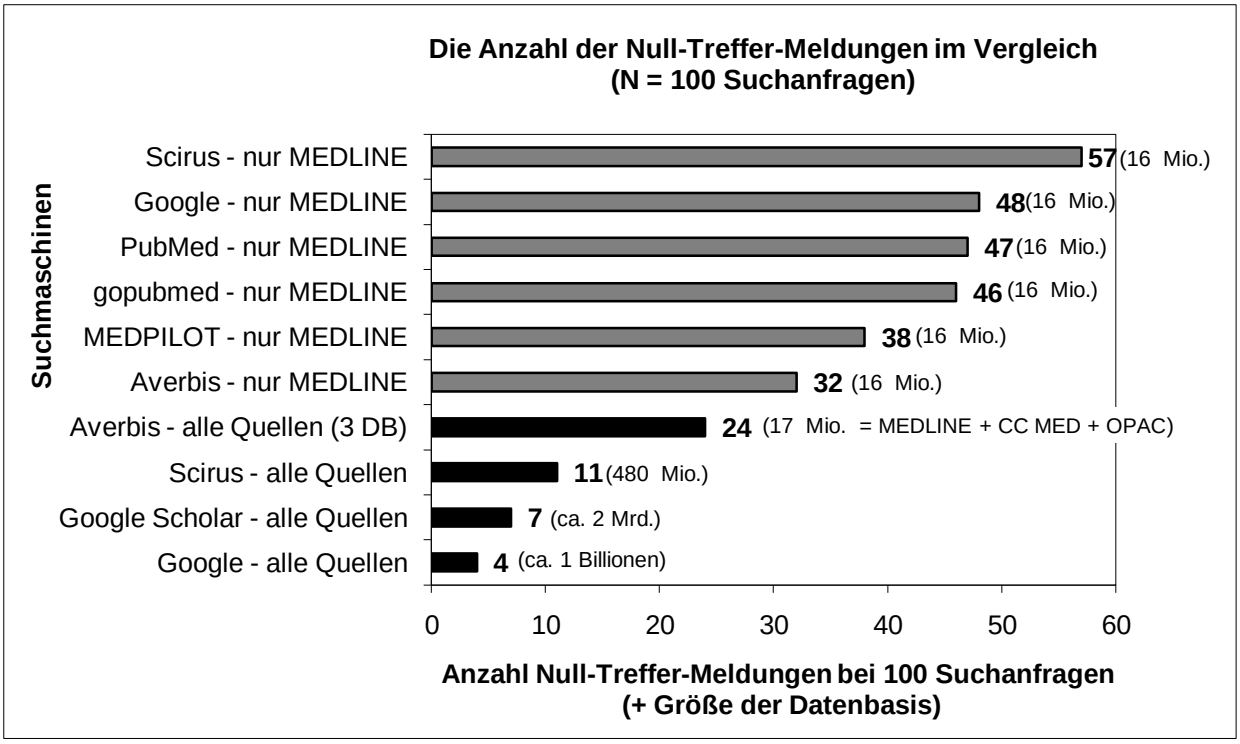

Abbildung 17. Anzahl der Null-Treffer-Meldungen im Vergleich. In Klammern: Datenbestand des Index. Graue Balken: Vergleich nur auf Grundlage von MEDLINE-Daten; schwarze Balken: Vergleich unter Berücksichtigung aller verfügbaren Daten (N = 100 Suchanfragen).

Bei gleicher Datengrundlage (also MEDLINE) schnitt der Averbis-Algorithmus am besten ab: Von 100 Suchanfragen der repräsentativen Testkollektion wurden nur 32 mit Null-Treffer-Meldungen quittiert (also 32%). Beim reinen MEDLINE-basierten Vergleich wies Google 48 Null-Treffer-Meldungen auf (48%). Bei Nutzung sämtlicher verfügbarer Datenquellen für den Index blieb Google bei der Frage der Null-Treffer-Meldungen jedoch ungeschlagen: Nur auf vier von 100 Suchanfragen konnte Google überhaupt keine Treffer liefern. Die Averbis-

Testsuchmaschine verbesserte sich jedoch ebenfalls: Durch die Hinzunahme von ca. einer weiteren Millionen Datenbankeinträgen sank die Anzahl der Null-Treffer-Meldungen von 32 auf 24. Die Ergebnisse der Retrieval-Tests und des Benchmarkings werden in Kapitel 6 diskutiert.

5.2.2.4 Durchschnittliche Treffermenge

Ein Blick auf die durchschnittliche Treffermenge, die bei einer Suchanfrage zurückgemeldet wird, zeigt uns das Potenzial der Suchmaschinen. Zwar besteht zwischen dem Umfang der Treffermenge und der Anzahl der relevanten Treffer kein zwingender Zusammenhang, doch korrelieren diese beiden Größen miteinander. Wie bereits erwähnt, erhöht sich mit der Größe der Datenbasis auch die Wahrscheinlichkeit relevante Treffer zu finden. Beim Vergleich auf Basis der MEDLINE-Daten zeigte sich die Averbis-Suchumgebung als äußerst effektiv (graue Balken in Abbildung 18).

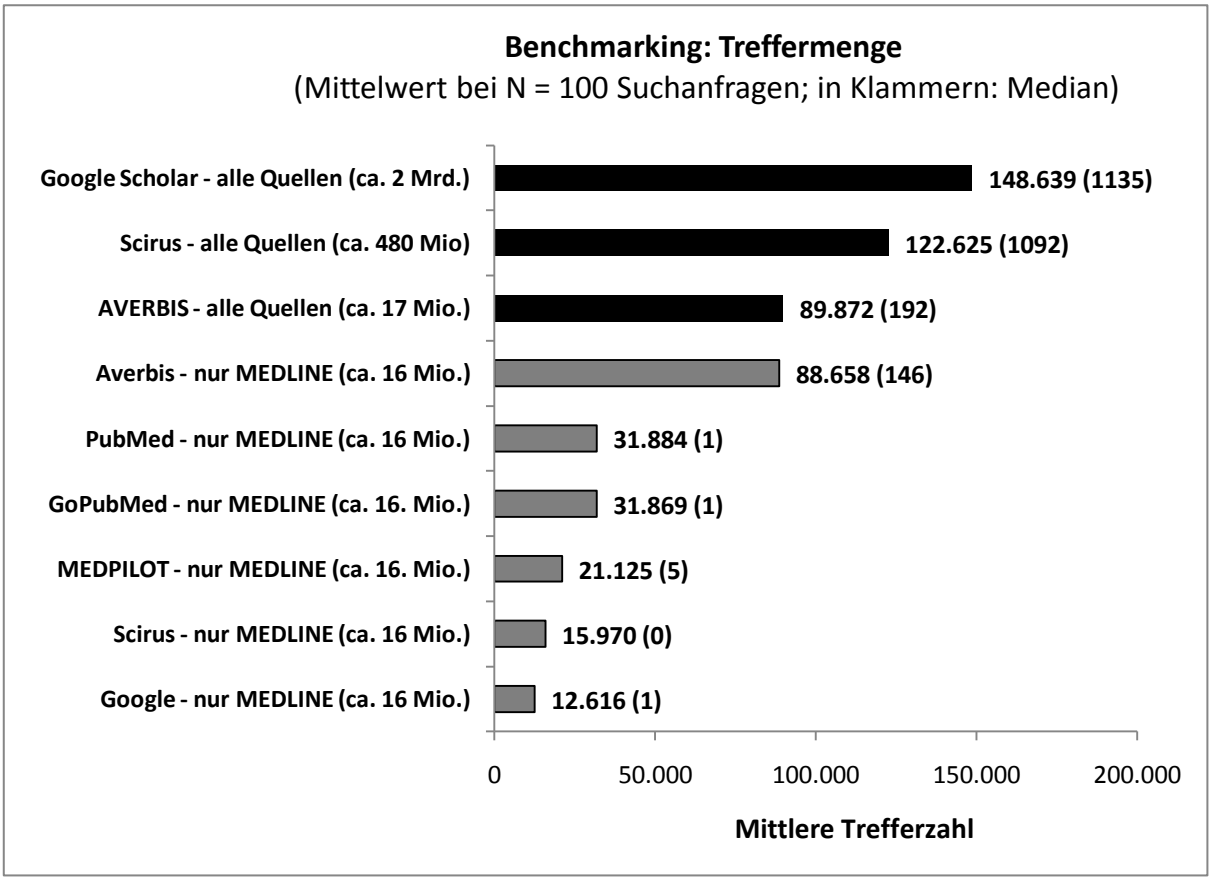

Abbildung 18. Vergleich der durchschnittlichen Treffermenge bei 100 Suchanfragen (Test mit repräsentativer Testkollektion). Graue Balken: Vergleich nur auf Grundlage von MEDLINE-Daten; schwarze Balken: Vergleich unter Berücksichtigung aller verfügbaren Daten (N = 100 Suchanfragen)

Von allen getesteten Suchmaschinen, meldete die Averbis-Suche die meisten Treffer. Bei 100 Suchanfragen waren dies im Schnitt 89872 Treffer pro Suchanfrage. Dieser Wert allein ist noch nicht sehr aussagekräftig. Da die Treffermenge in Abhängigkeit von der einzelnen Suchanfrage sehr streut, bildet der Mittelwert nur ein sehr grobes Maß zur Einschätzung der Treffereffektivität. Einen größeren Aussagewert besitzt hier der Median (Zentralwert) der Treffermenge. Dieser Kennwert beschreibt die Zahl, bei der 50% einer beobachten Stichprobe oberhalb und 50% unterhalb dieses Wertes liegen. Anders als das arithmetische Mittel besitzt der Median den Vorteil, robust gegenüber Extremwerten zu sein. PubMed findet bei 100 Suchanfragen zwar im Mittel 31884 Treffer, weist aber einen Median von 1 auf (in Abbildung 18: Zahl in Klammern hinter der Treffermenge). Das bedeutet, bei 50 von 100 Suchanfragen wurde höchstens ein Treffer gefunden. Insofern ist der Median hier ein besserer Indikator für die Bewertung der Retrieval-Leistung als der Mittelwert. Auch hier schneidet Averbis sehr gut ab: Der Median liegt bei 146 Treffern und ist damit der höchste Wert im Testfeld.

Zum Vergleich wurden in Abbildung 18 auch Tests mit Suchmaschinen widergegeben, die auf einen wesentlich größeren Index zurückgreifen können (schwarze Balken). Scirus erreichte hier bei einer Indexgröße von 480 Mio. einen Median von 1092 Treffern. Die Averbis-Testsuchmaschine kam dagegen mit 17 Mio. Dokumenten auf einen Median von 192. Obwohl der Index von Scirus rund 28-mal mehr Daten enthält, ist der Median-Wert nur ca. 5,7mal so groß wie der von Averbis. Diese Verhältnis von Indexgröße und erreichtem Medianwert macht nochmals deutlich, über welches Potenzial die Averbis-Technik verfügt, wenn die Indexgröße weiter ausgebaut wird.

5.3 Ergebnisse der Usability-Untersuchung

Im folgenden Kapitel werden die Ergebnisse der Usability-Untersuchung vorgestellt. Neben den Resultaten für die globale Einschätzung der Benutzerfreundlichkeit werden die Ergebnisse der Fragebogenteile, die Ergebnisse des szenariobasierten Usability-Tests sowie die Resultate für das fokussierte Interview berichtet. Die Reihenfolge der berichteten Ergebnisse orientiert sich dabei am Ablauf der Usability-Untersuchung (vgl. Kap. 4.3.2). Die statistischen Überprüfungen wurden auf dem üblichen Signifikanzniveau von fünf Prozent durchgeführt. Eingesetzt wurden t-Tests und einfache Varianzanalysen zur Testung der Mittelwertunterschiede sowie Chi-Quadrat-Tests zur Überprüfung auf Verteilungsunterschiede.

5.3.1 Globale Bewertung der Averbis-Testsuchmaschine

Fast alle Probanden haben das Portal der Testsuchmaschine positiv bewertet und konnten sich eine regelmäßige Nutzung dieser Plattform vorstellen: Eine Befragung im Anschluss an die Untersuchung ergab, dass sich von den 24 getesteten Probanden 20 Personen (83,3%) zukünftig eine regelmäßige Nutzung des MEDPILOT-Portals vorstellen konnten. Vier Testpersonen (16,7%) konnten sich dies „mit Einschränkungen“ vorstellen. Die Usability wurde vom überwiegenden Teil der Testpersonen ebenfalls sehr positiv bewertet. Dieses Fazit leitet sich zum einen aus den Ergebnissen des Usability-Tests ab, zum anderen aus den guten Bewertungen hinsichtlich verschiedener Usability-Dimensionen (vgl. Kap. 5.3.4) und den Erkenntnissen, die durch das fokussierte Interview gewonnen wurden (vgl. Kap. 5.3.3.4).

5.3.2 Demografische Variablen und Selbsteinschätzungsskalen

Durch einen Fragebogen wurden zu Beginn der Untersuchung Daten zur Demografie und zur Selbsteinschätzung der Internet- bzw. Literaturrecherchekompetenz erhoben. Zudem wurde gefragt, ob die Testpersonen MEDPILOT in der bisherigen Form bereits kannten und nutzten.

Demografische Variablen. Die Probandengruppe bestand aus 24 Medizinern mit unterschiedlichem Expertise-Status. Das heißt, die Gruppe bestand aus 12 berufstätigen Ärzten mit Hochschulabschluss sowie 12 Studierenden der Medizin, die sich im Durchschnitt im neunten Semester ihrer Ausbildung befanden (alle nach dem Physikum). Die Spannweite reichte dabei vom fünften bis zum vierzehnten Semester. Von den 12 berufstätigen Ärzten waren 10 in der Patientenbetreuung tätig und zwei in Forschung und Lehre. Die beruflichen Schwerpunkte verteilten sich auf ein breites Spektrum medizinischer Fachgebiete. Kein Fachgebiet war übermäßig häufig vertreten. Insgesamt setzte sich die Probandengruppe zu gleichen Teilen aus weiblichen wie männlichen Personen zusammen (vgl. Tabelle 15). Der Altersdurchschnitt der Testpersonen lag bei 28,25 Jahren (SD = 4,9), wobei die Studierenden im Schnitt 26 Jahre alt waren (SD = 5,1) und die Ärzte mit Abschluss 30,4 Jahre (SD = 3,94).

Tabelle 15. Stichprobe der Probanden für den Usability-Test (N=24).

		Geschlecht	
		weiblich	männlich
Expertise-Status	Ärzte mit Abschluss	n = 6	n = 6
	Studierende der Medizin (nach dem Physikum)	n = 6	n = 6

Bekanntheit von MEDPILOT. 16 von 24 Probanden kannten MEDPILOT bereits vor der Untersuchung als medizinische Suchmaschine (66,7%). Einem Drittel der Testpersonen (acht Probanden bzw. 33,3%) war MEDPILOT bisher nicht bekannt, wobei vier dieser acht Personen aus der Gruppe der Ärzte und vier aus der Gruppe der Studierenden stammten.

Nutzungsintensität. Zwei Testpersonen (8,3%) nutzten die Suchmaschine mindestens einmal pro Woche (ausschließlich Ärzte). Drei Probanden (12,5%) nutzten das Portal ein- bis zweimal im Monat und 11 Testpersonen (45,8%) setzten MEDPILOT seltener als einmal pro Monat ein, wobei sich keine signifikanten Unterschiede zwischen Ärzten und Studierenden zeigten (vgl. Abbildung 19).

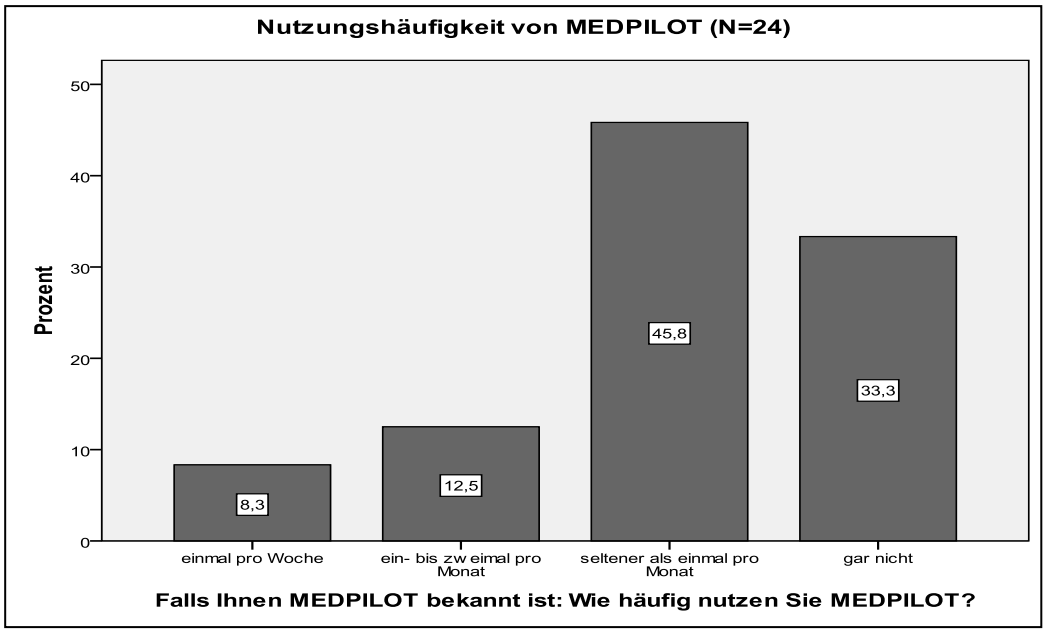

Abbildung 19. Nutzungshäufigkeit von MEDPILOT in Prozent (N=24).

Selbsteinschätzungsskalen. Im Fragebogen wurden u.a. eine Reihe von Variablen erhoben, die sich auf die *Selbsteinschätzung* der Probanden bezüglich ihrer Interneterfahrung und ihrer Internetkompetenz bezogen. Insgesamt nutzten die befragten Testpersonen das Internet seit M = 10,63 Jahren (SD = 2,32). Während die Ärzte im Schnitt bereits seit 11,3 Jahren Internetnutzer waren (SD = 2,06), gaben die Studierenden an, dass sie dieses Medium im Durchschnitt seit 9,92 Jahren nutzten (SD = 2,43). Bei der statistischen Überprüfung erwies sich dieser Unterschied als nicht signifikant (p = .14). Im Mittel lag die wöchentliche Nutzungsdauer des Internets ohne E-Mail-Nutzung bei den Ärzten bei 11,33 Stunden (SD = 6,41).

Die studentischen Testpersonen kamen hier auf einen Wert von 7,75 Stunden pro Woche (SD = 3,02). Der Unterschied war statistisch jedoch nicht signifikant (p = .09).

Selbsteinschätzung der Internet- und Literaturrecherchekompetenz. Auf einer Skala von 1 = eher gering bis 7 = eher hoch, stuften die Testpersonen ihre *Internetkompetenz* mit einem durchschnittlichen Wert von M = 4,79 (SD = 1,22) ein (vgl. Tabelle 16). Auf die Skala bezogen bedeutet dies, dass die Probanden ihre Kompetenz leicht über dem Durchschnitt einordneten, wobei die Studierenden sich als etwas kompetenter einschätzten (M = 5,00; SD = 1,35) als die bereits berufstätigen Ärzte (M = 4,79; SD = 1,22). Die Unterschiede waren jedoch nicht signifikant, p > .05.

Tabelle 16. Internet- und Recherchekompetenz der Probanden für den Usability-Test (N = 24).

	Studierende (der Medizin nach dem Physikum; n=12)		**Ärzte (mit abgeschl. Medizinstudium; n=12)**		**gesamt (N=24)**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Seit wie vielen Jahren benutzen Sie schon das Internet?	9,92	2,43	11,33	2,06	10,63	2,32
Wie viele Stunden pro Woche nutzen Sie das Web (die mit E-Mails verbrachte Zeit nicht mitgerechnet)?	7,75	3,02	11,33	6,41	9,54	5,23
Wie würden Sie Ihre eigene Internetkompetenz einschätzen? (von 1=eher gering bis 7=eher hoch)	5,00	1,35	4,58	1,08	4,79	1,22
Wie würden Sie speziell Ihre Fähigkeit zur Literaturrecherche einschätzen? (von 1=eher gering bis 7=eher hoch)	3,58	1,16	4,33	1,30	3,96	1,27
Wie schätzen Sie Ihre Kompetenz im Umgang mit Suchmaschinen ein? (von 1=eher gering bis 5=eher hoch)	3,08	0,67	3,25	0,75	3,17	0,70

M = Mittelwert, SD = Standardabweichung

Für die Selbsteinschätzung hinsichtlich der *Kompetenz zur Literaturrecherche* wurde für die Gesamtstichprobe ein Wert von M = 3,96 (SD = 1,27) festgestellt. Die Auswertung für die Studierenden ergab einen Wert von M = 3,58 (SD = 1,16). Die Ärzte stuften ihre Kompetenz mit M = 4,33 (SD = 1,3) etwas höher ein. Sie lagen damit aber ebenfalls nur knapp über dem Mittel der Skala (von 1=eher gering bis 7= eher hoch), p > .05.

Kompetenz im Umgang mit Suchmaschinen. Auf einer Skala von 1 = eher gering bis 5 = eher hoch lag der Wert für die Gesamtstichprobe bei M = 3,17 (SD = 0,7). Der Mittelwert für die Studierenden lag bei M = 3,08 (SD = 0,67) und für die Ärzte bei M = 3,25 (SD = 0,75). Damit bewegten sich diese Werte über dem Skalendurchschnitt. Die Unterschiede zwischen den beiden Subgruppen waren statistisch allerdings nicht signifikant, p > .05.

Nutzer mit wenig und mit viel Erfahrung. Aufgrund der Erkenntnisse zur Expertenforschung wurde davon ausgegangen, dass erfahrene Internetnutzer effektiver und effizienter im Gebrauch von Suchmaschinen sind als Laien bzw. unerfahrene Nutzer. Es wurde erwartet, dass die Anzahl korrekt beantworteter Fragen bei den Experten höher ausfallen würde als bei Usern, die eher als wenig erfahren gelten. Dass Nutzer mit mehr Erfahrung größere Erfolgsraten bei der Lösung von Standardaufgaben im Web aufweisen, konnten z.B. auch Nielsen und Loranger (2006) bestätigen. Um zwischen „Nutzern mit wenig Erfahrung" und „Nutzern mit viel Erfahrung" valide unterscheiden zu können, hat es sich in der Onlineforschung als sehr hilfreich erwiesen, dass die Nutzer – über globale Maße der Selbsteinschätzung hinaus – nach *konkreten Verhaltensweisen* im Umgang mit dem Internet und dem PC befragt werden (vgl. Tabelle 17).

Tabelle 17. Interneterfahrung der Probanden (N=24). Indikatoren nach Nielsen & Loranger (2006).

	Studierende (der Medizin nach dem Physikum; n=12)		**Ärzte (mit abgeschl. Medizinstudium; n=12)**		**gesamt (N=24)**	
	ja	**%**	**ja**	**%**	**ja**	**%**
Benutzen Sie das Web zum Chatten?	5	20,8	3	12,5	8	33,3
Benutzen Sie Bookmarks (Favoriten)?	8	33,3	6	25,0	14	58,3
Nehmen Sie selbstständig Browserupgrades vor?	7	29,2	6	25,0	13	54,2
Können Sie Webseiten erstellen?	2	8,3	3	12,5	5	20,8
Können Sie Computerprobleme selbstständig beheben?	5	20,8	7	29,2	12	50,0
Halten Sie sich über die Technik am Laufenden? (Haben Sie beispielsweise Computermagazine abonniert?)	3	12,5	5	20,8	8	33,3
Werden Sie von Freunden um Rat gefragt, wenn es um Computerprobleme geht?	4	16,7	7	29,2	11	45,8

So gehen Nielsen und Loranger (2006) davon aus, dass sich anhand dieser Praxiskompetenzen im Umgang mit dem Internet valide Subgruppen von „unterschiedlich erfahrenen" Internetusern identifizieren lassen. Um diese Annahme auch mit den Testpersonen der Usability-Untersuchung überprüfen zu können, sind die Probanden in Anlehnung an Nielsen und Loranger (2006, S. 23) in zwei Gruppen mit unterschiedlich hoher Interneterfahrung aufgeteilt worden. Als „wenig erfahren" wurden die Probanden eingestuft,

> „… wenn sie seit höchstens drei Jahre online waren, das Web weniger als zehn Stunden pro Woche nutzten, weniger als ein Drittel der fortgeschrittenen Verhaltensweisen zeigten, andere um die Behebung ihrer Computerprobleme baten und wenn sie nicht wegen Computerproblemen um Rat gefragt wurden" (Nielsen & Loranger, 2006, S. 23).

Zu der Gruppe der Nutzer „mit viel Erfahrung“ wurden die Testpersonen gezählt,

> „… wenn sie seit mindestens vier Jahren online waren, das Web seit mehr als zehn Stunden pro Woche nutzen, mehr als ein Drittel der genannten fortgeschrittenen Verhaltensweisen zeigten, ihre Computerprobleme selbst lösten und als Informationsquelle für andere dienten“ (Nielsen & Loranger, 2006, S. 23).

Für die Gesamtstichprobe von N = 24 zeigte sich folgendes Bild: Insgesamt konnten acht Personen mit „viel Erfahrung“ (33,3%) und 16 Personen mit eher „wenig Erfahrung“ (66,7%) identifiziert werden. Von den 16 Testpersonen mit wenig Interneterfahrung waren 10 Studierende und sechs Ärzte. Unter den acht Probanden mit viel Erfahrung waren sechs Ärzte und zwei Studierende. Tendenziell wies die Gruppe der acht Interneterfahrenen eine etwas intensivere Nutzung des Internets auf. Diese Gruppe hielt sich auch für kompetenter im Umgang mit dem Internet ($M = 5,75$; $SD = 0,89$ vs. $M = 4,31$; $SD = 1,08$), $p < .05$. Sie schätzte auch ihre Fähigkeit zur Literaturrecherche höher ein ($M = 4,88$; $SD = 0,99$) als die Gruppe mit weniger Erfahrung ($M = 3,5$; $SD = 1,15$), $p < .05$. Als ebenfalls signifikant erwies sich der Unterschied bezüglich der Einschätzung der eigenen Kompetenz im Umgang mit Suchmaschinen (Interneterfahrene: $M = 3,63$; $SD = 0,52$ vs. Internetunerfahrene: $M = 2,94$; $SD = 0,68$), $p < .05$. In Kapitel 5.3.3.3 werden die Ergebnisse für die Effizienz und Effektivität von Aufgabenlösungen in Abhängigkeit von der Interneterfahrung berichtet. Im Übrigen zeigten sich im direkten Vergleich für *die fortgeschrittenen Verhaltensweisen* keine signifikanten Unterschiede zwischen der Gruppe der Ärzte und der Studierenden (Chi-Quadrat-Test), $p > .05$.

5.3.3 Szenariobasierter Usability-Test

5.3.3.1 Erstkontakt mit der Website

Nach einem Erstkontakt von 30 Sekunden mit der Eingangsseite der Testsuchmaschine sollten die Testpersonen drei Fragen beantworten, die wichtige Informationen zum Verständnis der Website betrafen (vgl. Kap. 4.3.2):

1. Wer betreibt die Website? Richtige Antwort: „ZB MED“
2. Welche Art von Informationen können Sie hier recherchieren? Richtige Antwort: „medizinische Informationen“
3. Was können Sie mit MEDPILOT neben der Recherche noch tun? Richtige Antwort: „Durchführen von Bestellungen“ und „Aufruf von Volltexten“ (oft sogar kostenlos).

Nach der 30 sekündigen Einblendung der Einstiegsseite erfolgte ein einfacher Gedächtnis-Recall-Test zu den oben genannten Fragen. Die Ergebnisse für diesen Gedächtnistest sind in den Abbildungen 20 bis 22 zusammengefasst.

Zu Frage 1: Die Frage nach dem Anbieter der Website wurde von 58 % der Testpersonen richtig erinnert. 17% beantworteten diese Frage teilweise richtig. In diesen Fällen wurden beispielsweise neben der ZB MED noch andere Einrichtungen genannt. Von 25% der Probanden wurde die Frage nach dem Website-Betreiber eindeutig falsch beantwortet. Entweder konnten sich die Testpersonen nicht erinnern oder sie nannten einen falschen Namen (Abbildung 20).

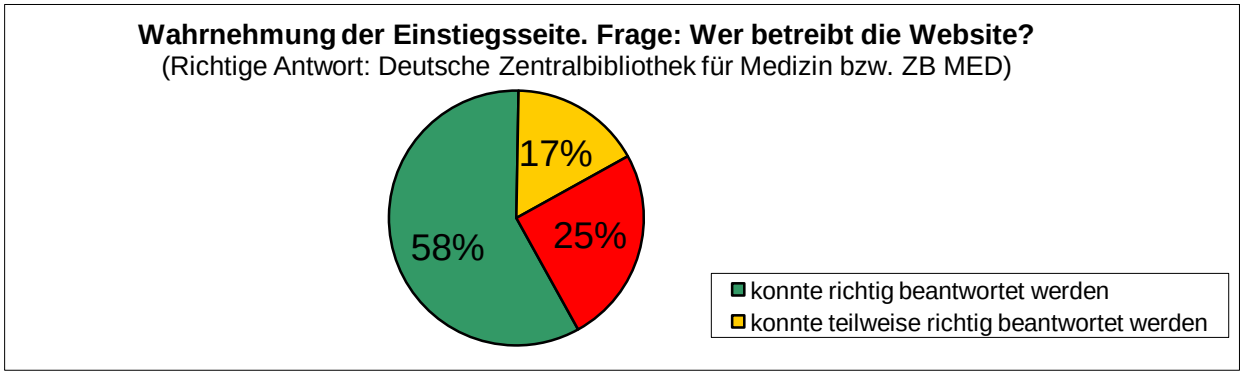

Abbildung 20. Wer betreibt die Website?

Zu Frage 2: Welche Art von Informationen können Sie hier recherchieren? Zwei Drittel (66,7%) der Testpersonen konnten diese Frage korrekt beantworten und wussten, dass die Website medizinische Informationen anbietet. 33% der Probanden konnten die Frage zumindest teilweise richtig beantworten. Keine der 24 Testpersonen gab hier eine eindeutig falsche Antwort (Abbildung 21).

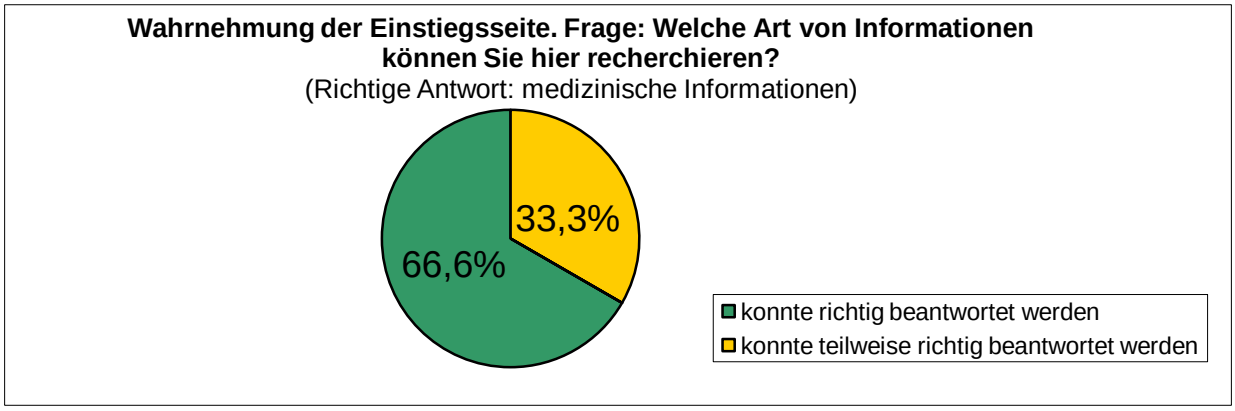

Abbildung 21. Welche Art von Informationen können Sie hier recherchieren?

Zu Frage 3: Auf die Frage, was man mit MEDPILOT neben der reinen Recherche noch tun kann, gaben 46% der Testpersonen richtige Antworten. Teilweise richtig waren die Antworten von 17% der Probanden und gänzlich falsch oder gar nicht beantwortet wurde diese Frage von 37% der Testpersonen (Abbildung 22).

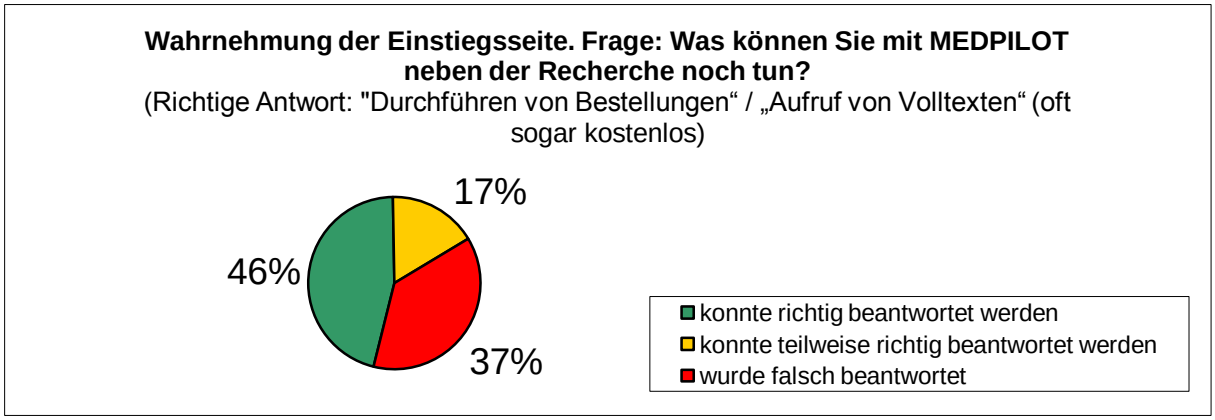

Abbildung 22. Was können Sie mit MEDPILOT neben der Recherche noch tun?

5.3.3.2 Explorative Aufgabe

Ein wichtiges Ziel der Untersuchung war die Klärung der Frage, ob die Unterstützungsfunktionen, die zur Verbesserung der Usability eingeführt wurden, den Nutzern überhaupt auffallen und tatsächlich auch genutzt werden. Während des freien Explorierens, bei dem die Testpersonen eine Frage aus einem Inhaltsbereich recherchieren sollten, der sie interessiert und bei dem sie sich gut auskannten, wurde von einer Protokollantin über einen zweiten Bildschirm beobachtet und festgehalten, ob die Testpersonen die Unterstützungsfunktionen benutzten oder zumindest während des „lauten Denkens“ erwähnten. Damit konnte indirekt darauf geschlossen werden, ob die Testpersonen die Funktionen auch wahrnahmen. Mit dieser Methode ist allerdings nicht direkt messbar, ob und was eine Versuchsperson tatsächlich gesehen bzw. wahrgenommen hat. Es ist auch nicht möglich, mit letzter Sicherheit festzustellen, dass ein Nutzer etwas nicht gesehen hat. Dennoch ist es ein vertretbarer Ansatz, wenn ein Eyetracker nicht zur Verfügung steht. Die Testpersonen hatten für die freie Exploration fünf Minuten Zeit. Dabei interessierten folgende Fragen:

1. Wurde die *MeSH-ICD-10-Autosuggest-Funktion* wahrgenommen und benutzt?
2. Wurde die *Schiebereglerfunktion zur Suchraumeingrenzung* wahrgenommen und benutzt?

3. Wurde die Funktion *Verwandte Suchbegriffe* wahrgenommen und benutzt?
4. Wurde die *Funktion zur Datenbankfilterung* wahrgenommen und benutzt?
5. Wurden die *Optionsfelder bei Sprache/Sortierung* wahrgenommen und benutzt?

Ergebnisse der Beobachtung der freien Exploration:

(1) *MeSH-ICD-10-Autosuggest-Funktion.* Sieben Testpersonen (29,2%) fiel die Vorschlagsfunktion auf. Fünf der sieben Personen, denen die Funktion aufgefallen war, benutzten sie auch tatsächlich (20,8 %).

(2) Schiebereglerfunktion zur Suchraumeingrenzung. 20 von 24 Personen (83,3%) ist der Schieberegler während der freien Exploration aufgefallen; genutzt wurde der Schieberegler von 19 Testpersonen (79,2%).

(3) *Verwandte Suchbegriffe*. Etwas mehr als der Hälfte der Testpersonen (54,3% = 13 Personen) ist die Funktion *Verwandte Suchbegriffe* aufgefallen. Während ihrer Recherche haben diese Funktion aber nur 7 Personen benutzt (29,2%).

(4) *Optionsfelder bei Sprache/Sortierung.* 14 Personen (58,3%) fielen diese Optionsfelder auf, wobei nur 10 Personen diese Optionsmöglichkeiten tatsächlich benutzten (41,7%).

(5) *Funktion zur Datenbankfilterung.* Die Funktion zur Datenbankauswahl, die sich rechts auf der Suchseite in einem grafisch abgegrenzten Kasten befand, nahmen nur neun Personen wahr (37,5%). Letztlich haben nur vier der 24 Testpersonen (16,7%) die Funktion zur Datenbankauswahl bzw. -filterung während der freien Exploration auch aktiv für ihre Recherche eingesetzt.

5.3.3.3 Die Rechercheaufgaben: Messung von Effektivität und Effizienz

Zur Messung der Effektivität des Informationsdesigns wurden den Testpersonen drei Recherchefragen mit unterschiedlichem Schwierigkeitsgrad gestellt. Um auch Aussagen zur Effizienz des Lösungsweges und damit auch indirekt zur Effizienz des Informationsdesigns der Website machen zu können, wurde zusätzlich die Zeit erfasst, die die Testpersonen zur Beantwortung der Fragen benötigten. Die Testpersonen lösten im Durchschnitt 2,04 von drei Fragen (SD = 0,55, N = 24). 66,7% der Probanden konnten *die leichte Recherchefrage* innerhalb der vorgegebenen Zeit von fünf Minuten beantworten. Bei der *mittelschweren Frage* waren dies 79,2% und bei der *schwersten Frage* konnten 14 Probanden (58,5%) die richtige Antwort finden (vgl. Tabelle 18).

Tabelle 18. Anteil der Probanden, die die Recherchefragen richtig oder falsch beantworteten (N=24).

	Frage richtig beantwortet		Frage falsch beantwortet		Zeit zur Beantwortung [in Sekunden]	
	n	%	n	%	M	SD
1. **Leichte Frage:** Welches ist die neueste in der ZB MED vorhandene Auflage von Pschyrembel, Willibald: Klinisches Wörterbuch? Richtige Antwort: „261. Auflage von 2007“.	**16**	**66,7**	**8**	**33,1**	**86,44**	**45,17**
2. **Mittelschwere Frage:.** Welche Erkrankung kann im Verbund mit Nasenpolypen und Aspirin-Intoleranz häufiger auftreten? Richtige Antwort: „Asthma“ od. „Asthma bronchiale“.	**19**	**79,2**	**5**	**20,8**	**127,68**	**43,80**
3. **Schwere Frage:** Wie viele Gene codieren für die Geschmacksrezeptoren der Fruchtfliege (Drosophila)? Hier waren mehrere richtige Antworten möglich: Je nach dem, aus welchem Jahr der recherchierte Artikel stammt, unterscheidet sich der Wissensstand bezüglich der Anzahl der Gene. Folgende Antworten wurden als korrekt gewertet: - 56 Gene, Artikel aus 2003 - 60 Gene, Artikel aus 2004 - 68 Gene, Artikel aus 2007 - 68 Gene, Artikel aus 2008	**14**	**58,3**	**10**	**41,7**	**127,21**	**65,69**

Die leichte Frage wurde im Durchschnitt schneller beantwortet als die mittelschwere oder schwere Frage. Die Probanden benötigten im Mittel 86,44 Sekunden zur Beantwortung der leichten Frage. Für die mittelschwere Frage brauchten sie 127,68 Sekunden und für die schwere Frage 127,21 Sekunden.

Diese obigen Werte beziehen sich auf diejenigen Testpersonen, die in der Lage waren, die Aufgaben zu lösen. Die Zeitaufwände der Personen, die die Aufgabe nicht schafften, überstiegen die angesetzte Zeitgrenze von fünf Minuten (300 Sek.) und wurden aus der Berechnung ausgeschlossen. Die relativ hohen Werte in der Standardabweichung weisen auf eine hohe Varianz in den Daten hin. Offenbar gibt es innerhalb der Gruppe der Testpersonen größere Unterschiede im Ausmaß der zur Lösung der Recherchefrage benötigten Zeit.

Einfluss der Interneterfahrung. Die Hypothese, dass interneterfahrene Nutzer Aufgaben im Web besser lösen können als unerfahrene Nutzer, konnte bestätigt werden: Die 16 eher unerfahrenen Internetnutzer lösten im Schnitt 1,87 (SD = 0,5) Aufgaben. Der Durchschnitt der acht erfahrenen Internetnutzer lag mit M = 2,38 (SD = 0,52) gelösten Aufgaben signifikant darüber, $F(1,23) = 5.21$, $p < .05$. In Bezug auf die benötigte Zeit zur Lösung aller drei Aufgaben konnte kein Unterschied zwischen erfahrenen (M = 109; SD = 42,77) und unerfahrenen Internetnutzern (M = 110,5; SD = 40,67) gefunden werden. Bei der Analyse der einzelnen Aufgaben zeigten sich ebenfalls keine statistisch signifikanten Unterschiede (alle Vergleiche: $p > .05$; s. Tabelle 19).

Tabelle 19. Vergleich der Lösungszeiten zwischen interneterfahrenen und unerfahrenen Nutzern. Durchschnittliche Antwortzeiten (in Sekunden) zur Beantwortung der drei Recherchefragen. Es wurden nur Antwortzeiten von Testpersonen berücksichtigt, die die Fragen in der vorgegebenen Zeit von 300 Sekunden auch korrekt beantworteten.

	Interneterfahrung	**N**	**M**	**SD**
Zeit für die leichte Rechercheaufgabe (in Sek.)	eher gering	9	70,67	26,11
	eher hoch	7	106,71	57,8
Zeit für die mittelschwere Rechercheaufgabe (in Sek.)	eher gering	12	135,83	43,23
	eher hoch	7	113,71	44,39
Zeit für die schwere Rechercheaufgabe (in Sek.)	eher gering	9	127,44	71,91
	eher hoch	5	126,80	60,68

.

Einfluss der inhaltlichen Expertise. Obwohl die Ärzte im Durchschnitt (M = 2,17; SD = 0,58) mehr Aufgaben lösten als die Studierenden (M = 1,92; SD = 0,52), war dieser Unterschied statistisch nicht signifikant (p = .275) (vgl. Tabelle 20). Bei der Analyse auf der Ebene der einzelnen Fragen wurde ebenfalls kein signifikanter Unterschied zwischen Studierenden und Ärzten festgestellt (alle p > .05).

Tabelle 20. Vergleich der korrekten Lösungsanteile zwischen Inhaltsexperten (Ärzten) und Studierenden für die einzelnen Recherchefragen.

	Antwort / Lösung	**Studierende (n=12)**		**Ärzte (n=12)**		**gesamt (N=24)**	
		n	**%**	**n**	**%**	**n**	**%**
1. Leichte Frage	**korrekt**	7	29,2 %	9	27,5 %	16	66,7 %
	falsch	5	20,8 %	3	12,5 %	8	33,3 %
2. Mittelschwere Frage	**korrekt**	9	37,5 %	10	41,7 %	19	79,2 %
	falsch	3	12,5 %	2	8,3 %	5	20,8 %
3. Schwere Frage	**korrekt**	7	29,2 %	7	29,2 %	14	58,3 %
	falsch	5	20,8 %	5	20,8 %	8	41,7 %

Bezüglich der benötigten Antwortzeiten konnten keine Unterschiede zwischen den beiden Gruppen mit unterschiedlich hoher inhaltlicher Expertise beobachtet werden. Die Ärzte benötigten im Mittel 110,56 Sekunden (SD = 41,2) und die Studierenden 108,43 Sekunden (SD = 41,2) zur Beantwortung der drei Recherchefragen. Bei Betrachtung der Antwortzeiten für die einzelnen Fragen (vgl. Tabelle 21) zeigten sich ebenfalls keine statistisch signifikanten Unterschiede (alle p > .05). Tendenziell brauchten die Studierenden mit zunehmender Schwierigkeit zwar mehr Zeit zur Beantwortung der Fragen, die Ergebnisse waren jedoch

nicht signifikant (p = .12). Ebenso konnten keine statistisch bedeutsamen Unterschiede bei den Antwortzeiten für die verschieden schwierigen Rechercheaufgaben bei der Gruppe der Ärzte nachgewiesen werden (alle p > .05).

Tabelle 21. Vergleich der Lösungszeiten zwischen interneterfahrenen und unerfahrenen Nutzern. Verglichen wurde die Zeit (in Sekunden), die zur Beantwortung der drei Fragen benötigt wurde.

	Expertise	N	M	SD
Zeit für die leichte Rechercheaufgabe (in Sek.)	Ärzte	9	94,78	55,70
	Studierende	7	75,71	26,97
Zeit für die mittelschwere Rechercheaufgabe (in Sek.)	Ärzte	10	126,50	46,75
	Studierende	9	129,00	43,07
Zeit für die schwere Rechercheaufgabe (in Sek.)	Ärzte	7	119,29	54,70
	Studierende	7	135,14	78,81

Anzahl der Suchworte. Die inhaltliche Expertise (Studierende vs. Ärzte) schien keinen Einfluss auf die Menge der Suchworte bzw. die Komplexität des Suchterms zu besitzen. Im Durchschnitt haben die 24 Testpersonen 2,56 Suchworte für die Antwortrecherche benutzt (mittlerer Wert für den ersten Suchterm für alle drei Recherchefragen). Studierende kamen auf M = 2,61 (SD = 0,4) und Ärzte auf M = 2,55 (SD = 0,60) Suchworte pro Query (Suchterm). Statistisch signifikant ist dieser Unterschied jedoch nicht. Der Durchschnittswert für die Anzahl der benutzten Suchworte im Usability-Test entspricht ungefähr dem Ergebnis, das zuvor durch die Analyse des MEDPILOT-Logfiles ermittelt wurde. Bei den MEDPILOT-Suchtermen wurden im Schnitt 2,6 Suchworte pro Query benutzt (vgl. Kap. 5.1.1). Dies kann als ein (indirekter) Beleg für eine valide Stichprobenzusammensetzung gewertet werden und unterstreicht nochmals, dass Mediziner sich in ihrem Suchverhalten kaum vom Durchschnitt der „Normaluser“ unterscheiden.

5.3.3.4 Unterstützungsmaßnahmen zur Verbesserung der Usability

Mithilfe eines *fokussierten Interviews* wurden die *Bewertungen sowie das Verständnis* der Testpersonen bezüglich der eingeführten *Unterstützungsmaßnahmen* zur Verbesserung der Usability erhoben.

Hilfen auf Suchschlitzebene. Durch die integrierte Drop-Down-Hilfe auf Suchschlitzebene konnten sich die Nutzer bei der Auswahl eines Suchwortes unterstützen lassen (vgl. Abbildung 23). Dabei wurden während der Eingabe eines Suchwortes automatisch möglichst

passende MeSH-Terme in Englisch und Deutsch sowie ICD-10-Terme eingeblendet (Auto-suggest-Funktion), die die Nutzer durch Anklicken zur Spezifikation ihres Suchterms verwenden konnten. Fast alle Personen fanden die Einblendung von passenden Suchwortvorschlägen hilfreich (91,7% von N = 24). Nur zwei Testpersonen waren sich nicht schlüssig, ob die Einblendungen der Vorschläge sie nicht zu sehr ablenke.

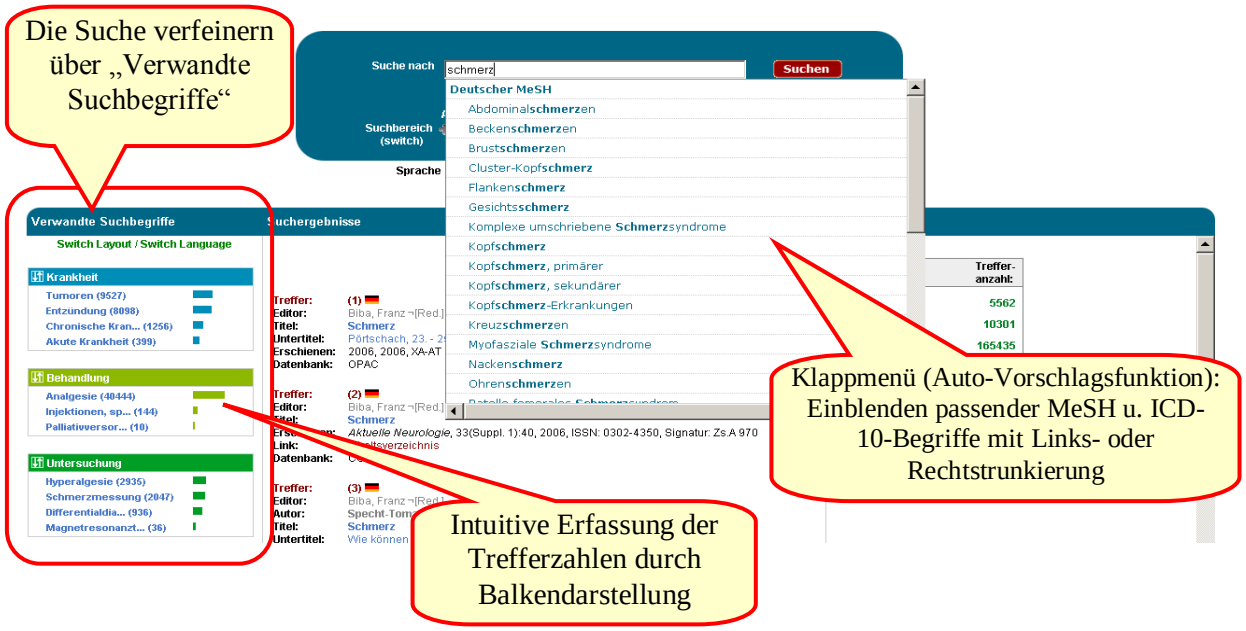

Abbildung 23. Verfeinerung der Suche über die Nutzung einer „Auto-Vorschlagsfunktion“, das Einblenden von „Verwandte Suchbegriffen“ sowie grafische Visualisierung von intuitiv erfassbaren Trefferzahlen.

Zusätzlich sollten die Testpersonen noch bewerten, ob sie eher eine *Links- oder eine Rechtstrunkierung* der zur Unterstützung eingeblendeten Suchworte wünschten. Hier ging es um die Frage, ob die Probanden eher eine Vorschlagsliste bevorzugen, die genau dem von ihnen eingegebenen Anfangsbuchstaben entspricht (Rechtstrunkierung) oder eher eine Einblendung von Begriffen bei der das wahrscheinlich gesuchte Wort sowohl am Anfang in der Mitte oder am Ende (eines Wortes) vorkommen kann. Diese sogenannte Linkstrunkierung sorgt zwar dafür, dass mehr passende Begriffe vorgeschlagen werden, doch kann es dazu führen, dass sich aufgrund der größeren Zahl eingeblendeter Begriffe ein Gefühl der Unübersichtlichkeit einstellt. Es zeigte sich, dass 69,6% der Testpersonen die Linkstrunkierung bevorzugten und nur 30,4% die Rechtstrunkierung. Allerdings konnten mit dem Begriff „MeSH-Term“ nur 29,2% der Testpersonen etwas anfangen. 70,2% der Probanden konnten sich unter „MeSH“ nichts vorstellen.

Hilfen zur Einschränkung des Suchraums. Zu den Optimierungsmaßnahmen bei der Benutzerführung gehörte insbesondere die *Einführung eines Schiebereglers* zur Einschränkung des Suchbereichs und damit auch der Treffermenge. Je nach Stellung des Reglers wurden unterschiedliche Bereiche der Metadaten durchsucht (z.B. nur Titel oder wahlweise Titel + Schlagwörter, vgl. Abbildung 7). Alternativ wurden den Probanden herkömmliche Optionsfelder präsentiert, durch deren Anklicken der Suchraum ebenfalls auf bestimmte Felder im Index eingeschränkt werden konnte (vgl. Abbildung 8).

Zwei Drittel der Testpersonen (16 von 24) bevorzugten den Schieberegler und bewerteten ihn positiv. Bei einer zukünftigen Integration des Schiebereglers in die MEDPILOT-Suchoberfläche würde die Mehrheit der Probanden (66,7%) den Schieberegler als Funktion zur Einschränkung des Suchraums nutzen. Die Begründung für diese Präferenz fiel den meisten Probanden jedoch eher schwer. Testpersonen, die den Schieberegler bevorzugten gaben sinngemäß zur Antwort, dass der Regler einfach „schicker" und „moderner" wirke. Einige dieser Probanden sprachen davon, dass die grafische Form des Schiebereglers „viel suggestiver" wirke als die herkömmlichen Optionsfelder. Der Schieberegler sei eine „leicht verständliche Alternative zur sonstigen Suchauswahl". Jedoch entschieden sich ein Drittel der Probanden klar für die Optionsfelder, da diese ihrer Ansicht nach leichter zu verstehen seien. Unter anderem wurde von diesen Testpersonen bemängelt, dass dass ein Schieberegler „eine kontinuierliche Einstellmöglichkeit suggeriert, die aber tatsächlich nicht vorhanden ist".

Präferierte Voreinstellung zur Einschränkung des Suchraums. Die Ergebnisse der Retrieval-Evaluation des MorphoSaurus-System zeigten, dass die Einstellung „im Titel suchen" zu den besten bzw. relevantesten Ergebnissen unter den ersten fünf Treffern der Trefferliste führte (vgl. Abbildung 16). Doch die Reaktion der Probanden auf diese Voreinstellung des Reglers war relativ eindeutig: Fast alle Testpersonen haben die Voreinstellung („im Titel suchen") vor dem Abschicken ihrer Suche verändert. Die meisten Probanden (66,7%) änderten die Voreinstellung von „Titel" nach „Titel + Schlagwort". 20,8% Prozent der Testpersonen wählten die Einstellung „Autor + Titel + Schlagwort + Abstract", weil sie sich hiervon die größtmögliche Treffermenge versprachen.

Sortierfunktion. Die Probanden wurden auch danach gefragt, welches Sortierkriterium sie für die Darstellung der Trefferliste als Voreinstellung präferieren würden. Die Testpersonen bevorzugten vor allem die *Relevanz* (54,2%) als Sortier- bzw. Ordnungskriterium bei der

Suche. 33,3% der Probanden wünschten sich eine *Kombination aus Relevanz und Aktualität* und nur 12,5% der Testpersonen präferierten *Aktualität* als Sortierkriterium.

Verständnis der An- oder Abwahl einer Datenbank zur Vorabfilterung. Die Ergebnisse der Explorationsaufgabe (Kap. 5.3.3.2) haben gezeigt, dass nur ein kleiner Teil der Probanden die Datenbankauswahl zur Einschränkung des Suchraums auch tatsächlich nutzte. Im Interview zeigte sich zum einen, dass die Testpersonen den „Datenbankkasten" häufig nicht bewusst gesehen haben und zum anderen war ihnen nicht auf Anhieb verständlich, wie die Filterung funktioniert: Die markierte (gefilterte) Auswahl wechselte zur Farbe rot (vgl. Abbildung 24). Vermutlich haben deshalb die Funktion nur vier Probanden aktiv eingesetzt. Es wurde kritisiert, dass die Farbe verwirrend sei. Für Rot-Grün-Blinde dürfte es in der Tat problematisch sein, die Funktion eindeutig zu verstehen. Von einigen der Probanden kam der Vorschlag auf die Farbe zu verzichten und die jeweils aktivierten Teile der Datenbank, also die Bezeichnungen einfach „fett" zu markieren und die unbenutzten Teile (Bezeichnungen) mit blassgrauer Schrift anzudeuten. Positiv wurde hingegen die schnelle Übersicht über die Anzahl der Treffer in den jeweiligen Datenbanken bewertet.

Abbildung 24. Datenbankauswahl bzw. -filterung zur Eingrenzung des Suchraums. Der Index der Testsuchmaschine bestand aus den drei Datenbanken ZB MED-OPAC, CC MED und MEDLINE.

Verwandte Suchbegriffe. Als weitere wichtige Unterstützung wurden die Funktion „Verwandte Suchbegriffe" eingeführt (vgl. auch Abbildung 9). Im Sinne eines „Drill Downs" dient sie der *Verfeinerung der Suche*. Dabei wurden zur Unterstützung und weiteren Spezifikation

des Suchbedürfnisses – unter verschiedenen inhaltsbezogenen Oberkategorien der Medizin – zum Suchterm passende Vorschläge eingeblendet. Obwohl nur knapp 30% der Probanden diese Funktion bei der Recherche auch tatsächlich nutzten, gaben im Interview 80% der Testpersonen an, dass sie sich durch die Einblendung der „Verwandten Suchbegriffe“ in ihrer Suche unterstützt fühlten – sobald sie den Sinn dieser Funktion begriffen. Diese Diskrepanz verweist darauf, dass die Selbsterklärungsfähigkeit dieser Funktion weiter verbessert werden muss. Eine weitere wichtige Frage in diesem Zusammenhang war die Untersuchung der optimalen Reihenfolge beim Erscheinen der „Verwandten Suchbegriffe“.

Sortierungsaufgabe „Verwandte Suchbegriffe“. Insgesamt stammen die „Verwandten Suchbegriffe“ aus neun für Mediziner wichtigen Oberkategorien. Bei einer Suchanfrage erschienen jeweils unterhalb dieser Oberkategorien eng mit der Suchanfrage assoziierte „Verwandte Suchbegriffe“ (vgl. Abbildung 23). Für die Testpersonen standen folgende Oberkategorien zur Auswahl:

1. Krankheit
2. Behandlung
3. Untersuchung
4. Medikamente
5. Journal
6. Autor
7. Anatomie
8. Persönliche Gesundheit
9. Ernährung

Die Testpersonen hatten die Aufgabe, neun Kärtchen mit den Kategorienamen so zu ordnen, dass diejenige Kategorie, die ihnen am wichtigsten war, ganz oben erschien. Alle anderen Kärtchen sollten je nach gewünschter Wichtigkeit darunter angeordnet werden. Die Ergebnisse dieser Sortieraufgabe sind in Abbildung 25 festgehalten. 58,3% der Testpersonen wünschten, dass die Kategorie „Krankheit“ als erste der neun Oberkategorien erscheint. Jeweils 29,2% der Probanden wollten, dass der Begriff „Behandlung“ auf Platz zwei oder drei eingeblendet wird. Die Oberkategorie „Untersuchung“ konnten sich die Testpersonen am besten auf den Rangplätzen zwei, vier oder sechs vorstellen (mit jeweils 25%).

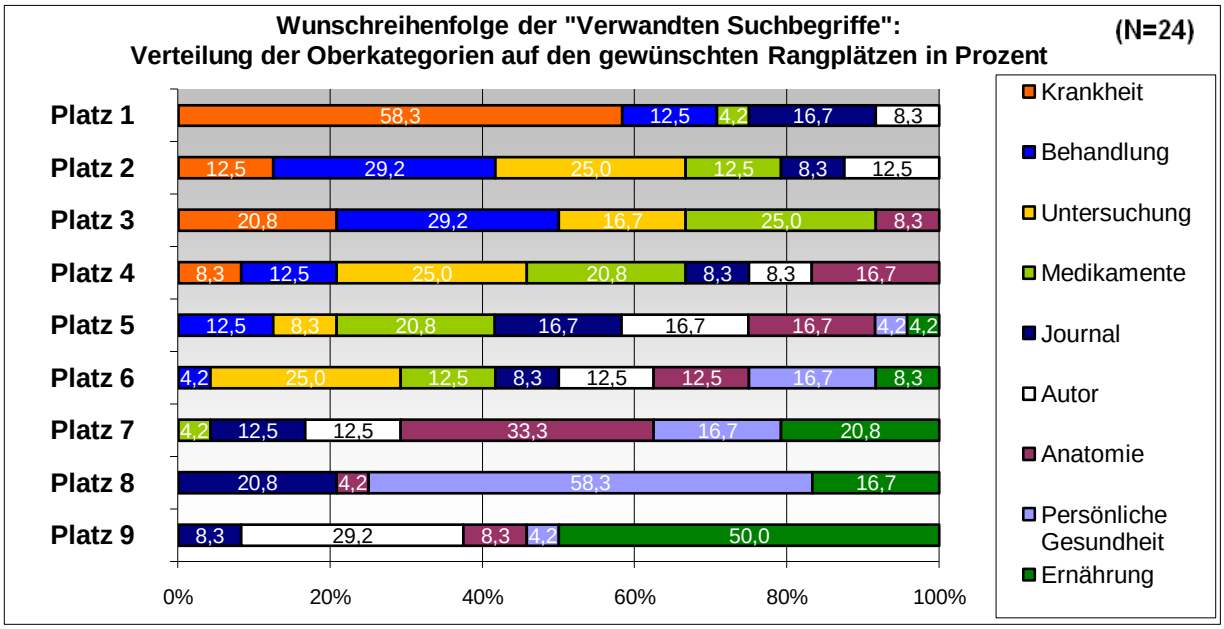

Abbildung 25. Gewünschte Reihenfolge der „Verwandten Suchbegriffe". Prozentuale Verteilung der Anzahl der Testpersonen (N=24), die die jeweilige Oberkategorie („Verwandte Suchbegriffe) auf einem bestimmten Rangplatz sehen wollten.

Die Kategorie „Medikamente" wünschten sich 25 % der Probanden auf Platz drei, wobei sich jeweils 20,8 % der Testpersonen auch einen vierten oder fünften Platz für diese Kategorie vorstellen konnten. Die Kategorie „Journal" erhielt die meiste Zustimmung für eine Anordnung auf Rangplatz acht (20,8 %) sowie Rangplatz eins und vier mit je 16,7 %. Den Begriff „Autor" wollten die Probanden am ehesten auf dem neunten Platz verortet wissen (29,2 %), wobei 16,7 % der Testpersonen diese Kategorie gerne auch auf dem fünften Platz sehen würden. Die Kategorie „Anatomie" wurde mehrheitlich auf den mittleren bis unteren Rangplätzen gesehen. 33,3 % plädierten hier für eine Anordnung auf Platz sieben und 16,75 % für Platz fünf. Im zusätzlichen Interview zeigten sich einige klare Tendenzen: Die meisten Probanden sprachen sich dafür aus, dass die Kategorien „Krankheit" und „Behandlung" auf den vorderen Rangplätzen zu finden sein sollten. Bei den mittleren Rängen waren die Voten nicht so eindeutig. Auf den hinteren Rangplätzen bevorzugte die Mehrzahl der Testpersonen die beiden Oberkategorien „Persönliche Gesundheit" sowie „Ernährung". Diese Verteilungstendenz lässt sich grafisch noch weiter verdeutlichen, wenn die ersten, mittleren und hinteren drei Rangplätze jeweils zusammengefasst dargestellt werden (vgl. Abbildung 26).

Personalisierung der Kategorienreihenfolge. Abschließend wurden die Testpersonen dazu befragt, wie sie die Möglichkeit bewerten, die Kategorien der „Verwandten Suchbegriffe“ nach eigenen Vorlieben ordnen zu können.

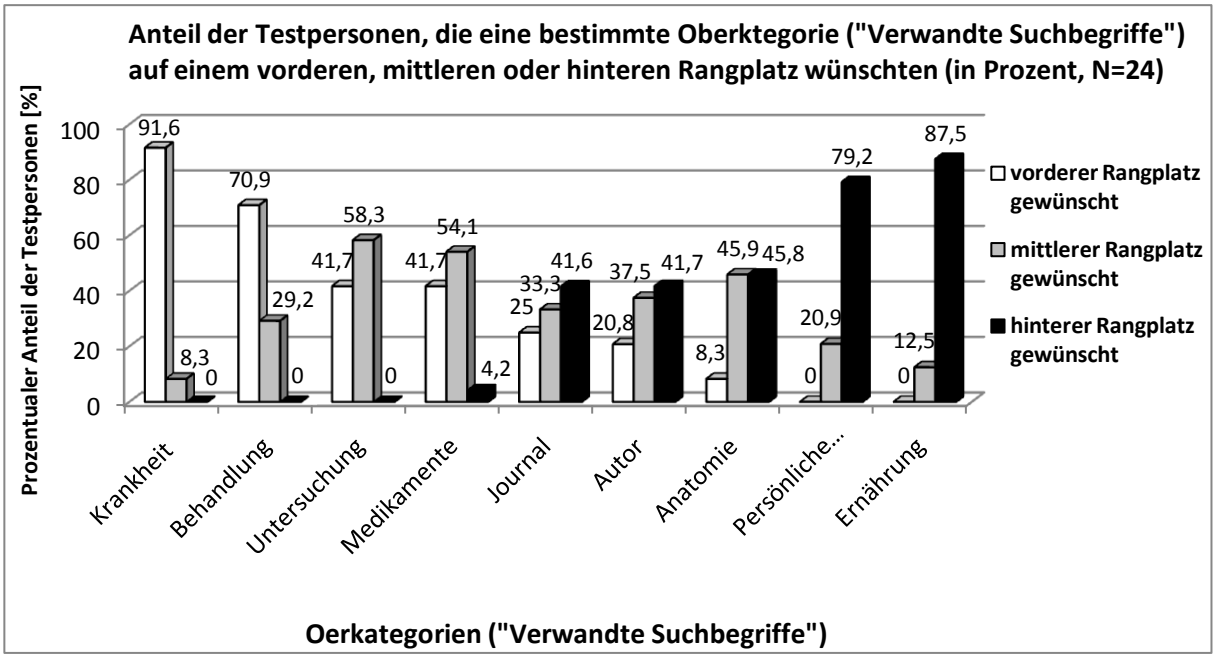

Abbildung 26. Gewünschte Verteilung der Oberkategorien („Verwandte Suchbegriffe“). Es wurden jeweils die drei vorderen, mittleren und hinteren Rangplätze zusammengefasst (N=24).

Personalisierbakeit der Kategorien. Die Probanden hatten die Möglichkeit, die Kategorien mithilfe der Maus an die von ihnen favorisierte Stelle verschieben (vgl. Abbildung 27).

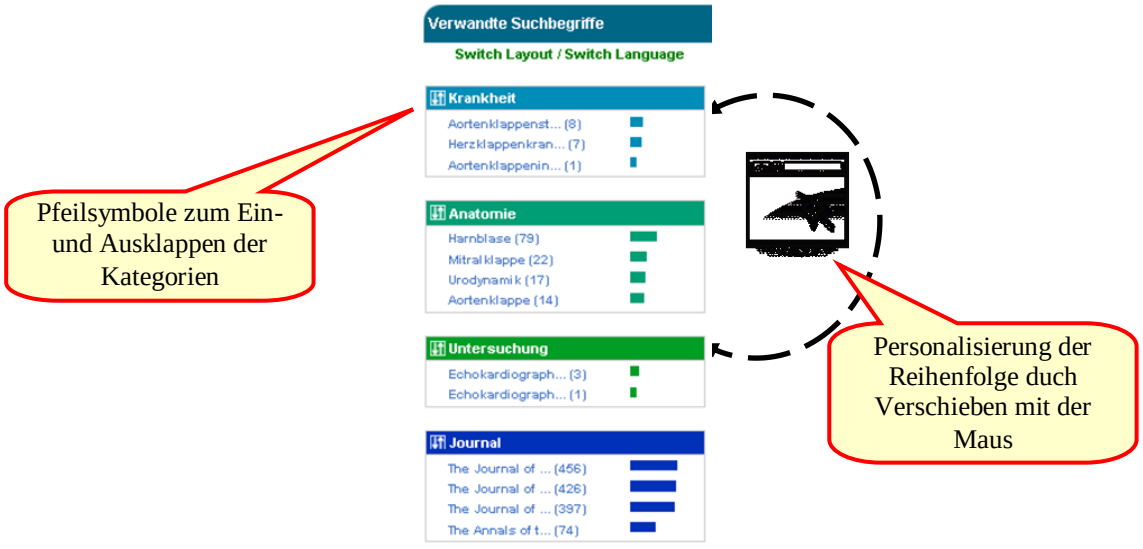

Abbildung 27. Möglichkeit der Personalisierung der Kategorien („Verwandte Suchbegriffe“) durch Verschieben mit der Maus. Doppelpfeile als Symbol zum Ein- oder Ausklappen der Kategorien.

22 von 24 Testpersonen (91,7%) bewerteten die Möglichkeit der persönlichen Reihenfolgebildung als positiv und zwei Personen war die Personalisierung egal (8,3%). Ohne Hinweise auf diese Funktion durch den Versuchsleiter fiel die Personalisierungsmöglichkeit allerdings nur zwei Probanden auf. Dies spricht dafür, dass die Selbsterklärungsfähigkeit in diesem Punkt noch weiter verbessert werden muss.

Es bleibt noch anzumerken, dass über die Hälfte der Probanden mit dem Begriff „Persönliche Gesundheit" nicht viel anzufangen wusste. Hauptsächlich wurden hier Begriffe wie „Fitness", „Wellness" und „persönliches Gesundheitspräventionsverhalten" assoziiert. Der Begriff „Ernährung" interessierte die Mediziner bei einer Literaturrecherche offensichtlich am wenigsten. Schließlich gaben eine Reihe von eher erfahrenen Medizinen noch an, dass sie sich die Kategorien „Journal" und „Autor" gerne grafisch abgesetzt von den anderen „Verwandten Suchbegriffen" vorstellen könnten. Dies ermögliche unter Umständen eine bessere Orientierung, so der Tenor der Kommentare. Zusätzlich wurde von einigen Testpersonen angeregt, auch Kategorien wie „clinical trials", „controlled clinical trials", „randomized clincal trials", „reviews" oder „meta analysis" zur Verfeinerung der Suche anzubieten. Dies erscheint sehr sinnvoll, da Forscher und Mediziner in der Regel sehr an den Ergebnissen evidenzbasierter Untersuchungen interessiert sind und eine Hilfestellung bei der Recherche in diesem Bereich sicher sehr zu schätzen wissen.

Verständnis der Pfeilsymbole. Zur Herstellung einer größeren Übersichtlichkeit bei den Kategorien hatten die Probanden die Möglichkeit, Pfeilsymbole am oberen rechten Rand der Kategorienamen anzuklicken. Dadurch klappten die zu dieser Oberkategorie passenden Treffer ein oder aus. Auch diese Option nutzten nur die wenigsten Testpersonen spontan (vgl. Abbildung 26). Danach befragt, was die Pfeilsymbole bedeuten, antworteten die meisten Probanden, dass sie damit wahrscheinlich die komplette Kategorie „wegklicken" könnten. Also war dieses Symbol nicht eindeutig genug, um selbsterklärend zu sein. Danach befragt, welches Symbol sie sich alternativ vorstellen könnten, antworteten ca. ein Drittel der Probanden, das sie sich ein ähnliches Symbol, wie bei der Benutzung von Fenstern des Windows-Betriebssystems vorstellen könnten, also ein „X" oder ein „–"**,** das dazu führt, dass sich die Kategorie schließt oder einklappt.

Verständnis der Balken. Alle Probanden verstanden auf Anhieb, dass die Balken eine zusätzliche intuitiv erfassbare Darstellung der unterschiedlichen Trefferzahlen repräsentierten.

Obwohl die Information über die Treffermenge durch die Anzeige von Trefferzahlen und Balken redundant war, kritisierte dies keine der Testperson. Viele Probanden berichteten unaufgefordert, dass sie die Balken schneller verstanden als die Trefferzahlen.

Präsentation der Trefferliste. Hier sollte vor allem die Frage beantwortet werden, ob die Testpersonen, das Format der Trefferliste als übersichtlich erlebten. 66,7 Prozent der Probanden hielten die Trefferliste für „eher übersichtlich“ (16 Personen) und nur 33,3 Prozent (acht Personen) stuften sie als „eher unübersichtlich“ ein. In Bezug auf die Verständlichkeit bewerteten 23 Probanden (95,8%) die Trefferliste als „eher verständlich“ und nur eine Testperson als „eher unverständlich“ (4,2%). Zusätzlich wurden die Probanden danach gefragt, ob mit der Trefferliste „angemessen viel Informationen“, „zu viele“ oder „zu wenige Informationen“ präsentiert wurden. 16 Testpersonen (66,7%) fanden, dass mit der Trefferliste „eher angemessen viel Informationen“ präsentiert wurden, sechs Probanden (25%) meinten, dass die Trefferliste „eher zu viel Informationen“ enthielt und für zwei Probanden bot die Trefferliste „eher zu wenig Informationen“. Fast alle Probanden (22) fanden die Worte ihres Suchterms leicht wieder, da die Suchworte durch ein *Highlighting* schnell zu identifizieren waren. Dabei wurden die Suchworte im Treffertext farblich markiert. Nur zwei Probanden wünschten sich eine noch deutlichere Hervorhebung der Suchworte.

Präsentation des Einzeltreffers. Wenn die Probanden sich zu einem Treffer näher informieren wollten, konnten sie durch Klicken auf den entsprechenden Titel in der Trefferliste zur Einzeltrefferdarstellung gelangen. Hier wurde neben den Informationen der Trefferliste auch ein vollständiges Abstract für jeden Treffer präsentiert. 18 von 24 Testpersonen hielten diesen Überblick für ausreichend (75%). Für fünf Testpersonen (20,8%) waren die Informationen der Einzeltrefferdarstellung als Überblick „nicht ausreichend“. Eine Person äußerte sich nicht zu der Frage. 22 Probanden (91,7%) hielten die Einzeltrefferdarstellung für „übersichtlich“ und nur zwei Probanden (8,3 %) stuften sie als „eher unübersichtlich“ ein. Ebenso bewerteten 22 Testpersonen (91,7%) die Einzeltrefferpräsentation als eher verständlich und nur zwei Testpersonen als „eher unverständlich“ (8,3%). Zur Frage, wie die Testpersonen die Menge der präsentierten Informationen einschätzten, gaben 21 Probanden (87,5%) an, dass die Darstellung „eher angemessen viel Informationen“ umfasste; eine Person (4,2%) war der Auffassung, dass die Präsentation „eher zu viele Informationen“ enthielt und

für zwei Personen (8,3%) wurden hier „eher zu wenig Informationen“ dargestellt. Auch hier bewerteten fast alle Testpersonen das Highlighting der Suchworte positiv.

Abschließendes Fazit der Testpersonen im Interview. Die Ergebnisse des szenariobasierten Usability-Tests haben gezeigt, dass die Nutzer die verschiedenen Unterstützungsmaßnahmen zur Verbesserung der Usability des Portals gut annahmen und zum größten Teil als sehr positiv bewerteten. Am Ende der Untersuchung wurden die Probanden gebeten, ein abschließendes Urteil zur Bewertung der Testsuchmaschine abzugeben. 21 von 24 Testpersonen (87,5%) beurteilten das Web-Angebot als „positiv“. Die übrigen drei Probanden (12,5%) stuften es als „neutral“ ein. Valider scheint hier allerdings die Frage danach zu sein, ob die Testpersonen sich vorstellen können, die Suchmaschine auch privat einzusetzen. Dies bejahten 20 Testpersonen (83,3%). Vier Probanden konnten sich zwar auch vorstellen, die Suchmaschine privat zu nutzen, dies aber nur „mit Einschränkungen“ (16,7%).

Besonders positiv wurde von vielen Testpersonen im Abschlussteil des Interviews die Fähigkeit zur automatischen Übersetzung von Suchbegriffen hervorgehoben. Einige der Probanden berichteten, dass sie im Rahmen einer Recherche nach englischsprachigen Artikeln oft erst in der Übersetztungs-Website „LEO“ (www.leo.org) nachsehen würden, um passende englischsprachige Suchterme zu finden. Insgesamt erlebte die überwiegende Zahl der Testpersonen (19 von 24) das Web-Angebot der Testsuchmaschine als sehr benutzerfreundlich. Eine detaillierte Diskussion der Ergebnisse der Usability-Untersuchung folgt in Kapitel 6.

Abbildung 28 zeigt nochmals das vollständige Design, das dem in der Untersuchung benutzten Suchmaschinen-Interface zugrunde lag. Dieses Design könnte nach den vorliegenden Erkenntnissen der Usability-Untersuchung ein guter Ausgangspunkt für die Gestaltung der zukünftigen Suchseite von MEDPILOT sein.

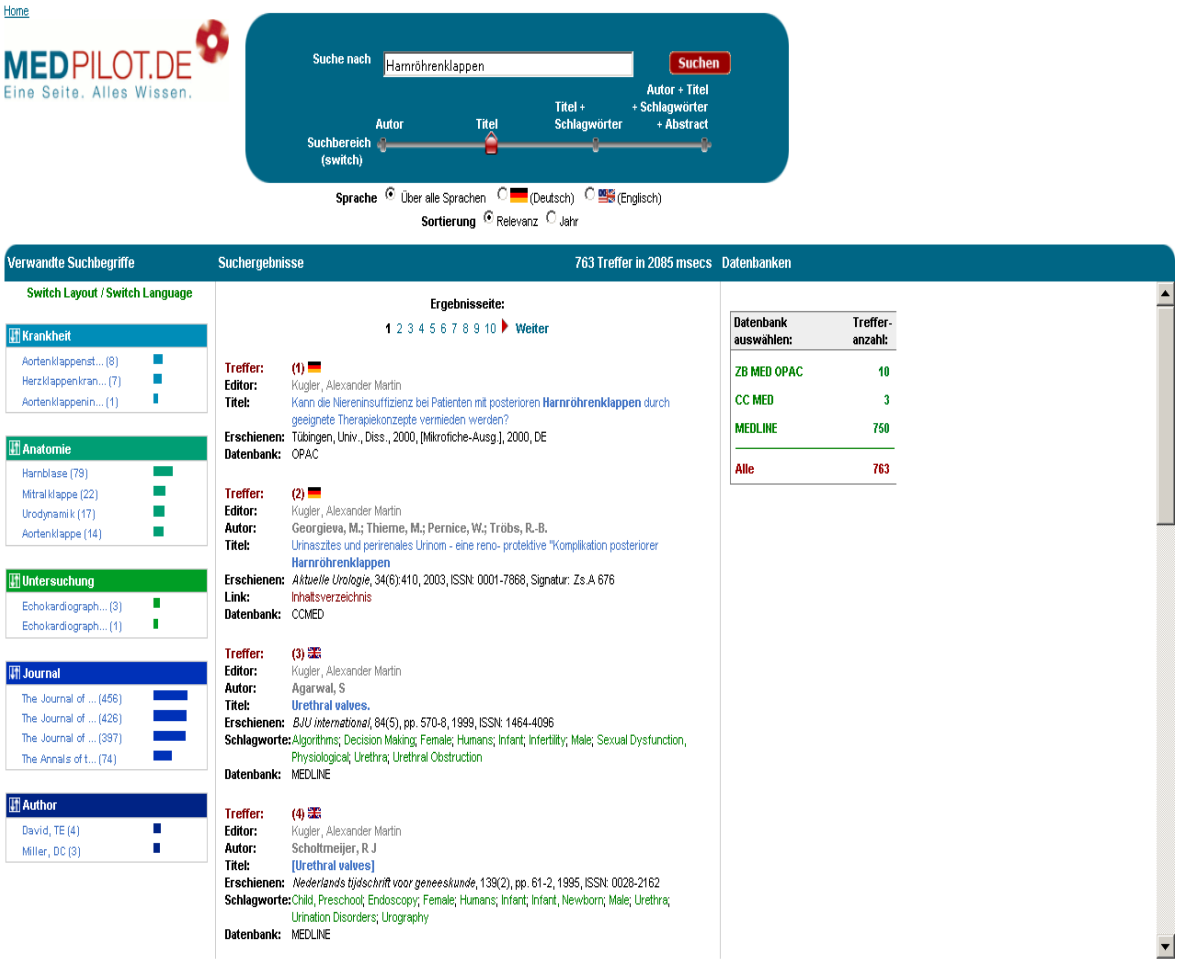

Abbildung 28. Interface der Testsuchmaschine.

5.3.4 Abschließender Fragebogenteil

5.3.4.1 Vergleich mit der Studie von El-Menouar (2004)

Beim Vergleich mit den Ergebnissen einer Fragebogenstudie aus 2004 zur Bewertung verschiedener Usability-Aspekte von MEDPILOT (El-Menouar, 2004), zeigten sich einige Unterschiede gegenüber dem Averbis-Testsystem (vgl. Abbildungen 29 und 30).

Die folgende Ergebnisdarstellung bezieht sich in erster Linie auf die Beschreibung des Anteils der Testpersonen, die der jeweiligen Aussage im Fragebogen tendenziell zustimmten („stimme voll zu“ und „stimme eher zu“). Die Ergebnisse für sämtliche Antworten können den beiden Abbildungen 29 und 30 entnommen werden. Der Anteil der Befragten, die der Aussage „Die Suchmaschine hat unklare Begriffe in der Bedienung“ zustimmten, fiel bei der

Testsuchmaschine etwas geringer aus (20,8%) als in der Befragung von El-Menouar 21,2%. Auf die Antwortmöglichkeiten „stimme nicht zu“ bzw. „stimme eher nicht zu“ entfielen bei der Testsuchmaschine 63%; bei El-Menouar waren es 56,5%. Hier schneidet die Testsuchmaschine etwas besser ab als das bisherige MEDPILOT-Portal.

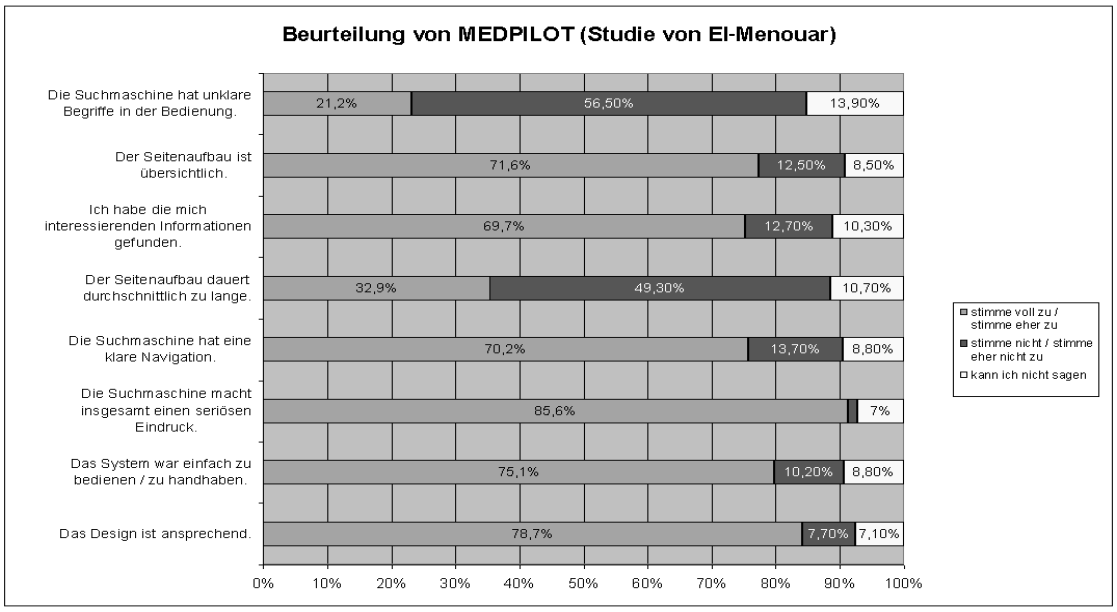

Abbildung 29. Bewertungsdimensionen für die Beurteilung von MEDPILOT. Fragebogenuntersuchung von El-Menouar, 2004; N = 1771).

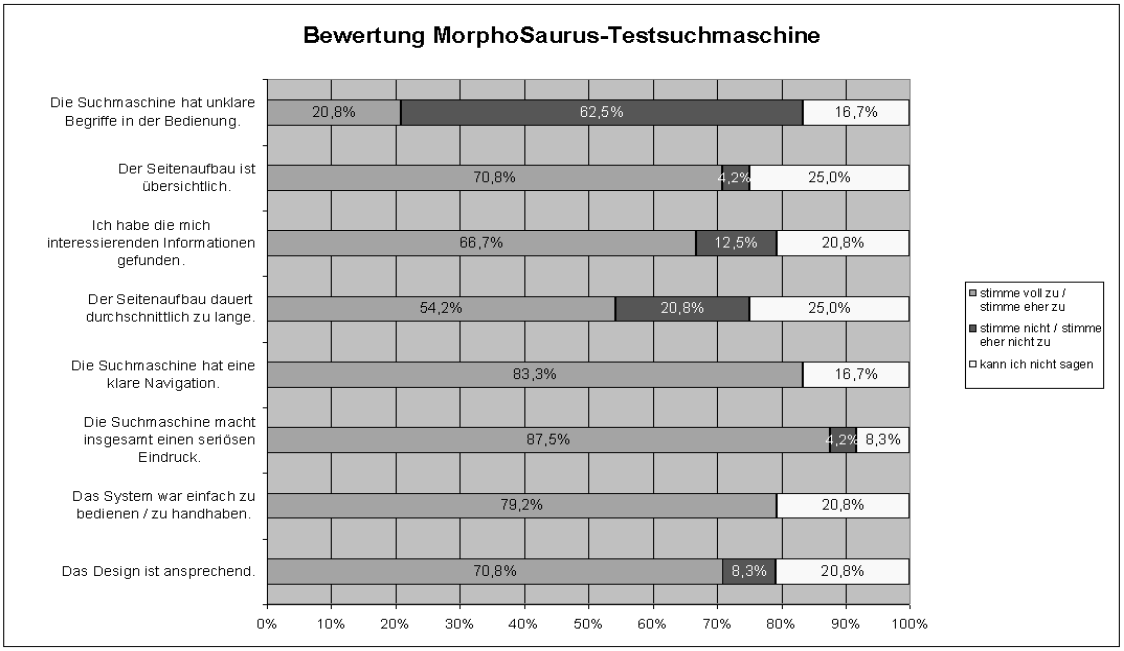

Abbildung 30. Bewertungsdimensionen für die Beurteilung der Averbis-Testsuchmaschine. Untersuchung MorphoSaurus-Projekt, 2008; N = 24).

Für das Item „Der Seitenaufbau ist übersichtlich“ zeigte sich kein großer Unterschied in den Zustimmungswerten (Testsuchmaschine: 70,8% vs. MEDPILOT: 71,6%). Ebenfalls gering war der Unterschied bei dem Item „Ich habe die mich interessierenden Informationen gefunden“ (Testsuchmaschine: 66,7% vs. 69,7% MEDPILOT). Das Item zur Bewertung der „Dauer des Seitenaufbaus“ kann hier nicht valide verglichen werden. Da das Testsystem noch nicht optimal skaliert war, dauerte der Suchvorgang bzw. der Seitenaufbau hier zum Teil unverhältnismäßig lange, was sich auch in den Bewertungen niederschlägt. Die Zustimmungswerte für das Item „Der Seitenaufbau dauert durchschnittlich zu lange“ lagen bei der Testsuchmaschine bei 54,2% und für MEDPILOT bei 32,9%. Im Punkt „Navigation“ scheint das Testsuchsystem der bisherigen MEDPILOT-Version aber deutlich überlegen zu sein. Der Aussage „Die Suchmaschine hat eine klare Navigation“ stimmten 70,2% der von El-Menouar Befragten zu. Dagegen waren es bei der Testsuchmaschine über 83,3% der Testpersonen, die dieser Aussage zustimmten. Einen seriösen Eindruck haben beide MEDPILOT-Varianten hinterlassen. Das Item lautete hier: „Die Suchmaschine macht insgesamt einen seriösen Eindruck“. El-Menouar stellte für diese Aussage 85,6% Zustimmung fest und für die Testsuchmaschine wurde ein Wert 87,5% ermittelt. Dem Item „Das System war einfach zu bedienen/zu handhaben“ stimmten bei der Studie von El-Menouar 75,1% zu; beim Testsystem waren es 79,2% der Testpersonen. Schließlich beurteilten 78,7% der von El-Menouar Befragten das Design als „ansprechend“. (Item: „Das Design ist ansprechend“). In der Averbis-Version waren es 70,8% der Probanden, die dem zustimmten. Jedoch war hier der Anteil derjenigen Personen größer, die sich in diesem Punkt nicht eindeutig äußern konnten (oder wollten); 20,8% wählten die Option „kann ich nicht sagen“ gegenüber 7,1% in der alten MEDPILOT-Version.

5.3.4.2 Beurteilung klassischer Usability-Dimensionen

Die Probanden bewerteten zum Schluss der Untersuchung die Benutzerfreundlichkeit der Testsuchmaschine hinsichtlich klassischer Usability-Dimensionen. Dies geschah mithilfe von fünfstufigen Einschätzskalen (von 1 = stimme gar nicht zu bis 5 = stimme voll zu), wobei fünf Items negativ und fünf positiv formuliert waren. Insgesamt wurde die Usability des Testportals sehr positiv bewertet (vgl. Abbildung 31). Die negativen Items blieben alle unterhalb des Mittelwerts der jeweiligen Skala. Die positiv formulierten Items erreichten alle Werte deutlich oberhalb des Mittelwertes. 21 von 24 Probanden konnten sich vorstellen, das

System häufiger zu nutzen (87,5 %). Genau so viele Testpersonen stimmten der Aussage „Die meisten Menschen würden den Umgang mit der Suchmaschine schnell erlernen“ zu („stimme zu“ oder „stimme voll zu“).

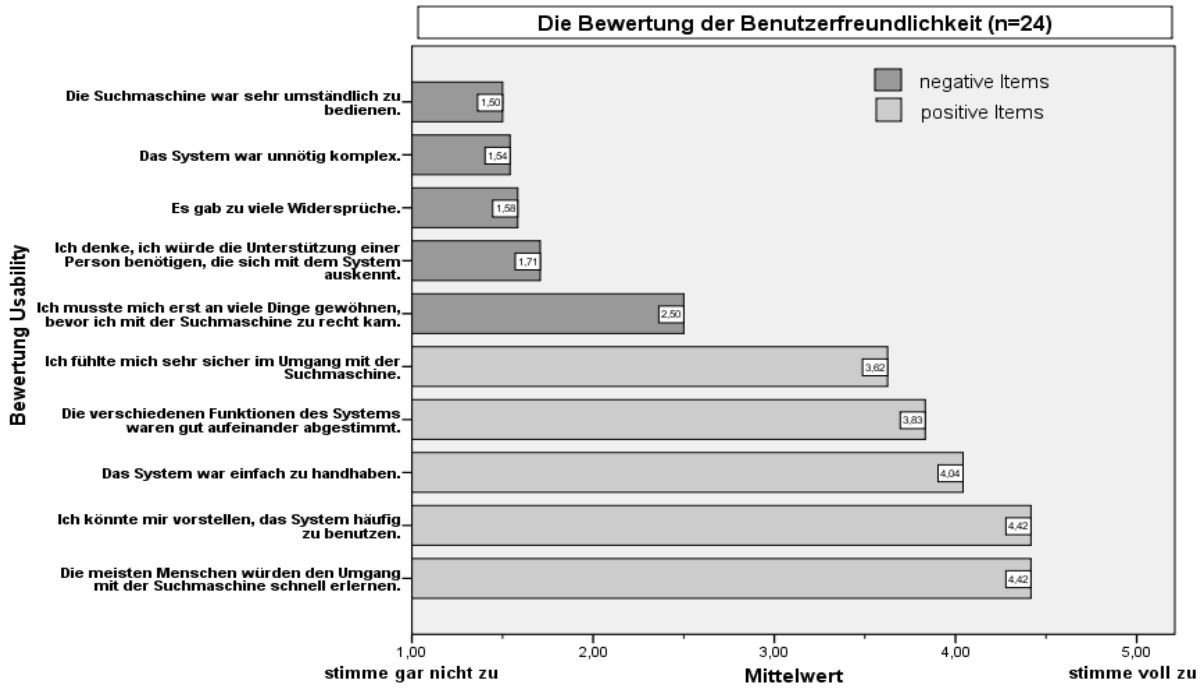

Abbildung 31. Bewertung der Benutzerfreundlichkeit mithilfe von Einschätzungsskalen (N = 24). Dargestellt sind hier die Mittelwerte der Skalen.

5.3.4.3 Bewertung des Images – Vergleich mit vascoda

Das Image der Testsuchmaschine wurde von den Probanden ebenfalls sehr positiv bewertet. Grundlage für die Beurteilung des Images war ein Polaritätenprofil mit positiven und negativen Attributen, welches auch schon für die Imagebeurteilung des Rechercheportals vascoda eingesetzt wurde (vgl. vascoda-Usability-Studie 2007, S. 142). Im Vergleich der beiden Imageprofile schnitt die Testsuchmaschine im Durchschnitt besser ab als vascoda (vgl. Abbildung 32). Besonders deutlich wird dieser Unterschied bei den Attributen „professionell“, „wichtig“, „nützlich“, „hochwertig“, „einfach“ und „modern“.

Die besten Werte im Bereich des positiven Pols lagen bei den Attributen „nützlich“ (M= 1,25; SD = 0,44), „professionell“ (M = 1,58; SD = 0,58) und „hochwertig (M = 1,75; SD = ,53). Dies kann als weiterer Beleg dafür gewertet werden, dass die Testsuchmaschine in Handhabung, Design und in den Trefferergebnissen sehr positiv bewertet wurde. Allerdings

ist eine unmittelbare Vergleichbarkeit mit den vascoda-Daten schwierig, da die Stichprobe für die Testpersonen in 2007 eher heterogen war, während zur Untersuchung der Testsuchmaschine Probanden aus einem homogenen Umfeld (ausschließlich Medizin) rekrutiert wurden.

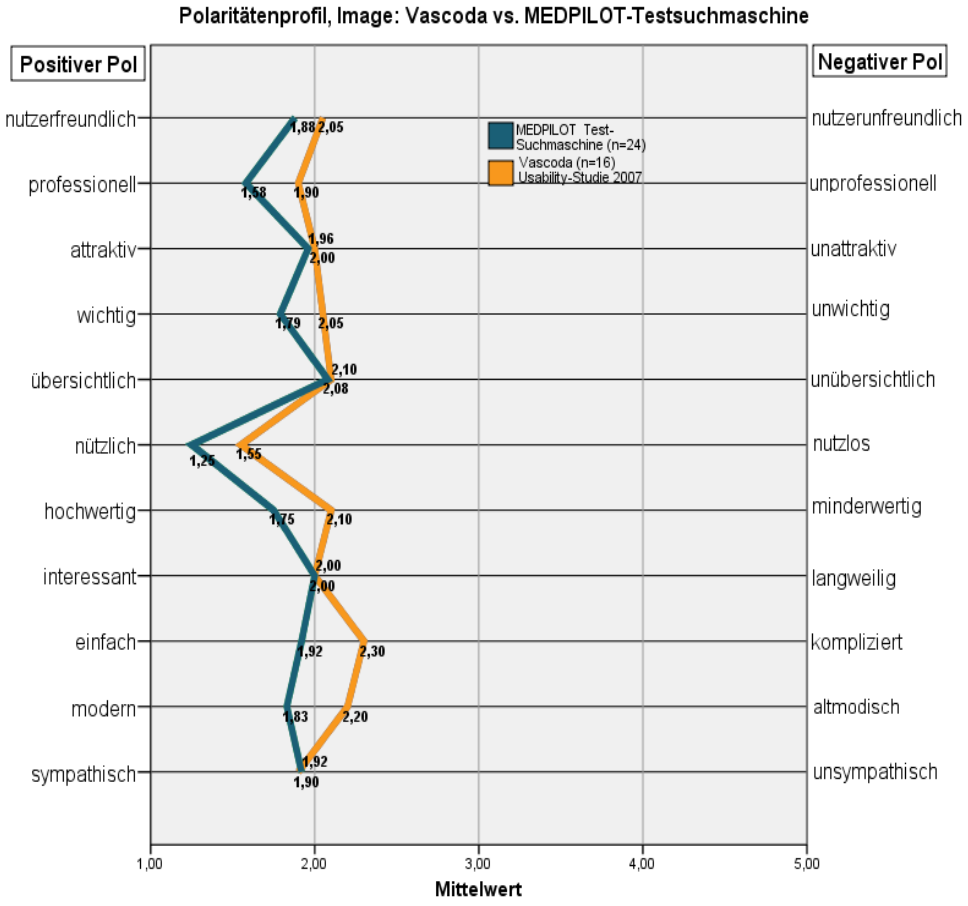

Abbildung 32. Imagevergleich zwischen der Testsuchmaschine und vascoda mithilfe eines Polaritätenprofils.

Bei einer Überprüfung auf Mittelwertunterschiede zwischen den beiden Gruppen der *Ärzte und Studierenden* konnten keine signifikanten Differenzen festgestellt werden. Ebenso ließen sich keine signifikanten Mittelwertunterschiede zwischen *interneterfahrenen und internet-unerfahrenen Probanden* feststellen.

6 Diskussion der Ergebnisse

Im Folgenden werden die Ergebnisse des MorphoSaurus-Projekts diskutiert. Im Mittelpunkt steht dabei die Bewertung der Ergebnisse im Hinblick auf die konkreten Fragestellungen der Untersuchung (vgl. Kap. 3.2). Das Resümee der Arbeit und damit auch die Frage, ob die angestrebten Ziele des Projekts auch erreicht wurden, wird in Kapitel 7 formuliert.

Zu den Ergebnissen der Retrieval-Tests. Als Fazit der Evaluation der Retrieval-Leistung lässt sich Folgendes festhalten: Bei der Verarbeitung von problematischen Sprachphänomen, mit denen Suchmaschinen in der Regel Schwierigkeiten haben, erwies sich die Averbis Core Engine mit der MorphoSaurus-Technik als sehr leistungsfähig. Auf dem Gebiet der Akronymbehandlung sowie bei der fehlertoleranten Behandlung von Rechtschreibfehlern wurde allerdings ein Optimierungsbedarf festgestellt. Hier sind aber durch zukünftige Weiterentwicklungen der Averbis Engine gute Fortschritte zu erwarten.

Das *Benchmarking auf der Grundlage der MEDLINE-Daten* zeigte, dass die Averbis-Testsuchmaschine sämtlichen Konkurrenten überlegen war. Mit einer Trefferrelevanz von 41 % unter den ersten fünf Treffern setzte sich die Testsuchmaschine gegenüber Google, MEDPILOT, GoPubMed, Scirus und PubMed klar durch (vgl. Kap. 5.2.2.1). Bei vergleichbaren Messbedingungen wurden mit der MorphoSaurus-Technik mehr relevante Treffer gefunden als mit den Suchmaschinen der Mitbewerber und auch die Anzahl der Null-Treffer-Meldungen war unter Einsatz der MorphoSaurus-Technik am geringsten.

Bei einem *Vergleich unter Einbeziehung sämtlicher zur Verfügung stehenden Quellen* gelang es der Testsuchmaschine – bei wesentlich geringerem Datenbestand – annähernd gleichhohe Relevanzwerte zu erzielen wie die Suchmaschinen Google oder Google Scholar (vgl. Kap. 5.2.2.2). Bei Vergrößerung des Index durch die Aufnahme weiterer Datenbanken würde sich die Retrieval-Leistung aller Voraussicht nach noch weiter steigern lassen, sodass eine sehr gute Chance bestünde, die Leistungsfähigkeit von Google oder Google Scholar im Bereich des medizinischen Information-Retrievals zu übertreffen. Systematische Tests bezüglich des Einflusses der Indexgröße auf die Retrieval-Leistung stehen allerdings noch aus.

Einfluss der Testkollektionen auf die Höhe der Precision-Werte. Beim direkten Einzelvergleich zwischen der Averbis-Testsuchmaschine und MEDPILOT (vgl. Kap. 5.2.1.8) wurden andere Precision-Werte für die Testsuchmaschine festgestellt als im später durchgeführten

Benchmarking (vgl. Kap. 5.2.2.1). Die Ursache dafür liegt in erster Linie darin begründet, dass hier mit verschiedenen Testkollektionen gearbeitet wurde. Die Retrieval-Tests zur Überprüfung der Leistungsfähigkeit zur Verarbeitung von problematischen Sprachaspekten wurden mit speziell auf die Untersuchungsfragen zugeschnittenen Testkollektionen durchgeführt und hatten nicht den Anspruch auf eine „allgemeine Repräsentativität". Obwohl hier auch mit Zufallsziehungen bei der Zusammenstellung der Suchterme gearbeitet wurde, stellen diese Testkollektionen primär ein auf den jeweiligen Sprachaspekt abgestimmten Pool von Suchtermen dar. Deshalb fielen die Precision-Werte, die hier gemessen wurden, auch relativ unterschiedlich aus.

Im Gegenzug dazu hatte die für das Benchmarking mit relativ hohem Aufwand entwickelte „repräsentative Testkollektion" die Aufgabe, eine valide Basis für einen allgemeinen Vergleich der Retrieval-Leistung verschiedener Suchmaschinen zu bilden. Bei diesem Vergleich stand also nicht die Untersuchung der einzelnen Sprachaspekte im Vordergrund, sondern vielmehr sollten mittels der Testkollektion sowohl die formalen als auch inhaltlichen Aspekte des MEDPILOT-Logfiles Berücksichtigung finden. Mit anderen Worten: Die „repräsentative Testkollektion" sollte möglichst gut die Suchterme typischer MEDPILOT-Nutzer abbilden. Daher ist erklärbar, dass die Höhe der Precision-Werte für die Sprachaspekte-Tests und die Benchmarking-Tests unterschiedlich ausfallen. Dies ist nach Ansicht des Autors keine methodische Schwäche, sondern spiegelt viel mehr ein gegenstandsangemessenes Vorgehen wider. Die Validität der Ergebnisse leitet sich vor allem aus der Standardisierung der Testbedingungen ab. Sämtliche Ergebnisse der Vergleiche zwischen den Suchmaschinen auf der Grundlage einer gemeinsamen Datenbasis besitzen eine sehr hohe Validität (entweder war die Vergleichsdatenbasis CC MED oder MEDLINE). Nicht direkt vergleichbar sind die Ergebnisse der Tests mit unterschiedlich vielen Quellen im Index (vgl. Kap. 5.2.2.2). Sie hatten einen eher heuristischen Wert und sollten Anhaltspunkte für weitere Untersuchungen liefern.

In den nächsten Abschnitten werden die Ergebnisse der Tests für verschiedene Problembereiche sprachlich schwieriger Suchanfragen im Einzelnen bewertet (vgl. Kap. 5.2).

Rechtschreibfehler. Die durchschnittliche Trefferrelevanz von p = 0,25 (unter den ersten fünf Treffern) fiel hier für die Testsuchmaschine nicht besonders hoch aus. Das heißt, von den ersten fünf Treffern waren im Durchschnitt nur ein Viertel relevant. Im Vergleich mit den Precision-Werten, die bei den anderen Sprachaspekten festgestellt wurden, ist das wenig, aber

im direkten Vergleich mit MEDPILOT, das so gut wie keinen Rechtschreibfehler tolerierte (p = 0,04), war es sehr viel. Im Vergleich zu Google ist die Leistung der Averbis-Testsuchmaschine in diesem Punkt noch ausbaufähig. Nach Auskunft von Averbis wurde die Rechtschreibfehler-Erkennung der Suchmaschine inzwischen stark verbessert. Erneute Tests stehen hier aber noch aus.

Akronyme und Abkürzungen. Ein ähnliches Fazit, wie für die Rechtschreibfehlererkennung lässt sich für den Bereich der Akronyme und Abkürzungen ziehen. Mit einem durchschnittlichen Precision-Wert von 0,35 war die Quote für die Trefferrelevanz nicht wirklich hoch. Hier konnte sich die Testsuchmaschine wiederum nicht wirklich von MEDPILOT absetzen (p = 0,33). Auch hier wurde bereits ein leistungsfähigeres Modul implementiert, das eine erheblich verbesserte Erkennungsquote von Akronymen und Abkürzungen verspricht. Ein erneuter Retrieval-Test steht hier ebenfalls noch aus.

Synonyme. Ein gutes Ergebnis erzielte die Testsuchmaschine bei der Auflösung bzw. Erkennung von Synonymen. Während Averbis von der Originalphrase (p = 0,60) zum Synonym (p = 0,42) nur gut 30% an Genauigkeit bzw. Trefferrelevanz verlor, waren es bei MEDPILOT von der Originalphrase (p = 0,39) zum Synonym (p = 0,14) fast 65%. Hier ist die Testsuchmaschine mit der MorphoSaurus-Technologie klar im Vorteil. Im Vergleich mit der Retrieval-Leistung bei anderen Sprachaspekten wäre jedoch eine Steigerung bei der Erkennensleistung von Synonymen wünschenswert. Das bedeutet, dass der Verlust an Genauigkeit (von 30%) von der Originalphrase zum modifizierten Suchterm (Synonym) zukünftig weiter reduziert werden müsste.

Komposita. Hier erzielten sowohl die Testsuchmaschine (p = 0,75) als auch MEDPILOT (p = 0.840) relativ hohe Werte. Entscheidend für den Vergleich der beiden Suchmaschinen ist hier jedoch die Feststellung, dass die Averbis-Testsuchmaschine bei den zerlegten Komposita in der Höhe der Trefferrelevanz sogar noch zulegte (auf p = 0,81) während MEDPILOT auf p = 0,34 zurückfiel: Die Trefferrelevanz (gemessen an Averbis) sank hier bei der Auflösung der Komposita um fast 58% während sie bei der Testsuchmaschine sogar leicht zulegte. Dies ist ein klarer Beleg dafür, dass die Testsuchmaschine kaum an Genauigkeit verliert, wenn Nutzer anstatt eines Kompositums einzelne Worte für ihre Suche verwenden.

Übersetzungsleistung. Insbesondere bei der Recherche in englischsprachigen oder internationalen Datenbanken darf aufgrund der Ergebnisse der Retrieval-Tests viel von der neuen

MorphoSaurus-Technik erwartet werden. Bei der Evaluation hat sich ausgezahlt, dass hier sowohl eine vorwiegend deutschsprachige Datenbank wie CC MED als auch eine vorwiegend englischsprachige Datenbank wie MEDLINE getestet wurde. Dies erlaubte ein differenzierteres Urteil über die Leistungsfähigkeit der Averbis Core Engine.

Ob mit deutsch- oder englischsprachigen Suchtermen gesucht wurde, spielte für das Averbis-System keine große Rolle. Die Testsuchmaschine kam jeweils auf Precision-Werte von ca. 0,9. Der Vergleich auf der Grundlage der MEDLINE-Datenbank zeigte auch, dass MEDPILOT größere Schwierigkeiten bei der korrekten Auflösung von englischsprachigen Suchanfragen in MEDLINE hatte als die Averbis-Testsuchmaschine (MEDPILOT: deutsche Anfragen: $p = 0,58$, englische Anfragen: $p = 0,40$; Averbis: deutsche Anfragen: $p = 0,88$, englische Anfragen: $p = 0,91$).

Weltweit gibt es wesentlich mehr englisch- als deutschsprachige medizinische Publikation. Insofern bietet die MorphoSaurus-Technik einen echten Mehrwert bei der Recherche nach englischsprachiger medizinischer Fachliteratur. Ein Arzt, der sich nicht sicher ist, ob er mit dem richtigen englischsprachigen Begriff nach einschlägiger Literatur sucht, kann nun getrost ein deutsches Fachwort verwenden, um an relevante englische Artikel zu gelangen – und dies ohne den Umweg über Wörterbücher oder Online-Hilfen wie LEO zu nehmen. Ebenso erhöht sich die Wahrscheinlichkeit, dass wissenschaftliche Autoren, die deutschsprachig publizieren, in der englischsprachigen Fachcommunity Beachtung finden, da sie von den englischen Kollegen eher gefunden werden.

Laien- und Expertensprache. Die Retrieval-Tests konnten belegen, dass die Averbis-Testsuchmaschine beim Wechsel von der Laien- zur Expertensprache keine Einbußen bei der Trefferrelevanz aufweist (Testkollektion Laiensprache: $p = 0,92$; Testkollektion Expertensprache: $p = 0,93$). Die Höhe der Precision macht zudem deutlich wie souverän die Averbis-Technologie die Verarbeitung der beiden Sprachformen meistert. Die Leistung von MEDPILOT in diesem Bereich ist zwar auch nicht schlecht (Laiensprache: $p = 0.79$; Expertensprache: $p = 0,73$), doch wird die MorphoSaurus-Technik erheblich besser mit der Transformation von der einen in die andere Sprachform fertig. Dies hängt nach Meinung des Autors nicht zuletzt mit der Performanz des Subwortthesaurus zusammen. Hier findet ja eine Abbildung von verschiedenen bedeutungshaltigen Wortenteilen auf eine Ebene von allgemeiner Bedeutung statt. Insofern sind Bestandteile von Laienbegriffen, die eine ähnliche oder

gleiche Bedeutung besitzen wie bestimmte Bestandteile von Fachbegriffen, im Subwortthesaurus aufeinander bezogen und semantisch angemessen abgebildet.

Die Untersuchung der Fähigkeit zur gleichwertigen Behandlung von Laien- und Expertensprache war zwar nicht die primäre Fragestellung der Arbeit, dennoch hat diese Eigenschaft der MorphoSaurus-Technik ein großes Potenzial, wenn es um Suchanwendungen im medizinischen Bereich geht, die auch für Laien oder Novizen zugänglich sein sollen. Ein Beispiel für eine solche Anwendung ist die „Weisse Liste“[30]. Hierbei handelt es sich um ein Website, mit deren Hilfe sich Patienten und Angehörige über das Leistungsangebot und die Qualität von Anbietern im Gesundheitsbereich informieren können. Das Portal wendet sich ausdrücklich auch an medizinische Laien. Die Averbis GmbH hat die Suchfunktionen der Website maßgeblich mitentwickelt und auch hier die MorphoSaurus-Technik zum Einsatz gebracht. Erste Rückmeldungen durch die Nutzer des Portals sind sehr positiv.

Grammatikalische Variationen. Die gleichwertige Behandlung von grammatikalischen Varianten im sprachlichen Ausdruck stellt viele Suchmaschinen vor massive Probleme. Die MorphoSaurus-Technologie erreichte im Umgang mit diesem Sprachaspekt zwar keinen Spitzenwert, dennoch ist die Leistung beachtlich. Zumal der Vergleich mit MEDPILOT sehr eindeutig ausfällt. Während die Averbis-Testsuchmaschine annähernd eine Precision von $p = 0{,}5$ erreichte und auch durch die Variantenbildung nicht viel an Trefferrelevanz einbüßte, zeigte MEDPILOT hier nur eine Precision von $p = 0{,}20$ (Originalphrase). Durch die Variantenbildung verringerte sich dieser Wert jedoch auf $p = 0{,}15$.

Um die Unterschiede bei den Ergebnissen der Retrieval-Tests zur Überprüfung der problematischen Sprachaspekte noch deutlicher hervorzuheben, gibt die folgende Liste (Tabelle 22) eine Rangreihe der getesteten Sprachaspekte wieder. Diese sind nach Größe der Precision-Werte, die mit der jeweiligen Testkollektion erreicht wurden, geordnet (jeweils nur für den Cut-Off-Wert von 5).

Es bleibt festzuhalten, dass Sprachaspekte, die eher von den *Übersetzungs- bzw. Abbildungseigenschaften* der MorphoSaurus-Technologie profitieren, besser in den Retrieval-Tests abschnitten als Aspekte wie Akronyme oder Rechtschreibfehler, welche stärker durch zusätzliche Analyseschritte verarbeitet werden (müssen).

[30] Homepage: http://www.weisse-liste.de/

Tabelle 22. Vergleich der Precision-Werte zwischen der Averbis-Testsuchmaschine und MEDPILOT für den Cut-Off-Wert von 5 (Trefferrelevanz bis zum fünften Treffer). Die Ergebnisse für die einzelnen Testkollektionen sind nach der Höhe der erreichten Precision-Werte geordnet.

	Precision bei Cut-Off-Wert 5	
	Averbis	**MEDPILOT**
Expertensprache	0,93	0,77
Übersetzung de - en	0,93	0,74
Laiensprache	0,92	0,79
Übersetzung en - de	0,92	0,67
Komposita (zerlegt)	0,81	0,34
Komposita (Original)	0,75	0,80
Synonyme (Original)	0,60	0,39
Grammatik (Original)	0,49	0,20
Grammatik (Variation)	0,46	0,15
Synonyme (modifiziert)	0,42	0,14
Akronyme	0,35	0,33
Rechtschreibfehler	0,26	0,04

Zu den Ergebnissen des Benchmarkings. Der Vergleich der konkurrierenden Suchmaschinen kommt zu relativ eindeutigen Ergebnissen. Der Retrieval-Vergleich auf der Grundlage der MEDLINE-Daten zeigt für die Averbis-Testsuchmaschine die besten Ergebnisse. Für die Effektivität des MorphoSaurus-Algorithmus spricht, dass die Testsuchmaschine die höchsten Werte bei der Trefferrelevanz erreichte, durchschnittlich die meisten Treffer zurückmeldete und zusätzlich die wenigsten Null-Treffer-Meldungen aufzuweisen hatte. Weitere Tests unter Vergrößerung des Datenindex durch Aufnahme zusätzlicher Quellen sowie der Vergleich mit den Retrieval-Werten anderer Suchmaschinen (mit unterschiedlich großem Index) erbrachten die Erkenntnis, dass die Vergrößerung des Suchmaschinen-Index sehr positive Effekte auf die Wirksamkeit des MorphoSaurus-Mechanismus besitzt. Diese Beobachtung sollte in weiteren Untersuchungen systematisch überprüft werden. Das Ergebnis zeigt aber auch: Wenn die Integration der MorphoSaurus-Technik in MEDPILOT abgeschlossen ist, ergeben sich weitere Potenziale für die Erhöhung der Retrieval-Effektivität durch die Erweiterung des Suchmaschinenindex mittels frei verfügbarer Datenbanken.

Zu den Ergebnissen der Usability-Untersuchung. Im Folgenden werden die wichtigsten Ergebnisse der Usability-Untersuchung diskutiert. Über 80% der Testpersonen konnten sich nach Teilnahme an der Untersuchung eine regelmäßige Nutzung eines verbesserten

MEDPILOT-Portals vorstellen. Dies steht in einem gewissen Gegensatz zu der tatsächlichen Nutzung von MEDPILOT zum Zeitpunkt der Untersuchung. Hier waren es über 79% der Testpersonen, die MEDPILOT „seltener als einmal pro Monat“ oder „gar nicht“ nutzten. Im Hinblick auf eine zukünftige Verbesserung der MEDPILOT-Oberfläche stimmt die positive Gesamtbeurteilung der Testsuchmaschine allerdings optimistisch.

Der erste wichtige Punkt im Ablauf der Usability-Tests war die Gedächtnisaufgabe bezüglich dreier wichtiger Informationen auf der Einstiegsseite. Die Ergebnisse weisen darauf hin, dass hier ein Verbesserungsbedarf in der inhaltlichen bzw. formalen Gestaltung der präsentierten Grundinformationen besteht. Immerhin erinnerten sich zwei Drittel der Testpersonen nach kurzer Seitenansicht daran, was die wichtigste Information der Eingangsseite darstellt, nämlich die Möglichkeit medizinische Informationen zu recherchieren. Die Möglichkeit Bestellungen durchführen zu können, wird noch ganz gut erinnert (46%), aber dass zusätzlich auch Volltexte angeboten werden, ist schon schwerer zu behalten (37%). Die wichtigsten Informationen so augenfällig zu präsentieren, dass die Nutzer sich diese merken können, ist Aufgabe eines guten Web-Designers. Der Bekanntheitsgrad des Dienstleistungsangebots könnte zudem durch zusätzliche Marketingmaßnahmen innerhalb der Zielgruppe gesteigert werden.

Sehr aufschlussreich war auch die Beobachtung der Testpersonen während der Phase der freien Exploration. Hier sollte sich zeigen, ob die zur Unterstützung der Nutzer integrierten Funktionen, intuitiv verständlich und selbsterklärend sind.

Die *automatische Vorschlagsfunktion* auf Suchschlitzebene wurde nur von ca. 20% der Testpersonen genutzt, aber 90% der Testpersonen fanden eine solche Funktion hilfreich – wenn sie sie erstmal verstanden haben. Der Wert von ca. 20% für die tatsächliche Nutzung der Funktion fällt deutlich geringer aus als z. B. in der Studie von Anick und Kantamneni (2008), die durch eine Analyse des Yahoo-Logfiles herausfanden, dass 30-37 Prozent der User dynamische Tools gebrauchten. Es kann mehrere Gründe geben, warum dieser Wert in der vorliegenden Untersuchung so gering ausfällt. Zum einen könnte es daran liegen, dass die Probanden mit der Testsuchmaschine noch nicht genügend vertraut sind, wobei auch eine Rolle spielt, dass die Einführung von dynamischen Unterstützungsfunktionen in bekannten Suchmaschinen noch nicht etabliert ist. Die Technik wird in größerem Umfang erst seit ein bis zwei Jahren eingesetzt. Insofern hat hier auf breiter Ebene noch keine „Adaption“ des

Verhaltens im Umgang mit Suchmaschinen stattgefunden. Zum anderen ist es bei der Einführung neuer Techniken wichtig, die Nutzer mittels eines entsprechenden Informationsdesigns auf die Existenz dieser Funktion hinzuweisen und ihnen zu vermitteln, dass sie durch den Gebrauch der Vorschlagsfunktion Zeit und kognitive Mühe sparen können; beispielsweise durch entsprechende Hinweise unterhalb des Suchschlitzes. Bei einem Relaunch wäre es sinnvoll, während einer gewissen „Einführungszeit" die neuen Elemente der Suchmaschine mit zusätzlichen kurzen Erklärungen zu versehen.

Anders verhält es sich bei dem zur Eingrenzung des Suchraums integrierten *Schieberegler*. Diesen nutzten spontan fast 80 % der Testpersonen im Rahmen ihrer Recherchen. Zudem wurde diese Funktion von zwei Dritteln der Probanden positiv bewertet. Hier zeigt sich, dass mit einfachen Mitteln eine Verbesserung der Suchmaschinen-Usability erreicht werden kann. Obwohl der Schieberegler keine anderen Funktionen bietet als die herkömmlichen Optionsfelder, wurde er von vielen Untersuchungsteilnehmern in der Bedienung als viel intuitiver erlebt. Da ca. ein Drittel der Testpersonen jedoch explizit die Optionsfelder präferierten, ist für einen zukünftigen Relaunch anzuraten, diese Funktion in einer personalisierbaren Form zu implementieren, die den Anwendern die Wahl zwischen herkömmlich gestalteten Auswahloptionen und dem Einsatz eines Schiebereglers lässt.

Die Funktion der *Verwandten Suchbegriffe* wurde von nur knapp 30 % der Testpersonen während der Exploration auch genutzt. Hier zeigt die Diskrepanz zur späteren positiven Bewertung durch 80 % der Testpersonen, dass ebenso wie bei der Autosuggest-Funktion die Selbsterklärungsfähigkeit des Unterstützungsangebots verbessert werden muss. Es stellt sich die Frage, für wen und in welcher Situation ein solches „Drill Down" zur Fokussierung des eigenen Suchbedürfnisses eine Hilfe darstellt? Aus den Interviewdaten ging hervor, dass gerade die Studierenden oder allgemeiner, Personen, die noch über keine langjährige Expertise in einer Inhaltsdomäne verfügen, durch Einführung solcher Hilfen stark profitieren könnten. Dies hat sicherlich mit der kognitiven Entlastung der Nutzer durch die Einblendungen zu tun. Wenn solche Funktionen in ein Suchinterface eingeführt werden, müssen sie über ein hohes Maß an Selbsterklärungsfähigkeit verfügen. Dies war in der vorliegenden Untersuchung nur teilweise der Fall. Immerhin nutzten spontan 30 Prozent der Testpersonen die Funktion der *Verwandten Suchbegriffe*, dennoch ist hier durch eine grafisch und textlich eindeutigere Gestaltung eine Optimierung der Selbsterklärungsfähigkeit vorstellbar (z. B.

durch Einsatz zusätzlicher kurzer Alternativ-Texte, die eingeblendet werden, sobald die User mit der Maus über die Funktion gleiten).

Etwas über 40% der Testpersonen nutzten die Möglichkeiten der *Sortierung* nach Sprache / Relevanz oder Aktualität. Für die meisten der Testpersonen war die Relevanz das wichtigste Sortierkriterium (54,2%). Die restlichen Probanden wünschten sich eine Kombination aus „Relevanz und Aktualität“ oder „Aktualität“ als Sortierkriterium. Deshalb empfiehlt sich aus empirischer Sicht die Voreinstellung nach Relevanz. Dennoch sollten die Nutzer auch diese Funktion personalisieren können, um ihre Präferenz beim nächsten Besuch der Site als Grundeinstellung vorzufinden.

Nur ganz wenige Probanden nutzten auch die Option zur *Auswahl bestimmter Datenbanken* (16,7%). Die grafische Realisierung wurde kritisiert, aber durch die Interviews wurde ebenso deutlich, dass die Testpersonen eine solche Funktion grundsätzlich begrüßten.

Ein Interface weist dann ein gutes Informationsdesign bzw. eine gute Usability auf, wenn die Funktionen und Navigationsstrukturen der Website auch tatsächlich schnell und einfach von den Nutzern durchschaut werden können und der Mehrwert einer Funktion unmittelbar begreifbar ist, sodass diese mit hoher Wahrscheinlichkeit auch schon beim ersten Websitebesuch genutzt wird. Insofern war aufgrund der Diskrepanz zwischen tatsächlichem Verhalten und der positiven Bewertungen durch die Testpersonen eine gewisse Skepsis angezeigt. Hier kommt sicherlich auch der aus den Sozialwissenschaften bekannte Effekt der „sozialen Erwünschtheit“ zum Tragen. Natürlich ist der Anteil der Bewertung, der auf diesen Effekt zurückgeht, schwer zu beziffern. Deshalb wird an dieser Stelle dafür plädiert, dass in ähnlichen Untersuchungen stets auch der konkrete Umgang mit dem Web-Angebot beobachtet wird, weil dadurch wesentlich validere Vorhersagen des tatsächlichen Nutzerverhaltens möglich werden.

Die Ergebnisse des szenariobasierten Usability-Tests, bei dem die Testpersonen Rechercheaufgaben lösten, zeigen, dass die Testsuchmaschine durchaus über eine hohe Effektivität und Effizienz verfügt. Die leichte Recherchefrage wurde von zwei Dritteln der Testpersonen innerhalb der vorgegebenen Zeit von fünf Minuten richtig beantwortet. Die mittelschwere Frage konnten sogar 79,2% der Probanden richtig beantworten. Die schwere Recherchefrage lösten immerhin noch 58% der Untersuchungsteilnehmer im vorgegebenen Zeitfenster. Auch hier gibt es sicherlich Verbesserungsmöglichkeiten. Diese hängen aller-

dings unmittelbar mit dem Design und den Hilfsfunktionen zusammen, die bei einem späteren MEDPILOT-Relaunch eingeführt werden. Spätestens dann sollte der Usability-Test wiederholt werden.

Zur Imagemessung. Die Imagemessung mittels eines Polaritätenprofils zeigt in der Regel eine Momentaufnahme der Attribute und Eigenschaften, die Nutzer mit einer Website verbinden. Um hier zu validen Schlussfolgerungen zu gelangen, wäre es nötig, solche Beobachtungen auf eine breitere Datengrundlage zu stellen. Dazu sollte die Stichprobengröße in künftigen Untersuchungen wesentlich erhöht werden. Hiermit wären dann auch statistisch valide Aussagen zu Unterschieden zwischen Subgruppen möglich (Ärzte vs. Studierende, Interneterfahrene vs. Internetunerfahrene, Forscher vs. Kliniker usw.).

Zum Schluss der Diskussion sollen an dieser Stelle nochmals die Vorteile der Anwendung der MorphoSaurus-Technik als innovative Technologie im Bereich des medizinischen Information-Retrievals zusammengefasst werden:

1. Die *Suche ist sehr schnell* (aufgrund der indexbasierten Suche im Millisekundenbereich). Zudem sind mit Lucene als OpenSource-Anwendung die Kosten für eine moderne Suchmaschinentechnologie überschaubar.

2. Bei fast allen problematischen Sprachaspekten konnte eine bessere Verarbeitungsleistung als bei den Konkurrenten beobachtet werden. Die Averbis Core Engine lieferte *mehr und relevantere Treffer*, wobei letztlich besonders die hohe Trefferrelevanz unter den ersten fünf Hits von großer Bedeutung ist. Zudem zeigten sich unter Einsatz der MorphoSaurus-Technologie *die wenigsten Null-Treffer-Meldungen*.

3. Besonders gute Ergebnisse hinsichtlich der Trefferrelevanz wurden bei den Sprachaspekten *Übersetzungen*, *Laien-Expertensprache*, *Komposita*, *Synonyme* und *sprachliche Varianten* erzielt. Ausbaufähig sind die Leistungen in den Bereichen *Akronyme / Abkürzungen* sowie *fehlertolerante Rechtschreibung*.

4. Die Averbis Core Engine bietet eine *intelligente Treffer-Rangordnung* durch einen durchdachten Katalog von Gewichtungskriterien, wobei Suchworte die im Titel eines Artikels vorkommen, eine wichtige Rolle im Relevanzalgorithmus spielen.

5. Die Ergebnisse der vorliegenden Untersuchung konnten zeigen, dass die Kombination aus technischen Innovationen (MorphoSaurus-Technik) und Maßnahmen zur Verbes-

serung der Suchmaschinen-Usability dazu beitragen kann, die *Zufriedenheit der Nutzer* im Umgang mit einem wissenschaftlichen Suchsystem zu erhöhen.

Für eine genaue Einschätzung der Leistungsfähigkeit der MorphoSaurus-Technik bei der Verarbeitung sprachlich problematischer Suchanfragen wäre es wünschenswert, die vorliegenden Testkollektionen (zur Überprüfung der Sprachaspekte) auch anhand anderer Suchmaschinen, wie PubMed, Scirus, und Google bzw. Google Scholar zu überprüfen. Aufgrund der begrenzten Ressourcen war dies innerhalb des MorphoSaurus-Projekts leider nicht möglich.

7 Fazit und Desiderata

Das letzte Kapitel dieser Arbeit enthält das Fazit zu den zahlreichen Ergebnissen des MorphoSaurus-Projekts. Zum einen bleibt zu resümieren, ob die angestrebten Ziele erreicht wurden und zum anderen sollen hier auch mögliche Zielperspektiven und Desiderata für nachfolgende Untersuchungen in diesem Bereich formuliert werden. Schließlich wird ein Ausblick auf Entwicklungen und Tendenzen im webbasierten Information-Retrieval gegeben sowie die Übertragbarkeit der MorphoSaurus-Technik auf andere Projekte und Inhaltsdomänen diskutiert.

Nach den Erkenntnissen, die das Evaluationsteam der ZB MED durch die Untersuchung der Retrieval-Leistung gewonnen hat, besitzt der MorphoSaurus-Ansatz ein großes Potenzial für die Optimierung der Verarbeitungsleistung sprachlich problematischer Suchanfragen. In der Domäne des medizinischen Information-Retrievals scheint ist die Morpho-Saurus-Technik schon jetzt so leistungsfähig, wie keine der hier getesteten Konkurrenten. Dies legen zumindest die Ergebnisse für das Benchmarking mit MEDLINE als Vergleichsbasis nahe. Hier wurde für die Averbis-Testsuchmaschine im Durchschnitt sowohl die höchste Precision, die wenigsten Null-Treffer-Meldungen als auch die meisten Treffer festgestellt.

Dennoch unterstreichen die Ergebnisse der durchgeführten Usability-Untersuchung die Notwendigkeit der Einbeziehung von empirischen Daten des Userverhaltens. Erst wenn sich ein System auch in der Benutzung durch die User bewährt, kann von einem Erfolg eines Information-Retrieval-Systems gesprochen werden. Daher hat sich, das dieser Arbeit zugrundeliegende Rahmenmodell zur Beschreibung des Erfolgs webbasierter Retrieval-Systeme gut bewährt. Erst durch die Vielfalt der hier eingesetzten Evaluationsmethoden, wurde das Projekt der Komplexität des untersuchten Gegenstandsbereichs gerecht und konnte seine Ziele erreichen.

Zu Ziel 1: *Implementierung moderner Suchmaschinentechnologie*. Die Averbis-Testsuchmaschine setzte auf der Open-Source-Java-Bibliothek Lucene auf. Dabei handelt es sich um ein leistungsstarkes Suchmaschinen-Framework, das sich durch eine hohe Performanz und gute Skalierbarkeit auszeichnet. Es ist sehr gut geeignet, um auch sehr große Textkorpora zu indexieren und für die Suche aufzubereiten. Von den flexiblen Möglichkeiten dieses Frame-

works profitierte das Projekt sehr. Die Integration der MorphoSaurus-Technik in das Suchmaschinen-Framework verlief ohne große Schwierigkeiten.

Zu Ziel 2: *Entwicklung valider Evaluationsmethoden*. Kernstück von Retrieval-Tests sind u. a. qualitativ hochwertige Testkollektionen, mit denen die Leistungsfähigkeit von Retrieval-Systemen valide überprüft werden kann. Die im Rahmen des MorphoSaurus-Projekts entwickelten Testkollektionen waren sowohl von ihren Inhalten als auch von der gewählten Konstruktions- und Messmethode gegenstandsangemessen. Mithilfe von inhaltsanalytischen Methoden wurden aus dem MEDPILOT-Logfile (bzw. anhand einer ausreichend großen Stichprobe) die Test-Suchterme zur Überprüfung der Retrieval-Leistung für die Verarbeitung sprachlicher Problemfälle extrahiert. Anschließend wurden die spezifischen Testkollektionen durch eine Zufallsauswahl aus den extrahierten Suchtermen zusammengestellt, sodass die jeweiligen Kollektionen die zu untersuchenden Fragestellungen gut abbildeten. Im zweiten Schritt wurde nach dem gleichen Prinzip aus dem *Gesamtpool* der extrahierten Suchterme eine zusätzliche *repräsentative* Testkollektion erstellt. Dieses Vorgehen berücksichtigt zum einen, dass *alle Suchterm-Inhalte die gleichen Chancen* zur Aufnahme in die Testkollektion erhalten (Zufallsauswahl) und zum anderen stellt dieses Vorgehen sicher, dass auch die *Verteilung von formalen Aspekten* der Suchterme (im Logfile) bei der Konstruktion der Testkollektion Beachtung findet.

Ein Verbesserungsbedarf hinsichtlich der Methodik wird allerdings bei der Beurteilung der Trefferrelevanz gesehen. In der vorliegenden Untersuchung war es so, dass ein zurückgemeldeter Treffer als relevant oder eben nicht relevant eingestuft worden ist. Es gibt aber auch Fälle, in denen ein Treffer als mehr oder weniger relevant eingestuft werden kann. Ein Bewertungssystem, das auf einer kontinuierlichen Einschätzskala für die Relevanzurteile aufbaut, böte hier sicherlich Vorteile.

Die Gültigkeit des Urteils bei der Einschätzung der Relevanz eines Treffers hängt auch immer von einer korrekten Interpretation des „Information Need“ ab, also davon, ob das Suchbedürfnis, welches durch den Inhalt des Suchterms ausgedrückt wird, vom Bewerter eines Treffers angemessen beurteilt wird. Deshalb sollte die „semantische Reichweite“ des im Suchterm formulierten Informationsbedürfnisses zwecks Abgrenzung genau umrissen werden. Dazu ist es sinnvoll, die inhaltliche Kernbedeutung des Suchbedürfnisses zu beschreiben und darüber hinaus auch die semantischen Abgrenzungen zu anderen Suchinhalten, sodass

bestimmte Treffer zweifelsfrei als nicht relevant eingestuft werden können. Dieses Vorgehen wurde im MorphoSaurus-Projekt vor allem bei der Konstruktion der „repräsentativen Testkollektion“ gewählt. Daher gilt diese Testkollektion aus Sicht des Evaluationsteams als ein sehr valides Instrument zur Überprüfung der Retrieval-Leistung der untersuchten Suchsysteme. Von Zeit zu Zeit muss aber auch eine solche Testkollektion an die sich wandelnden Suchbedürfnisse der Nutzer angepasst werden, um weiterhin als valide zu gelten. Dies geschieht am besten über eine regelmäßige Analyse des Logfiles, welche Hinweise auf Veränderungen bei stark nachgefragten Inhalten liefert. Eine engmaschige Beobachtung des MEDPILOT-Logfiles böte die Möglichkeit, zeitnah auf die Bedürfnisse der Nutzer zu reagieren und die Angebotspalette im Sinne der Kundenzufriedenheit zu erweitern oder zu fokussieren.

Wie aus den Ergebnissen der Retrieval-Tests hervorgeht, hängt die Höhe der festgestellten Precision-Werte – neben den unterschiedlichen Fähigkeiten der Suchmaschinen zur Verarbeitung der Suchterme – sowohl von der Beschaffenheit der eingesetzten Testkollektion als auch von der Größe des Index des durchsuchten Datenkorpus ab. Valide Aussagen zur Retrieval-Leistung von Suchmaschinen können nur unter der Einhaltung von gleichen Testbedingungen gemacht werden. Dazu gehört in erster Linie eine vergleichbare Datenbasis. Mit MEDLINE als Vergleichsgrundlage ist dem Projekt die Schaffung von gleichen Rahmenbedingungen für die Retrieval-Tests gut gelungen. Systematische Untersuchungen zum Effekt der Steigerung der Retrieval-Leistung durch die Vergrößerung der Datenbasis (vgl. 5.2.2.2) stehen allerdings noch aus.

Aus methodischer Sicht hat sich die ergänzende Durchführung einer Usability-Untersuchung sehr bewährt. Hier ist es vor allem der szenariobasierte Usability-Test zu nennen, der Auskunft darüber geben kann, ob die Interface-Gestaltung dazu beiträgt, ob die Nutzer ihre Ziele (Literaturrecherchen) auch zügig erreichen können. Die Methode des „lauten Denkens“ während des fokussierten Interviews hat sich als sehr effektiv erwiesen, um viele wertvolle Hinweise zur Verbesserung der Selbsterklärungsfähigkeit der Suchmaschinenoberfläche zu gewinnen. Die Erhebung von Fragebogendaten zur Beurteilung klassischer Usability-Eigenschaften erlaubte darüber hinaus einen Vergleich mit ähnlichen Untersuchungen. Ob die Nutzer bestimmte Hilfs- und Unterstützungsangebote auch annehmen, kann letztlich nur

empirisch überprüft werden. Diesen Nachweis zu führen, ist mit den eingesetzten Methoden gut gelungen.

Zu Ziel 3: *Evaluation der MorphoSaurus-Technologie*. Das Projekt hat mit der gewählten Methodik eine umfangreiche Evaluation auf hohem Niveau umgesetzt. Der Vergleich der Retrieval-Leistung von MEDPILOT und der Averbis-Testsuchmaschine zielte darauf ab, Stärken und Schwächen der beiden Suchsysteme zu identifizieren (vgl. Kap. 5.2). Es wurden verschiedene Retrieval-Tests durchgeführt, wobei spezielle Testkollektionen zur Überprüfung sprachlich problematischer Suchanfragen zum Einsatz kamen. Dadurch konnte sehr genau nachgewiesen werden, dass es sich bei der getesteten MorphoSaurus-Technologie um einen innovativen Ansatz handelt, der die Leistung einer medizinischen Suchmaschine erheblich verbessern kann. Durch den Einsatz der Averbis Core Engine konnten zum einen *mehr relevante Treffer* unter den ersten fünf bis 20 Hits gefunden werden zum anderen ließen sich auch *generell mehr Treffer* finden. Zusätzlich nahm durch den Einsatz der MorphoSaurus-Technologie der Anteil der *Null-Treffer-Meldungen* ab (vgl. Kap. 5.2.2).

Zu den Kernzielen des MorphoSaurus-Projekts zählten zwei Punkte:

1. *Die Etablierung der Mehrsprachigkeit*. Hierbei ging es um den Nachweis, dass mithilfe der MorphoSaurus-Technologie die Anfragesprache in der der Suchterm formuliert wird, an Bedeutung verliert. Durch die Zerlegung der Anfrage in „Morpheme“ bzw. Subwörter und durch den Einsatz eines interlingualen Subwortthesaurus konnten hier bemerkenswerte Ergebnisse in der Übersetzungsleistung erzielt werden. Die gelungene Etablierung der Mehrsprachigkeit des Suchsystems konnte durch die durchgeführten Tests hinlänglich belegt werden.

2. *Die Normalisierung von Sprachvarianten*. Ein großes Problem von aktuellen Suchmaschinen liegt bisher im Versagen bei der Normalisierung sprachlicher Varianten. Auch wenn eine Suchanfrage sprachlich nur leicht modifiziert wird, haben viele der bekannten Suchmaschinen massive Probleme mit der gleichwertigen Behandlung einer veränderten Anfrage: Durch die Variation gelingt es diesen Suchmaschinen nicht, die gleichen Treffer zu finden; auch die Trefferanzahl verändert sich. Wie die aufwendigen Untersuchungen zeigen konnten, besitzt die Averbis-Technologie die Fähigkeit zur Normalisierung von Sprachvarianten und kann diesen Vorteil, etwa bei der Erkennung von Synonymen, von grammatikalischen Variationen oder bei der Auf-

lösung von Komposita sowie bei der gleichwertigen Behandlung von Laien- und Expertensprache, voll ausspielen.

Im Bereich der *Akronym-Auflösung* und in der *Rechtschreibfehler-Verarbeitung* gibt es allerdings noch einen Optimierungsbedarf. Zur Verbesserung der Verarbeitungsleistung in diesen Punkten sind von Averbis bereits aktuellere Module entwickelt worden.

Benchmarking. Auch der Vergleich mit anderen konkurrierenden Suchmaschinen bestätigte die Leistungssteigerung im medizinischen Information-Retrieval durch den Einsatz der Averbis-Technologie. Hier zeigte sich die Testsuchmaschine der von Mediziner oft genutzten PubMed-Anwendung weit überlegen. Ein Suchmaschinen-Vergleich auf der Grundlage von MEDLINE-Daten zeigte auch hier eine größere Trefferrelevanz, insgesamt mehr Treffer und weniger Null-Treffer-Meldungen als bei den Konkurrenten. Durch eine Vergrößerung des Index könnte es MEDPILOT mit moderner Suchmaschinen-Technologie und integrierter MorphoSaurus-Technik in naher Zukunft sogar gelingen, die Leistung von Google im Retrieval medizinisch relevanter Literatur zu übertreffen. Insofern ist der Einsatz der neuen Technologie als voller Erfolg zu bezeichnen.

Zu Ziel 4: *Optimierung der Benutzerfreundlichkeit (Usability) der Testsuchmaschine*. Die Erwartungshaltungen der User bzw. der jeweiligen Zielgruppe sollten stets mit bedacht werden. Eine Suchumgebung, die zwar sehr gute Ergebnisse liefert, aber ansonsten eher benutzerunfreundlich ist, würde ihr Ziel verfehlen. Mit verschiedenen Vorschlägen zur Optimierung der Usability hat das Evaluationsteam diesem Anspruch Rechnung getragen. Der durchgeführte Usability-Test konnte zeigen, dass die Verbesserungsmaßnahmen zur Unterstützung der User von der Gruppe der Mediziner sehr positiv aufgenommen wurden. Die Erkenntnisse trugen dazu bei, das Interaktionsdesign der Testsuchmaschine auf einer empirisch gestützten Basis zu verbessern. Insbesondere folgende Unterstützungsmaßnahmen wurden von den Testpersonen als positive Verbesserung der Benutzerfreundlichkeit erlebt: die „Autosuggest-Funktion“, die Einführung eines „Schiebereglers“ zur Eingrenzung des Suchraums sowie die Einblendung von „Verwandten Suchbegriffen“ zur Verfeinerung der Suche.

Aus Sicht des Autors sind Usability-Tests im Rahmen von Design-Veränderungen bei Suchmaschinenoberflächen unverzichtbar. Ohne empirische Rückmeldungen durch die User sowie kontinuierliche Analysen der Logfiles, bleiben Überlegungen zum Erfolg eines

Retrieval-Systems im Bereich der Spekulation. Deshalb kann vor dem Hintergrund der Erfahrungen des MorphoSaurus-Projekts nur empfohlen werden, bei der Entwicklung und Veränderung von Suchmaschinen oder Portalen, möglichst früh empirisch begründetes Wissen über die Nutzer zu sammeln. Von den Kosten-Nutzen-Relationen her gesehen, trägt eine frühzeitige Investition in eine bessere Benutzerfreundlichkeit wesentlich dazu bei, Produkte zu entwickeln, die sich am Markt behaupten können.

Entwicklungen und Tendenzen. Die ZB MED bietet mit GREENPILOT[31] seit Juni 2009, parallel zu MEDPILOT, ein eigenes Portal zu den Inhaltsbereichen „Ernährung, Umwelt und Agrarwissenschaften“ an. Die Entwicklung von GREENPILOT orientierte sich dabei an den positiven Erfahrungen im Zusammenhang mit dem Betrieb von MEDPILOT. Darüber hinaus ist hier erstmals der Versuch unternommen worden, die Möglichkeiten der MorphoSaurus-Technik auf eine andere Inhaltsdomäne zu übertragen. Die Schwierigkeit dabei war, die Heterogenität der Fachtermini aus den Bereichen Ernährung, Umwelt und Agrarwissenschaften in einen gemeinsamen domänenspezifischen Thesaurus zu übertragen, sodass hier die MorphoSaurus-Technologie eingesetzt werden konnte. Die notwendige fachspezifische Anpassung des Subwortthesaurus basierte wesentlich auf der Einbindung des sogenannten Agrovoc-Thesaurus (Multilingual agricultural thesaurus), der von der „Food and Agriculture Organization of the United Nation (FAO)“ entwickelt wurde. Erste Evaluationen der neuen Suchmaschine waren sehr ermutigend. Dennoch werden für ein ähnlich hochdifferenziertes Thesaurus-System, wie es bereits für den Bereich der Medizin vorliegt, Anpassungen durch inhaltliche Experten notwendig werden, um die Retrieval-Leistung noch weiter zu verbessern.

Die Implementation der MorphoSaurus-Technik in die bestehende MEDPILOT-Umgebung sowie die Integration moderner Suchmaschinentechnologie wird aufgrund der positiven Projektergebnisse vorangetrieben und von der EDV-Abteilung der ZB MED bis voraussichtlich Ende 2009 umgesetzt. Hierbei werden zunächst die Datenbanken MEDLINE, ZB MED OPAC und CC MED in einem gemeinsamen Index zusammengefasst. Ein Ausbau des Index mit zusätzlichen Datenbanken ist geplant. Die Erkenntnisse zur Akzeptanz von innovativen Unterstützungsfunktionen, die durch die Usability-Studie gewonnen wurden, können bei

[31] Bei GREENPILOT (www.greenpilot.de) handelt es sich um die Virtuelle Fachbibliothek für Ernährung, Umwelt und Agrar. Sie soll die wissenschaftlichen Informationen der Fächer Ernährungs-, Umwelt-, und Agrarwissenschaften bündeln und über ein Portal den interessierten Nutzern zur Verfügung stellen.

einem zukünftigen Relaunch von MEDPILOT wichtige Impulse für das Interfacedesign liefern.

Die Methoden, die zur Bewertung der Retrieval-Leistung erarbeitet worden sind sowie das Know-how, das bei der Optimierung und Evaluation der Usability erworben wurde, können auch für andere Projekte mit Gewinn eingesetzt werden und ersparen dadurch längere Anlaufzeiten für ähnlich gelagerte Vorhaben.

Empfehlungen für ein künftigen MEDPILOT-Relaunch. Auch wenn die Inhaltsanalyse gezeigt hat, dass vorwiegend originär medizinische Inhalte gesucht werden, wird sich in den kommenden Jahren das Wachstum der biomedizinischen Forschung stark beschleunigen. Mediziner, die in der Forschung aktiv sind, sind auch an Grenzgebieten der Medizin interessiert (z. B Biochemie, Chemie, Biophysik, Gentechnik, Psychologie, Sozialwissenschaften etc.). Deshalb wäre hier langfristig an eine Erweiterung des MorphoSaurus-Subwort-Lexikons zu denken, da diese Technologie erheblich zur Verbesserung der Verarbeitung auch heterogener Wissensinhalte beitragen kann. Ebenso wäre es wünschenswert, den indexbasierten Teil der Suchmaschine (mit MorphoSaurus-Technik) auf möglichst viele der frei verfügbaren Datenbanken zu erweitern, um die Retrieval-Leistung weiter zu steigern.

Insgesamt haben die Evaluationsergebnisse gezeigt, dass die MorphoSaurus-Technologie sehr gut geeignet ist, wesentliche Probleme des medizinischen Information-Retrievals zu bewältigen. Insbesondere die Leistungsfähigkeit im Umgang mit Synonymen, Komposita und sprachlichen Variationen sowie die Fähigkeit zum Auffinden fremdsprachlicher Artikel heben dieses System von anderen Lösungsansätzen ab. Daher bietet diese Technologie verschiedene Alleinstellungsmerkmale, die die Attraktivität der MEPILOT-Nutzung deutlich erhöhen könnten. Aus den begleitenden Interviews im Rahmen der Usability-Untersuchung ging deutlich hervor, dass Mediziner sehr positiv auf die Übersetzungsfähigkeiten der Testsuchmaschine reagierten. Zudem ist es vorstellbar, dass mit einer zukünftig verbesserten Usability auch neue Nutzergruppen für MEDPILOT erschlossen werden könnten. Gerade auch Personen mit eher geringerer Interneterfahrung sowie Studierende, die noch keine inhaltlichen Experten ihrer Domäne sind, würden von einem solchen, auf optimale Benutzerfreundlichkeit ausgerichteten System, profitieren. Der Einstieg in die professionelle medizinische Literaturrecherche setzt aufseiten der Nutzer die Überwindung von Hemmschwellen voraus. Ein MEDPILOT, das durch die Optimierung der Trefferrelevanz und der Gebrauchstauglichkeit

den Nutzern bezüglich ihrer Erwartungen bestmöglich entgegenkommt, hätte die besten Voraussetzungen, die Standardsuchmaschine für den Bereich der medizinischen Literaturrecherche zu werden.

Zum Schluss soll noch ein kleiner Ausblick auf *generelle Tendenzen* im Information-Retrieval gegeben werden. Sehr interessant zeigt sich die aktuelle Entwicklung im Bereich der biomedizinischen Informationsverarbeitung. Aktive Projekte auf diesem Gebiet verfolgen u.a. ontologiebasierte Suchansätze (vgl. z. B. Dietze & Schroeder, 2009). Doch bisher zeichnet sich keine umfassende Lösung für die vielfältigen Probleme im Information-Retrieval ab. Es lässt sich jedoch die Tendenz erkennen, dass eine Kombination aus Techniken, wie dem MorphoSaurus-Ansatz, Textmining-Methoden, dem Natural Language Processing (NLP) und den semantik- bzw. ontologiebasierten Ansätzen den größten Erfolg verspricht. Am Ende dieser Entwicklung könnte ein System stehen, dass es Suchmaschinen eines Tages ermöglicht, natürlichsprachliche Anfragen zu „verstehen" und adäquat zu beantworten. Doch bis dahin scheint es noch ein weiter Weg zu sein.

Andere Projekte zeigen, dass die Berücksichtigung der psychologischen Erkenntnisse des Nutzerverhaltens sowie technologieorientierte Hilfen, wie die Nutzerunterstützung durch Sprachtechnologien immer breitere Akzeptanz finden. Ein Beispiel dafür ist die Weiterentwicklung der Webinitiative der australischen Regierung, die mit „HealthInsite" [32] ein Portal geschaffen hat, in dem Verbraucher nach vertrauenswürdigen Gesundheitsinformationen suchen können. Buckley-Smith und Deacon (2008) berichten z. B. vom Einsatz eines Thesaurus zur Unterstützung der Nutzer bei der Suche.

Aber auch in der Bibliothekslandschaft setzt sich der Trend zu einem modernen, nutzerorientierten Information-Retrieval immer mehr durch (vgl. z. B. Blenkle, Ellis & Haake, 2009a, 2009b). Die Staats- und Universitätsbibliothek Bremen (SuUB Bremen) bietet ihren Nutzern mit der Elektronischen Bibliothek (E-LIB[33]) verschiedene Unterstützungsmöglichkeiten an. Dazu gehören auch eine Autosuggest-Funktion und Web 2.0-Applikationen wie die Integration von Wikipedia, LibraryThing und Google Books. Darüber hinaus kann ein Suchbegriff über eine Wortwolke (Tag Cloud) in seinem semantischen Umfeld dargestellt werden,

[32] Homepage: http://www.healthinsite.gov.au/

[33] Homepage: http://suche3.suub.uni-bremen.de/

was wiederum Anlass zur Verfeinerung der Suche sein kann. Interessant scheinen zusätzliche Funktionen, wie die „Publikationsgeschichte des Suchbegriffes“, wobei nach einer Suche die zeitliche Verteilung der Treffer als Grafik angezeigt wird. So erhalten die Nutzer Informationen zum chronologischen Verlauf des Auftretens eines Begriffes in der Literatur. Eine weitere wichtige Funktion aus Sicht der Nutzer sind die „Favoriten“. Dabei werden die stark nachgefragten Titel in Verbindung mit einem Suchbegriff zurückgemeldet.

Diese kleine Auswahl von Beispielen sollte zeigen, dass Bibliotheken und andere öffentliche Einrichtungen darum bemüht sind, neueste Entwicklungen im Bereich des Information-Retrieval aufzunehmen und im Sine einer größeren Kundenzufriedenheit in ihren Portalen umzusetzen. Mit seinem spezifischen Ansatz stellt das MorphoSaurus-Projekt daher ein gelungenes Beispiel für die Innovationsfähigkeit von Bibliotheken im Bereich der öffentlichen Informationsversorgung dar.

8 Literatur

Alpert, J. & Hajaj, N. (2008). We knew the web was big... Posting on Googleblog (25.07.2008). *Onlinedokument*. Online verfügbar unter: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html (Stand 01.10.2009)

Aly, A. A. (2008). Using a Query Expansion Technique to Improve Document Retrieval. *International Journal "Information Technologies and Knowledge"*, 2 (4), 343-348.

Anderson, J. R. & Bower, G. H. (1972). Recognition and retrieval processes in free recall, *Psychological Review, 79*, 97-123.

Anick, P. & Kantamneni, R. (2008). A longitudinal study of real-time search assistance adoption. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR´08)*. ACM Press, New York, NY, USA, 701-702.

Barnett, V. & Lewis, T. (1978). *Outliers in statistical data*. New York: Wiley.

Beitzel, S., Jensen, C., Chowdhury, A., Grossman, D. & Frieder, O. (2004). Hourly Analysis of a Very Large Topically Categorized Web Query Log. *Proceedings of the 2004 ACM SIGIR Conference on Research and Development in Information Retrieval (ACM-SIGIR)*, Sheffield, UK, 321-328.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. *Scientific American, 284* (5), 34-43.

Blenkle, M., Ellis, R. & Haake, E. (2009a). E-LIB Bremen – Automatische Empfehlungsdienste für Fachdatenbanken im Bibliothekskatalog / Metadatenpools als Wissensbasis für bestandsunabhängige Services. *Bibliotheksdienst, 43*(6), 618-627. Online verfügbar unter: http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte2009/Erschliessung010609BD.pdf (Stand 01.10.2009)

Blenkle, M., Ellis, R. & Haake, E. (2009b). Next-generation library catalogues: review of E-LIB Bremen. *The Journal for the Serials Community, 22*(2), 178-181.

Buckley Smith, J & Deacon, P. (2008). Thesaurus support for quality health information searching on HealthInsite. *11th European Conference of Medical and Health Libraries (EAHIL 2008)*, 23th-28th of June 2008, Helsinki, Finland. Online verfügbar unter:

http://www.eahil.net/conferences/helsinki_2008/www.terkko.helsinki.fi/bmf/EAHILpapers/Jill_Buckley_Smith_paper.pdf (Stand 01.10.2009)

Butzlaff, M., Telzerow, A., Lange, S. & Krüger, N. (2001). Ärzte, Internet und neues Wissen. Nutzung und Effizienz von neuen Weiterbildungsmedien im Krankenhaus. *Medizinische Klinik*, 96 (6), 309-320.

Christmann, U. & Groeben, N. (1999). Psychologie des Lesens. In B. Franzmann, K. Hasemann, D. Löffler & E. Schön (Hrsg.), *Handbuch Lesen* (S. 145-223). München: Saur.

Daumke, P. (2007). *Das MorphoSaurus-System. Lösungen für die linguistischen Herausforderungen des Information Retrievals in der Medizin*. Universität Freiburg. Online verfügbar unter: http://www.freidok.uni-freiburg.de/volltexte/4932/ (Stand 01.10.2009)

Daumke, P., Schulz, S., Müller, M. L., Dzeyk, W., Pacheco, E. J., Cancian, P. S., Nohama, P. & Markó, K. (in Druck). Subword-based Semantic Retrieval of Clinical and Bibliographic Documents. *Methods of Information in Medicine*.

Davies, K. (2007). The information-seeking behaviour of doctors: a review of evidence. *Health Information and Libraries Journal, 24*, 78-94.

Dietze, H. & Schroeder, M. (2009). GoWeb: a semantic search engine for the life science web. *BMC Bioinformatics, 10* (10):S7. Online verfügbar unter: http://www.biomedcentral.com/1471-2105/10/S10/S7 (Stand 08.10.2009)

Dzeyk, W. (2001). Ethische Dimensionen der Onlineforschung. *Kölner Psychologische Studien, Jahrgang VI*, Heft 1, 1-30. Online verfügbar unter: http://kups.ub.uni-koeln.de/volltexte/2008/2424/ (Stand 01.10.2009)

Dzeyk, W. & Markó, K. (2008). Optimizing and evaluating the MEDPILOT search engine. Boosting medical information retrieval by using a morpheme thesaurus. *Journal of EAHIL (Journal of the European Association for Health Information and Libraries), 4* (1), 14-19. Online verfügbar unter: http://www.eahil.net/newsletter/journal_2008_vol4_n1.pdf (Stand 01.10.2009)

El-Menouar, Y. (2002). *Was erwarten Nutzerinnen und Nutzer vom Internet-Angebot medizinischer Bibliotheken? Ergebnisse einer internetbasierten Umfrage*. Köln: Deutsche

Zentralbibliothek für Medizin. Online verfügbar unter: http://www.zbmed.de/fileadmin/pdf_dateien/menouar_02.pdf (Stand 01.10.2009)

El-Menouar, Y. (2004). *Evaluation der Virtuellen Fachbibliothek Medizin „MedPilot". Ergebnisse einer internetbasierten Nutzerbefragung*. Köln: Deutsche Zentralbibliothek für Medizin. Online verfügbar unter: http://www.zbmed.de/fileadmin/pdf_dateien/medpilot_Evaluationsstudie_2004.pdf (Stand 01.10.2009)

eResult-Studie (2008). Wording-Studie 3.0 - Verständnisprobleme: Wichtige Web 2.0 und E-Commerce Begriffe sind Nutzern immer noch unklar!, *Studie*. Online verfügbar unter: http://www.eresult.de/studien_artikel/studienbaende/wording_studie_3_0.html (Stand 01.10.2009)

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised edition). Cambridge, MA: MIT Press.

Erlhofer, S. (2008). *Suchmaschinenoptimierung für Webentwickler* (4. Aufl.). Bonn: Galileo Press.

Eshet-Alkali, Y. & Amichai-Hamburger, Y. (2004). Experiments in digital literacy. *Cyberpsychology & Behavior, 7* (4), 421-430.

Fischer, M. (2006). *Website Boosting* (1. Aufl.). Frechen: mitp-Verlag.

Fourie I. (2006). Learning from web information seeking studies: some suggestions for LIS practitioners. *The Electronic Library, 24* (1), 20-37.

Friedel, J. E. F. (2007). *Reguläre Ausdrücke* (3. Aufl.). Köln: O`Reilly.

Gilster, P. (1997). *Digital literacy*. New York: Wiley.

Giustini, D. & Barsky, E. (2005). A look at Google Scholar, PubMed, and Scirus: comparisons and recommendations. *JCHLA/JABSC*, 26, 85-89. Online verfügbar unter: http://pubservices.nrc-cnrc.ca/rp-ps/absres.jsp?jcode=jchla&ftl=c05-030&lang=eng (Stand 01.10.2009)

Glotz, P. (2001). Medienkompetenz als Schlüsselqualifikation. In I. Hamm (Hrsg.), *Medienkompetenz* (S. 16-37). Gütersloh: Verlag Bertelsmann Stiftung.

Griesbaum, J., Bekavac, B. & Rittberger, M. (2009). Typologie der Suchdienste im Internet. In D. Lewandowski (Hrsg.), *Handbuch Internet-Suchmaschinen* (S. 18-52). Heidelberg: AKA Verlag.

Grimes, C., Tang, D. & Russell, D. M. (2007). Query logs alone are not enough. WWW 2007, *Workshop on Query Logs Analysis: Social and Technological Challenges*, Banff, Canada.

Groeben, N. (1982). *Leserpsychologie. Textverständnis – Textverständlichkeit*. Münster: Aschendorff.

Groeben, N. (2004). Medienkompetenz. In R. Mangold, P. Vorderer & G. Bente (Hrsg.), *Lehrbuch der Medienpsychologie* (S. 27-49). Göttingen: Hogrefe.

Groeben, N. & Rustemeyer, R. (2001). Inhaltsanalyse. In E. König & P. Zedler (Hrsg.), *Qualitative Forschung* (S. 233-258). Weinheim: UTB.

Hahn, U., Wermter, J., DeLuca, D. S., Blasczyk. R., Poprat, M., Bajwa, A. & Horn, P. (2007). StemNet: An Evolving Service for Knowledge Networking in Life Sciences. *Proceedings of the German e-Science Conference 2007 (GES2007)*; 2.-4. Mai 2007; Baden-Baden, Deutschland. Online verfügbar unter: http://edoc.mpg.de/316590 (Stand 01.10.2009)

Heijnk, S. (2002). *Texten fürs Web. Grundlagen und Praxiswissen für Online-Redakteure*. Heidelberg: dpunkt.verlag.

Heindl, E. (2003). *Logfiles richtig nutzen. Webstatistiken erstellen und auswerten*. Bonn: Galileo Press.

Heinold, E. F. & Spiller, U. (2007). Virtuelle Fachbibliotheken im System der überregionalen Literatur- und Informationsversorgung. *Studie zu Angebot und Nutzung der Virtuellen Fachbibliotheken*. ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften und Heinold, Spiller & Partner Unternehmensberatung GmbH. Online verfügbar unter: http://www.zbw.eu/ueber_uns/projekte/vifasys/gutachten_vifasys_2007_3_5.pdf (Stand 01.10.2009)

Hellbusch, J. E. & Mayer, T. (2006). *Barrierefreies Webdesign – Webdesign für Menschen mit körperlichen Einschränkungen* (4. Aufl.). Osnabrück: Know-Ware Verlag.

Hersh, W. (2004). Health care information technology: progress and barriers. *Journal of the American Medical Association, 292*, 2273-2274.

Herskovic, J. R., Tanaka, L. Y., Hersh, W. & Bernstam, E. V. (2007). A day in the life of PubMed: analysis of a typical day's query log. *J Am Med Inform Assoc, 14*, 212-220.

Höchstötter, N. (2007). Suchverhalten im Web – Erhebung, Analyse und Möglichkeiten. *Information: Wissenschaft und Praxis, 58* (3), 135-140.

Hölscher, C. & Strube, G. (2000). Web search behavior of Internet experts and newbies. *Proceedings of the 9th International World Wide Web Conference, 33* (1-6), 337-346.

Huffman, S. B. & Hochster, M. (2007). How Well does Result Relevance Predict Session Satisfaction? *SIGIR'07*, July 23–27, 2007, Amsterdam, The Netherlands: ACM.

iProspect (2006). *Search Engine User Behavior Study*. Online verfügbar unter: http://www.iprospect.com/about/whitepaper_seuserbehavior_apr06.htm (Stand 01.10.2009)

Jacobsen, J. (2005). *Website-Konzeption. Erfolgreich Web- und Multimedia-Anwendungen entwickeln* (3. erw. Auflage). München: Addison-Wesley.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research, 28*, 407-432.

Jansen, B. J., Spink, A. & Saracevic, T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing & Management, 36* (2), 207-227.

Jansen, B. J., Spink, J. & Pedersen, J. (2005). A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology*, 56 (6), 559-570.

Jelitto, M. (2007). *Evaluation von Web-Auftritten im Gesundheitswesen. Strategien der Qualitätssicherung.* Eine Expertise im Auftrag der Bundeszentrale für gesundheitliche Aufklärung (BZgA). Köln: BZgA.

Kaczmirek, C. (2003). Gebrauchstauglichkeit der Ergebnisseiten von Suchmaschinen. In: G. Szwillus, J. Ziegler (Hrsg.), *Mensch & Computer 2003: Interaktion in Bewegung* (S. 337-347). Stuttgart: B. G. Teubner.

Kalbach, J. (2008). *Handbuch der Webnavigation. Die User-Erfahrung optimieren*. Beijing; Köln [u.a.]: O'Reilly.

Krause, J. & Mayr, P. (2006). Allgemeiner Bibliothekszugang und Varianten der Suchtypologie - Konsequenzen für die Modellbildung in vascoda. *IZ-Arbeitsbericht*, Nr. 38. Bonn: IZ Sozialwissenschaften.

Krug, S. (2002). *Don't make me think! Web Usability – Das intuitive Web* (1. Auflage). Bonn: mitp-Verlag.

Langer, I., Schulz von Thun, F. & Tausch, R. (1993). *Sich verständlich ausdrücken*. München: Reinhardt.

Leroy, G., Xu, J., Chung, W., Eggers, S. & Chen, H. (2007). An end user evaluation of query formulation and results review tools in three medical meta-search engines. *International Journal of Medical Informatics, Volume 76* (11-12), 780-789.

Lewandowski, D. (2005). *Web Information Retrieval: Technologien zur Informationssuche im Internet*. Frankfurt am Main: DGI, 2005 (Informationswissenschaft; 7), Online verfügbar unter: http://www.durchdenken.de/lewandowski/web-ir/download/Web-IR-Buch.pdf (Stand 01.10.2009)

Lewandowski, D. (2006). Zur Bewertung der Qualität von Suchmaschinen. In J. Eberspächer & S. Holtel (Eds.), *Suchen und Finden im Internet* (pp. 195-199). Heidelberg: Springer.

Lewandowski, D. (2007a). Mit welchen Kennzahlen lässt sich die Qualität von Suchmaschinen messen? In M. Machill und M. Beiler (Hrsg.), *Die Macht der Suchmaschinen - The Power of Search Engines* (S. 243-258). Köln: Halem.

Lewandowski, D. (2007b). Nachweis deutschsprachiger bibliotheks- und informationswissenschaftlicher Aufsätze in Google Scholar. *IWP - Information: Wissenschaft und Praxis, 58* (3), 165-168.

Lewandowski, D. (2009). Spezialsuchmaschinen. In D. Lewandowski (Hrsg.), *Handbuch Internet-Suchmaschinen* (S. 3-17). Heidelberg: AKA Verlag.

Lewandowski, D. & Höchstötter, N. (2008). Web Searching: A Quality Measurement Perspective. In A. Spink & M. Zimmer (Eds.), *Web Search. Multidisciplinary Perspectives* (pp. 309-340). Berlin: Springer.

Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., Granka, L., Pellacini, F. & Pan, B. (2008). Eye tracking and online search: Lessons learned and

challenges ahead. *Journal of the American Society for Information Science & Technology, 59* (7), 1041-1052.

Maaß, C., Skusa, A., Heß, A. & Pietsch, G. (2009). Der Markt für Internetsuchmaschinen. In D. Lewandowski (Hrsg.), *Handbuch Internet-Suchmaschinen* (S. 3-17). Heidelberg: AKA Verlag.

Machill, M. & Welp, C. (2003). *Wegweiser im Netz. Qualität und Nutzung von Suchmaschinen*. Gütersloh: Verlag Bertelsmann Stiftung.

Markó, K. (2008). *Foundation, Implementation and Evaluation of the MorphoSaurus System. Subword Indexing, Lexical Learning and Word Sense Disambiguation for Cross-Language Information Retrieval*. Friedrich-Schiller-Universität Jena. Online verfügbar unter: http://www.db-thueringen.de/servlets/DocumentServlet?id=12506 (Stand 01.10.2009)

Navarro-Prieto, R., Scaife, M. & Rogers, Y. (1999). Cognitive Strategies in Web Searching. *Proceedings of the 5thConference on Human Factors and the Web*, Juni 1999, Gaithersburg. Online verfügbar unter: http://zing.ncsl.nist.gov/hfweb/proceedings/navarro-prieto/index.html (Stand 01.10.2009)

Nielsen, J. (1993). *Usability Engineering*. Boston: Academic Press.

Nielsen, J. & Loranger, H. (2006). *Web Usability*. München: Addison-Wesley.

OneStat.com (2007). Most people use 2 and 3 word phrases in search engines according to OneStat.com. *Onlinedokument* (31.10.2007). Online verfügbar unter: http://www.onestat.com/html/aboutus_pressbox56-word-phrases-in-search-engines.html (Stand 01.10.2009)

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G. & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication, 12* (3). Online verfügbar unter: http://jcmc.indiana.edu/vol12/issue3/pan.html (Stand 01.10.2009)

Price, D. (1963). *Little Science, Big Science*. New York: Columbia University Press.

Reng, C. M., Friedrich, H. J., Timmer, A. & Schölmerich J. (2003). Fachinformationen für Ärztinnen und Ärzte in Deutschland. Akzeptanz, Qualität und Verfügbarkeit von Fach-

informationen unter besonderer Berücksichtigung der elektronischen Informationsmedien. *Medizinische Klinik, 98* (11), 648-655.

Rustemeyer, R. (1992). *Praktisch-methodische Schritte der Inhaltsanalyse. Eine Einführung am Beispiel der Analyse von Interviewtexten*. Münster: Aschendorff.

Schmider, E. (2003). *Handbuch für Webtexter. So schreiben Sie fürs Internet*. Berlin, Heidelberg: Springer.

Schmidt-Mänz, N. (2007). *Untersuchung des Suchverhaltens im Web – Interaktion von Internetnutzern mit Suchmaschinen*. Hamburg: Verlag Dr. Kovac.

Schmidt-Maenz, N. & Bomhardt, C. (2005). Wie suchen Onliner im Internet? *Science Factory/Absatzwirtschaft 2*, 5-8.

Schneider, S. (2004). Virtuelle Fachbibliothek Medizin: Effiziente medizinische Recherche. *Deutsches Ärzteblatt, 101* (43), 2870-2874. Online verfügbar unter: http://www.aerzteblatt.de/v4/archiv/artikel.asp?src=suche&p=schneider+medpilot+2004&id=43955 (Stand 01.10.2009)

Schreier, M. & Appel, M. (2002). Realitäts-Fiktionsunterscheidungen als Aspekt einer kritischkonstruktiven Mediennutzungskompetenz. In N. Groeben und B. Hurrelmann (Hrsg.), *Medienkompetenz. Voraussetzungen, Dimensionen, Funktionen* (S. 231-254). Weinheim: Juventa.

Schulz, U. (2001a). Search Engine Usability - über die Nutzungsqualität von Suchmaschinen. In: Schmidt, R. (Hrsg.), *Proceedings Information Research & Content Management; Orientierung, Ordnung und Organisation im Wissensmarkt*; 23. Online-Tagung der DGI, S. 74-83. Frankfurt a. M.: DGI. Online verfügbar unter: http://www.bui.haw-hamburg.de/pers/ursula.schulz/publikationen/searchengine_usability.pdf (Stand 01.10.2009)

Schulz, U. (2001b). Usability-Kriterien für Suchmaschinen. *NfD*, S. 467-469. Online verfügbar unter: http://www.bui.haw-hamburg.de/pers/ursula.schulz/publikationen/suchmakriterien.pdf (Stand 01.10.2009)

Schulz, U. (2002). "Das stiehlt meine Zeit" - Über die Nutzungsqualität von Bibliotheks-websites. *BuB, 54* (4), 224-229. Online verfügbar unter: http://www.bui.haw-hamburg.de/pers/ursula.schulz/publikationen/bibliothekswebsites.pdf (Stand 01.10.2009)

Shneiderman, B. (1994). Dynamic queries for visual information seeking. *IEEE Software 11* (6), 70-77.

Shultz, M. (2007). Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association* (JMLA), 95 (4), 442-445.)

Spink, A. & Jansen, B. J. (2004). *Web Search: Public Searching of the Web*. Dordrecht: Kluwer.

Spink, A., Jansen, B. J., Wolfram, D. & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer, 35* (3), 107-109.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

vascoda-Usability-Studie, eResult (2007). Ergebnisband. Eine Usability-Studie zum Internetportal www.vascoda.de im Auftrag der ZB MED: Deutsche Zentralbibliothek für Medizin (www.zbmed.de). *Onlinedokument*. Online verfügbar unter: http://edok01.tib.uni-hannover.de/edoks/e01vascoda/Studien_Untersuchungen/vascoda_Usability-Studie_Dez07.pdf (Stand 01.10.2009)

WebHits (2009). Web-Barometer. Online verfügbar unter: http://www.webhits.de/deutsch/index.shtml?webstats.html (Stand 01.10.2009)

Weist, D. (2004). *Accessibility – Barrierefreies Internet: Hintergründe, Technik, Lösungen für Menschen mit Behinderungen*. Berlin: VDM Verlag Dr. Müller.

White, H, Wright, T. & Chawner, B. (2006). Usability evaluation of library online catalogues. ACM International Conference Proceeding Series; Vol. 169, *Proceedings of the 7th Australasian User interface conference* - Volume 50, Hobart, Australia, 69-72.

Whitelaw, C, Hutchinson, B., Chung, G. Y. & Ellis, G. (2009). Using the Web for Language Independent Spellchecking and Autocorrection. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore*, 6-7 August, 890-899. Online verfügbar unter: http://www.aclweb.org/anthology/D/D09/D09-1093.pdf (Stand 01.10.2009)

Willson, R. & Given, L. M. (2008). The Effect of Misspellings on Information Retrieval in Online Public Access Catalogues. *Proceedings of the 36th annual conference of the*

Canadian Association for Information Science (CAIS), University of British Colombia, Vancouver, June 5-7, 2008. Online verfügbar unter: http://www.cais-acsi.ca/proceedings/2008/willson_2008.pdf (Stand 01.10.2009)

Wirth, T. (2004). *Missing Links. Über gutes Webdesign* (2., erw. Aufl.). Hanser: München, Wien.

Wirth, W. & Schweiger, W. (Hrsg.). (1999). *Selektion im Internet. Empirische Analysen zu einem Schlüsselkonzept*. Opladen: Westdeutscher Verlag.

Wolfram, D., Spink, A., Jansen, B. J. & Saracevic, T. (2001). Vox Populi: The public searching of the Web. *Journal of the American Society for Information Science and Technology, 52* (12), 1073–1074.

Zaiß, A., Graubner, B., Ingenerf, J., Leiner, F., Lochmann, U., Schopen, M. Schrader, U. & Schulz, S. (2004). Medizinische Dokumentation, Terminologie und Linguistik. In: T. Lehmann, (Hrsg.), *Handbuch der Medizinischen Informatik* (2. überarb. Aufl.). Hanser: München.

9 Anhang

A. Repräsentative Testkollektion

Im nachfolgenden Teil des Anhangs (Tabelle 23) wird die „repräsentative Testkollektion“ wiedergegeben. Die Testkollektion war die Grundlage für die Einschätzung der Trefferrelevanz im Benchmarking-Test.

- Die Tabelle enthält die 100 Suchanfragen der repräsentativen Testkollektion, davon wurden 50 zufällig aus den 142.922 Suchanfragen des Logfiles ausgewählt und durch weitere 50 Varianten dieser Originalsuchterme ergänzt (z. B. bioaktive stoffe im gemüse vs. "bioaktive stoffe" +gemüse).
- Zusätzlich zur Zufallsauswahl wurde bei der Zusammenstellung der repräsentativen Testkollektion auch berücksichtigt, mit welchen Anteilen bestimmte „Suchterm- bzw. Sprachphänomene“ im Logfile vertreten waren (auf Grund der Ergebnisse der Inhaltsanalyse).
- Die Bedeutung der 50 Varianten sind in Spalte 3 formlos beschrieben. Da, wo es für die Relevanzeinschätzung aus inhaltlichen Gründen nötig war, wurden für den Beurteiler Erklärungen bzw. Kommentare eingefügt.
- Spalte 6 enthält einen inhaltlichen Kommentar zur Variantenbildung aus dem Originalsuchterm.

Tabelle 23. Repräsentative Testkollektion. Suchterme, die für die Retrieval-Tests des Benchmarkings verwendet wurden. Aufgelistet sind jeweils die Originalsuchterme sowie die zugehörigen Varianten.

lauf. Nr.	Original-suchterm	Inhaltlicher Kommentar zur Suchanfrage oder methodischer Kommentar zur Variantenbildung	lauf. Nr.	Variante	Art der Ver-änderung
1	46. ICAAC	´+#{ICAAC San Francisco: 46. Interscience Conference on Antimicrobial Agents and Chemotherapy. In diesem Zusammenhang ist NUR die 46. Konferenz relevant, die vom 17-20. Sept. 2006 in San Francisco stattgefunden hat. NICHT relevant sind die ICAAC-Konferenzen in den Jahren 1974, 1995, 1999, welche ebenfalls in San Francisco stattgefunden haben.}	51	46. Interscience Conference on Antimicrobial Agents and Chemotherapy	Abkürzung aufgelöst
2	"anabolika"	´+#{Stereoid. Substanzen, die den Aufbau von körpereigenem Gewebe vorwiegend durch eine verstärkte Proteinsynthese fördern (die also eine so genannte anabole Wirkung haben) }	52	anabolika	Anführungsstriche weggelassen
3	anorexia nervosa osteoporose empfängnis-verhütung	´+#{In der Bewertung des Nutzen/Risiko Verhältnisses bei Frauen mit osteoporotischen Risikofaktoren sollten andere Methoden der Empfängnisverhütung als Depocon in Erwägung gezogen werden. Die Anwendung von Depocon kann bei Patientinnen mit Osteoporose-Risikofaktoren (z.B. Knochenstoffwechsel-Erkrankung, chronischer Alkohol- oder Tabakkonsum, Anorexia nervosa, schwere osteoporotische Erkrankungen in der Familiengeschichte oder chronische Anwendung von Arzneimitteln wie Antikonvulsiva oder Kortikosteroide, welche die Knochendichte reduzieren können) ein zusätzlicher Risikofaktor sein}	53	magersucht osteoporose empfängnisverhütung	Synonym magersucht eher laiensprachlich
4	"aspergersyndrom"		54	"asperger-syndrom"	Bindestrich eingefügt
5	atlas of human sperm	+#{Atlas of Human Sperm Morphology. by Roelof Menkveld. Purchase at: Amazon.com. ASIN: 0683059254}	55	"atlas of human sperm"	Anführungsstriche eingefügt, Phrasenbildung
6	"bauch	Autor – keine anatomische Bedeutung	56	bauch	Anführungsstriche links weggelassen
7	bioaktive stoffe im gemüse		57	"bioaktive stoffe" +gemüse	optimierte Suchanfrage durch Phrasenbildung und Booleschen Operator
8	Carotis-desobliteration	+#{Desobliteration: Syn.: Endarteriektomie; Beseitigen eines durch Thrombose, Embolie oder Arteriosklerose entstandenen Arterienverschlusses.}	58	Desobliteration Carotis	Wortzusammen-setzung aufgelöst = Dekomposition
9	computer gestützte Therapie		59	computergestützte Therapie	Wortzusammen-setzung gebildet
10	constant score power	+#{Entwicklung und Evaluierung eines Fragebogens basierend auf dem Constant-Murely Score zur Selbstbeurteilung der Schulterfunktion durch den Patienten. --- Die von C. R. Constant für ein irisches Kollektiv erhobenen alters- und geschlechtsabhängigen Normalwerte differieren aufgrund der Kraftwerte zum Teil deutlich vom deutschen Referenzkollektiv. --- Geht anstelle der Kraft das Drehmoment in den Constant-Score ein, entsprechen 50 Nm 25 Punkten. Die mit dem Constant-Score erhobenen Gesamtpunktwerte waren ...}	60	"constant-score" +power	optimierte Suchanfrage durch Phrasenbildung, Bindestrich, Booleschen Operator, also disambiguiert

lauf. Nr.	Original-suchterm	Inhaltlicher Kommentar zur Suchanfrage oder methodischer Kommentar zur Variantenbildung	lauf. Nr.	Variante	Art der Ver-änderung
11	dehydration im alter		61	dehydrierung im alter	Synonym
12	riffel pflegedokumentation	´+#{CC MED_MEDPILOT: Riffel, Thomas: Pflegedokumentation: Von der Pflege zur DV-gestützten Dokumentation. Erscheinungsjahr/ Band (Heft) Seitenzahl(en) 2000 / 2 (4) 61-84 Signatur: Zs.A 5131}	62	"pflegedokumentation" autor=riffel	Vergleich von freier Suche und Feldsuche (kann averbis Feldsuche?)
13	hepatopulmonary syndrome	´+#{Hepatopulmonary syndrome is a syndrome characterized by chronic liver disease with pulmonary vascular dilatations and reduced arterial oxygenation. It is seen in patients with advanced liver disease, such as cirrhosis or portal hypertension. Patients can present with platypnea (difficulty breathing in an upright position and relieved by laying down) due to opening of pulmonary vascular shunts, and similarly have orthodeoxia (reduction of blood oxygen levels when sitting upright).}	63	hepatopulmonales Syndrom	Übersetzung En -> De
14	herbert schraube	´+#{Herbert-Schraube. Diese Schraube ist ein speziell für Kahnbeinbrüche entwickeltes System, das an diesem problematischen Handwurzelknochen sich als Standardverfahren etabliert hat.}	64	herbert-schraube	Bindestrich eingefügt = Rechtschreibung korrigiert, also disambiguiert
15	"incineration" +"fats"	´+#{incineration = die Verbrennung}	65	"Verbrennung" +"Fette"	Übersetzung En -> De
16	"inhaltsstoffe" +"der" +"kartoffel"		66	"inhaltsstoffe der kartoffel"	optimierte Suchanfrage durch Phrasenbildung
17	"kindeswohl*" +"diagnos*"		67	kindeswohl* +diagnos*	Anführungsstriche weggelassen
18	"meaureter"	´+#{Als Megaureteren bezeichnet man Harnleiter, die eine Weite von über 10 mm aufweisen. Dauerhafte abnorme Lichtungserweiterung des Harnleiters (Megaureter)}	68	"megaureter"	Rechtschreibung korrigiert
19	Pflegezeitschrift;juni 2002	´+#{Pflegezeitschrift - die Fachzeitschrift für den Berufsalltag in der stationären und ambulanten Pflege - So wird das gesuchte Heft Juni 2002 der Pflegezeitschrift in der CC MED angegeben, wobei 55 der Jahrgang ist und 453 eine beliebige Seite: 2002}	69	Pflegezeitschrift 6 2002	Semikolon durch Leerzeichen ersetzt. "Juni" durch Zahl 6 ersetzt.
20	pms therapie	´+#{Abkürzung für "prämenstruelles Syndrom" sowie für "pregnant mare serum gonadotropin")	70	prämenstruelles syndrom therapie	Abkürzung aufgelöst
21	"prostatatektomie"	´+#{Muss Prostatektomie heißen}	71	"prostatektomie"	Rechtschreibung korrigiert
22	sporttraumatologie		72	sport traumatologie	Wortzusammen setzung aufgelöst
23	steatorrhoe	´+#{Als Steatorrhoe (auch Steatorrhö) bzw. Pankreasstuhl oder Fettstuhl wird eine pathologische Erhöhung des Fettgehalts im Stuhl bezeichnet.}	73	Fettstuhl	Synonym (eher laiensprachlich)
24	stoßwellen	´+#{Die extrakorporale Stoßwellentherapie (ESWT) ist eine Stoßwellenbehandlung, mit der folgende Erkrankungen behandelt werden können …}	74	shock waves	Übersetzung De -> En

lauf. Nr.	Original-suchterm	Inhaltlicher Kommentar zur Suchanfrage oder methodischer Kommentar zur Variantenbildung	lauf. Nr.	Variante	Art der Ver-änderung
25	"syspur derm"	´+#{SYSpur-derm ist eine zweischichtige Kompresse aus Polyurethan-Weichschaum ohne medikamentöse Zusätze.}	75	"SYSpur-derm"	Bindestrich eingefügt = Rechtschreibung korrigiert
26	THA	´+#{Total hip arthroplasty (THA) is one of the most common and successful orthopedic procedures performed in the U.S., oder THA = tetrahydro-anthracene}	76	total hip arthroplasty	Abkürzung aufgelöst
27	tumor*		77	tumor	Rechtstrunkierung weggelassen (trunkiert die Suchmaschine automatisch?)
28	"wirelsäule"		78	"wirbelsäule"	Rechtschreibung korrigiert
29	adherence to antiretroviral therapy n update of current concepts	´+#{Rechtschreibung: An update --- Adherence to Antiretroviral Therapy: An Update of Current Concepts. Gregory M. Lucas, MD, PhD, Albert W. Wu, MD, MPH, and Laura W. Cheever, MD ScM ...}	79	adherence to antiretroviral therapy an update of current concepts	Rechtschreibung korrigiert
30	baha	´+#{BAHA = Akronym für Bone Anchored Hearing Aid}	80	bone anchored hearing aid	Abkürzung aufgelöst
31	copd"	´+#{Chronisch obstruktive Lungenerkrankung – engl.: chronic obstructive pulmonary disease, Abkürzung: COPD, seltener auch chronic obstructive lung disease, COLD – bezeichnet als Sammelbegriff eine Gruppe von Krankheiten, die durch Husten, vermehrten Auswurf und Atemnot bei Belastung gekennzeichnet sind. In erster Linie sind die chronisch-obstruktive Bronchitis und das Lungenemphysem zu nennen. Beide Krankheitsbilder sind dadurch gekennzeichnet, dass vor allem die Ausatmung (Exspiration) behindert ist. Umgangs-sprachliche Bezeichnungen sind „Raucherlunge“ für die COPD und „Raucherhusten“ für das Hauptsymptom.}	81	Chronic Obstructive Pulmonary Disease	Abkürzung aufgelöst
32	Halluzination nach Konsum von Spitzkegeliger Kahlkopf		82	Halluzination nach Konsum von Spitzkegeligem Kahlkopf	grammatikalische Varianten: reformulierte Suchanfrage
33	harnblasensenkung und nierenstau	´+#{Fachbegriff für Harnblasensenkung = Zystozele. Fachbegriff für Nierenstau ungefähr = Hydronephrose (genau genommen für Sackniere)}	83	harnblasensenkung +nierenstau	Boolescher Operator
34	harnröhrenklappen	´+#{Als Urethralklappen (Syn. Harnröhrenklappen) bezeichnet man segelartige Vorsprünge in der Harnröhre (Urethra), die unterhalb des Samenhügels liegen und den Harnfluss behindern. Im Extremfall findet sich ein vollständiger Verschluss der Harnröhre.}	84	harnröhrenklappe	Plural zu Singular

lauf. Nr.	Original-suchterm	Inhaltlicher Kommentar zur Suchanfrage oder methodischer Kommentar zur Variantenbildung	lauf. Nr.	Variante	Art der Ver-änderung
35	wegener's disease	´+#{The classical type of Wegener's granulomatosis encompasses a necrotizing granulomatous vasculitis of the upper and lower respiratory tracts and a necrotizing glomerulonephritis. As a rule, an initial involvement of the upper respiratory system in the nasal mucosa and paranasal sinuses is found. The terms "granuloma gangraenescens" and "midline disease" are frequently used at this stage of the disease. However, they should be abandoned and "Wegener's granuloma" preferred instead.Wegener's granulomatosis is part of a larger group of vasculitic syndromes, all of which feature the presence of an abnormal type of circulating antibody termed ANCAs (antineutrophil cytoplasmic antibodies) and affect small and medium-size blood vessels.}	85	Wegener's granulomatosis	Synonym (eher expertensprachlich)
36	Myelodysplastic Syndrone	´+#{Rechtschreibung: Myelodysplastic syndroMe: A group of bone marrow disorders characterized by the underproduction of one or more types of blood cells due to dysfuntion of the ...}	86	Myelodysplastic Syndrome	Rechtschreibung korrigiert
37	Oberarmkopf-prothese		87	Oberarmkopf Prothese	Wortzusammen-setzung aufgelöst
38	obere untere Extremitäten vergleich		88	vergleich der oberen und unteren extremität	reformulierte Suchanfrage
39	PET CT Herz	´+#{PET = Positronen-Emissions-Tomographie, CT = Computertomographie}	89	Positronen-Emissions-Tomographie CT Herz	Abkürzung aufgelöst
40	vitiligo	´+#{Vitiligo (Leucopathia acquisita) oder auch Weißfleckenkrankheit sowie Scheckhaut genannt ist eine chronische, nicht ansteckende Hauterkrankung}	90	Leucopathia acquisita	Synonym (eher expertensprachliches Synonym)
41	Clinical and Demographic Predictors of Long-term Disability in Patients With Relapsing-Remitting Multiple Sclerosis	+#{Langer-Gould A, Popat RA, Huang SM, Cobb K, Fontoura P, Gould MK, Nelson LM "Clinical and Demographic Predictors of Long-term Disability in Patients With Relapsing-Remitting Multiple Sclerosis: A Systematic Review." Arch Neurol. 2006; 63: 12: 1686-91.}	91	Predictors Long-term Disability Relapsing-Remitting Multiple Sclerosis	Suche nach genau einem Titel, jedoch nur mit Teilen der Phrase
42	"Familienstrukturen"		92	Familien Strukturen	Wortzusammen-setzung aufgelöst, Anführungsstriche entfernt
43	"gentech*" +"*pflanze" +"*gesetz*"	´+#{Variation der Suchanfrage (mit Trunkierung links und rechts) durch Weglassen der Anführungszeichen}	93	gentech* +*pflanze +*gesetz*	Anführungsstriche entfernt
44	plomin	´+#{Robert Plomin is an American psychologist best known for his work in twin studies and behavior genetics.}	94	Plomin R.	Anfangsbuchstabe des Vornamens ergänzt

lauf. Nr.	Original-suchterm	Inhaltlicher Kommentar zur Suchanfrage oder methodischer Kommentar zur Variantenbildung	lauf. Nr.	Variante	Art der Ver-änderung
45	Running velocities and heart rates at fixed blood lactate concentrations in young soccer players	´+#{Running velocities and heart rates at fixed blood lactate concentrations in young soccer players. Guner R, Kunduracioglu B, Ulkar B.}	95	Running velocities and heart rates soccer players	Suche nach genau einem Titel, jedoch nur mit Teilen der Phrase
46	kunsttherapie		96	kunsttherapei	Rechtschreibfehler willkürlich eingebaut
47	Leitlinien für die vrkehrsmedizinische Begutachtung	´+#{Rechtschreibung vErkehrsmedizinische}	97	Leitlinien für die verkehrsmedizinische Begutachtung	Rechtschreibung korrigiert
48	OSAS bei Kindern		98	obstruktives Schlafapnoe-Syndrom bei Kindern	Abkürzung aufgelöst
49	Weltgesundheitsorganisation	´+#{gleiche Treffer bei Weltgesundheitsorganisation und welt gesundheit organisation (aufgelöstes Wort)?}	99	Welt gesundheit organisation	Wortzusammen-setzung aufgelöst = Dekomposition
50	bechterew OR spondylitis ankylosans	´+#{Morbus Bechterew synonym zu Spondylitis ankylosans}Operator OR (für die Suche nach synonymen Begriffen). Die Spondylitis ankylosans (latinisiertes Griechisch: Spondylitis „Wirbelentzün-dung“ und ankylosans „versteifend“) ist eine chronisch entzündliche rheumatische Erkrankung mit Schmerzen und Versteifung von Gelenken. Synonyme sind Morbus Bechterew oder Bechterewsche Krankheit (nach Wladimir Michailowitsch Bechterew), Bechterew-Strümpell-Marie-Krankheit, ankylosierende Spondylitis, rheumatoide Spondylitis und Spondylarthritis ankylopoetica. Sie gehört zur Gruppe der Erkrankungen der Wirbelsäulengelenke (Spondylarthropathien) und betrifft vorwiegend die Lenden- und Brustwirbelsäule und die Kreuz-Darmbeingelenke. Außerdem kann es auch zu Entzündungen der Regenbogenhaut des Auges und selten auch anderer Organe kommen.}	100	bechterew spondylitis ankylosans	Boolescher Operator entfernt

B. Eingangsfragebogen

Sehr geehrte Untersuchungsteilnehmerin, sehr geehrter Untersuchungsteilnehmer, im Folgenden möchten wir Sie bitten, einige Fragen zu Ihrer Person und Ihrer Internetnutzung zu beantworten.
Vielen Dank für Ihre Mühe!

1. Ihr Geschlecht?

○ Weiblich ○ Männlich

2. Ihr Alter?

3. Geben Sie bitte Ihren beruflichen Hintergrund an.

○ Student ○ Abgeschlossenes Medizinstudium

4. Falls Sie Student/in sind: In welchem Semester sind Sie?

Fachsemester

5. Falls Sie ihr Studium schon abgeschlossen haben: Was ist Ihr Schwerpunkt?

○ Forschung und / oder Lehre ○ Sonstiges

○ Patientenversorgung

6. Falls Sie ihr Studium schon abgeschlossen haben: In welchem Fachgebiet sind Sie tätig?

7. Seit wie vielen Jahren benutzen Sie schon das Internet?

Jahre

8. Wie viele Stunden pro Woche nutzen Sie das Web (die mit E-Mails verbrachte Zeit nicht miteingerechnet)?

Stunden / Woche

9. Benutzen Sie das Web zum Chatten?

◉ Ja ○ Nein

10. Benutzen Sie Bookmarks?

○ Ja ◉ Nein

11. Nehmen Sie selbstständig Browserupgrades vor?

○ Ja ○ Nein

12. Können Sie Webseiten erstellen?

○ Ja ○ Nein

13. Können Sie Computerprobleme selbstständig beheben?

◉ Ja ○ Nein

14. Halten Sie sich über die Technik am Laufenden? Haben Sie beispielsweise Computermagazine abonniert?

○ Ja ○ Nein

15. Werden Sie von Freunden um Rat gefragt, wenn es um Computerprobleme geht?

○ Ja ○ Nein

16. Wie würden Sie ihre eigene Internetkompetenz einschätzen auf einer Skala von 1 bis 7 (wobei 1 bedeuten würde: eher gering und 7: eher hoch)?

Eher gering	1	2	3	4	5	6	7	Eher hoch
	○	○	○	○	○	○	○	

17. Wie würden Sie speziell ihre Fähigkeiten zur Literaturrecherche einschätzen (wieder auf einer Skala von 1 bis 7, wobei 1 bedeuten würde: eher gering und 7: eher hoch)?

Eher gering	1	2	3	4	5	6	7	Eher hoch
	○	○	○	○	○	○	○	

18. Wie schätzen Sie ihre Kompetenz im Umgang mit Suchmaschinen ein?

Anfänger	1	2	3	4	5	Experte
	○	○	○	○	○	

19. Kennen Sie MEDPILOT?

○ Ja ○ Nein

20. Falls Ihnen MEDPILOT bekannt ist: Wie häufig nutzen Sie MEDPILOT?

○ Täglich	○ Mehrmals pro Woche	○ Etwa einmal pro Woche	○ Etwa alle 14 Tage	○ Etwa ein bis zweimal pro Monat	○ Seltener als einmal pro Monat	○ Gar nicht	○ Keine Angabe

Vielen Dank für Ihre Mühe bei der Beantwortung unserer Fragen!

C. Ablauf der Usability-Untersuchung

Die Usability-Untersuchung dauerte pro Proband durchschnittlich eine Stunde. Der Ablauf gliederte sich in folgende Schritte:

- Vorbereitung der Untersuchung (5 Min.)
- Einstieg in die Homepage – Wahrnehmung der Grundinformationen (30 Sek.)
- Explorative Aufgabe (Recherche zu eigener Frage) (5Min.)
- Drei Rechercheaufträge (jeweils ca. 3 Min.)
- Interview zu den einzelnen Funktionen der Website (ca. 15-25 Min.)
    - Autosuggest-Funktion
    - Schieberegler (Einschränkung des Suchraums)
    - Verwandte Suchbegriffe
    - Verwandte Suchbegriffe – Farben für Kategorien
    - Verwandte Suchbegriffe – Grafik für Häufigkeit
    - Verwandte Suchbegriffe – Auf- und Zuklappen der Kategorien
    - Darstellung der Trefferliste
    - Rahmen Trefferzahlen der einzelnen Datenbanken (3. Spalte)
    - Einzeltrefferdarstellung
- Letzter Untersuchungsteil: Abschlussfragen (ca. 10 Min.)
    - Vergleich mit der Bewertung von MEDPILOT (El-Menouar, 2004)
    - Messung klassischer Usability-Aspekte
    - Abschlussfragebogen (Imagemessung – Vergleich mit vascoda)
- Debriefing bzw. Aufklärung
- Auszahlung der Aufwandsentschädigung

D. Abschlussfragebogen

Abschlussfragebogen

**Sehr geehrte Untersuchungsteilnehmerin, sehr geehrter Untersuchungsteilnehmer,
Sie hatten nun Gelegenheit, sich mit einer überarbeiteten Version von MEDPILOT auseinanderzusetzen. Nun möchten wir im Folgenden gerne Ihre Eindrücke diesbezüglich erfassen.**

1. Ich könnte mir vorstellen, das System häufig zu benutzen.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu
	○	○	○	○	○	

2. Das System war unnötig komplex.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu
	○	○	○	○	○	

3. Das System war einfach zu handhaben.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu
	○	○	○	○	○	

4. Ich denke, ich würde die Unterstützung einer Person benötigen, die sich mit dem System auskennt

stimme gar nicht zu	1	2	3	4	5	stimme voll zu
	○	○	○	○	○	

5. Die verschiedenen Funktionen des Systems waren gut aufeinander abgestimmt.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu
	○	○	○	○	○	

6. Es gab zu viele Widersprüche.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu
	○	○	○	○	○	

7. Die meisten Menschen würden den Umgang mit der Suchmaschine schnell erlernen.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu
	○	○	○	○	○	

8. Die Suchmaschine war sehr umständlich zu bedienen.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu
	○	○	○	○	○	

9. Ich fühlte mich sehr sicher im Umgang mit der Suchmaschine.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu
	○	○	○	○	○	

10. Ich musste mich erst an viele Dinge gewöhnen, bevor ich mit der Suchmaschine zurecht kam.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu
	○	○	○	○	○	

11. Welche Art der Sortierung würden Sie bei der Trefferliste bevorzugen?

- ○ Die zur Suchanfrage passendsten Treffer zuerst (Relevanz)
- ○ Alphabetische Sortierung der Treffer nach Autorenname
- ○ Die aktuellsten Treffer zuerst (Jahr)

12. Das Design ist ansprechend.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu

13. Die Suchmaschine macht insgesamt einen seriösen Eindruck.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu

14. Die Suchmaschine hat eine klare Navigation.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu

15. Der Seitenaufbau dauert durchschnittlich zu lang.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu

16. Ich habe die mich interessierenden Informationen gefunden.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu

17. Der Seitenaufbau ist übersichtlich.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu

18. Die Suchmaschine hat unklare Begriffe in der Bedienung.

stimme gar nicht zu	1	2	3	4	5	stimme voll zu

Im Folgenden möchten wir gerne von Ihnen wissen, mit welchen Eigenschaften Sie die Suchmaschine in Verbindung bringen.

19. Ich finde die Suchmaschine:

sympathisch	1	2	3	4	5	unsympathisch

20. Ich finde die Suchmaschine:

modern	1	2	3	4	5	altmodisch

21. Ich finde die Suchmaschine:

einfach	1	2	3	4	5	kompliziert

22. Ich finde die Suchmaschine:

interessant	1	2	3	4	5	langweilig

23. Ich finde die Suchmaschine:

hochwertig	1	2	3	4	5	minderwertig

24. Ich finde die Suchmaschine:

nützlich	1	2	3	4	5	nutzlos

25. Ich finde die Suchmaschine:

übersichtlich	1	2	3	4	5	unübersichtlich

26. Ich finde die Suchmaschine:

wichtig	1	2	3	4	5	unwichtig

27. Ich finde die Suchmaschine:

attraktiv	1	2	3	4	5	unattraktiv

28. Ich finde die Suchmaschine:

professionell	1	2	3	4	5	unprofessionell

29. Ich finde die Suchmaschine:

nutzerfreundlich	1	2	3	4	5	nutzerunfreundlich

Vielen Dank für Ihre Mühe bei der Beantwortung unserer Fragen!

Über das Buch:

Der Titel „Effektiv und nutzerfreundlich“ bringt die wesentlichen Erfolgskriterien für eine moderne webbasierte Literatur- und Informationssuche auf den Punkt. Eine Literatur-Suchmaschine im Wissenschaftsbereich muss relevante Treffer liefern und zugleich leicht bedienbar sein. Moderne Suchmaschinen sind mächtige Werkzeuge für die Informationssuche im Internet, dennoch scheitern auch sie regelmäßig an den sprachlich problematischen Suchanfragen der Nutzer. Dazu gehören:

- das Erkennen von relevanten fremdsprachlichen Treffern,
- der tolerante Umgang mit lexikalischen Varianten und Rechtschreibfehlern,
- das automatische Erkennen von synonymen Suchbegriffen oder zusammengesetzten Wörtern,
- die gleichwertige Behandlung von Laien- und Expertenanfragen sowie
- die korrekte Auflösung von Akronymen und Abkürzungen.

Im vorliegenden Buch beschreibt der Autor Waldemar Dzeyk die Ergebnisse eines Projekts an der Deutschen Zentralbibliothek für Medizin in Köln, das zum Ziel hatte, diese Schwachstellen von Suchmaschinen anzugehen. Unter Einsatz innovativer Sprachtechnologien und unter Einbeziehung von Erkenntnissen der Usability-Forschung wurde eine Testsuchumgebung exemplarisch für den Bereich der Medizin entwickelt und mit konkurrierenden Suchsystemen verglichen. Dabei zeigte sich, dass die Test-suchmaschine durch die von der Firma Averbis entwickelte Sprachtechnologie die meisten Treffer lieferte, den höchsten Anteil an relevanten Treffern bot und darüber hinaus die wenigsten Null-Treffer-Meldungen aufwies. Die Testsuchmaschine erreichte nicht nur bessere Ergebnisse als die von vielen Medizinern genutzte Suchumgebung PubMed, sondern war auch leistungsfähiger als das große Google.

Das Buch ist für alle ein Gewinn, die sich für die Anwendung moderner Sprachtechnologie in der fachwissenschaftlichen Suche interessieren. Darüber hinaus liefert die durchgeführte Usability-Studie viele aufschlussreiche Erkenntnisse zur optimalen Unterstützung der Nutzer im Suchprozess.

Über den Autor:

Waldemar Dzeyk studierte Psychologie, Germanistik sowie Theater- Film- und Fernsehwissenschaften in Bochum und war als wissenschaftlicher Mitarbeiter am Psychologischen Institut der Universität zu Köln tätig. Dort promovierte er zum Thema „Vertrauen in Internetangebote“. Als Projektleiter bei der Deutschen Zentralbibliothek für Medizin in Köln setzte er sich intensiv mit den Möglichkeiten der webbasierten Literatur- und Informationssuche auseinander. 2009 gründete er die Agentur XPERSITE, die sich mit Fragen der Usability und Onlinekommunikation beschäftigt.