GESIS

IZ InformationsZentrum Sozialwissenschaften

IZ-Arbeitsbericht Nr. 18

Projektbericht Vergleichsuntersuchung MESSENGER - FULCRUM

Gisbert Binder, Matthias Stahl, Lothar Faulborn

April 2000

InformationsZentrum
Sozialwissenschaften

Lennéstraße 30
D-53113 Bonn
Tel.: 0228/2281-0
Fax.: 0228/2281-120
email: binder@bonn.iz-soz.de
stahl@bonn.iz-soz.de
Internet: http://www.social-science-gesis.de

ISSN: 1431-6943

Herausgeber: Informationszentrum Sozialwissenschaften der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI)

Druck u. Vertrieb: Informationszentrum Sozialwissenschaften, Bonn
Printed in Germany

Inhalt

1 Fragestellung und Durchführung der Untersuchung

1.1 Fragestellung

Bei der vorliegenden Vergleichsuntersuchung zwischen den Retrievalinstrumenten MESSENGER (MES) und FULCRUM (FUL) handelt es sich um eine Studie im Rahmen des Projekts GIRT (German Indexing and Retrieval Testdatabase), bei dem es allgemein darum geht, die Leistungsfähigkeit moderner, intelligenter Indexierungs- und Retrievalsysteme im Vergleich mit herkömmlichen Standardsystemen zu überprüfen[1]. Im vorliegenden Fall geht es darum, in einem Benutzertest die Leistungsfähigkeit eines auf automatischer Indexierung basierenden Retrievalinstruments, das die Inhaltsdeskribierung des IZ nicht benutzt und ein nach Relevanz geranktes Suchergebnis liefert (FUL), mit der Standard-Freitextsuche, die um die intellektuell vom IZ vergebenen Deskriptoren ergänzt ist (MES), zu vergleichen.

1.2 Ergebnisse des Pretests

Die Vergleichsuntersuchung MES und FUL nimmt Ergebnisse und Erfahrungen eines Pretests auf, bei dem MES mit der Indexierungs- und Retrievalsoftware freeWAISsf, einem Vertreter des statistisch-quantitativen Retrievals, verglichen wurde. Die Ergebnisse sind veröffentlicht in einem Arbeitspapier von Elisabeth Frisch und Michael Kluck (vgl. Fußnote 1). Der Pretest hatte im wesentlichen die Aufgabe, ein für vergleichbare Folgestudien verwendbares methodisches Design zu entwickeln bzw. zu erproben. Die im Ergebnisbericht des Pretests vorgeschlagenen Veränderungen bei der Versuchsdurchführung (z.B. Fortfall der Begrenzung des Rechercheergebnisses auf 30 Treffer) wurden weitgehend berücksichtigt und werden in den folgenden Abschnitten dieses Berichts im einzelnen erläutert.

---

[1] Vgl. dazu: Elisabeth Frisch und Michael Kluck: Pretest zum Projekt German Indexing and Retrieval Testdatabase (GIRT) unter Anwendung der Retrievalsysteme Messenger und freeWAISsf, IZ-Arbeitsbericht Nr. 10, Bonn, 2.Auflage Oktober 1997.

1.3 Testdatenbank SOLFOR

Der Benutzertest wurde auf der Grundlage einer eigens für Testzwecke eingerichteten Datenbank SOLFOR durchgeführt. Dabei handelt es sich um Auszüge aus den sozialwissenschaftlichen Datenbanken SOLIS (Sozialwissenschaftliches Literaturinformationssystem) und FORIS (Forschungsinformationssystem Sozialwissenschaften) mit Schwerpunktsetzung auf den Themenbereichen Industrie- und Betriebssoziologie, Frauenforschung sowie Migration und ethnische Minderheiten aus den Jahren 1990 bis 1996. Die Testdatenbank enthält knapp 13.000 Dokumentationseinheiten (DE).

Die Datenbank liegt in zwei Versionen vor. Für FUL stehen als Suchfelder zur Verfügung: Titel und Untertitel, Kurzreferat bzw. Projektbeschreibung, normalerweise in deutsch, bei SOLIS-DE teilweise zusätzlich auch in englisch, Namen von Autoren bzw. Projektmitarbeitern, bei FORIS teilweise geographischer Raum. Die Version für MES enthält darüber hinaus zusätzliche Deskriptorenfelder, z.B. inhaltskennzeichnende Schlagwörter, die intellektuell vergeben worden sind. Im Vergleich zum Pretest stand bei MES ein erweiterer Bestand von Datenfeldern zur Verfügung, der dem Datenbankangebot des IZ Sozialwissenschaften bei den Hosts ähnelt.

1.4 Eingesetzte Retrievalsoftware

Das Retrievalinstrument FULCRUM, genauer: FULCRUM SearchServer 3.0, ist ein Client-Server-basiertes Volltextdatenbanksystem, das als zentrales Speicherungsprinzip ein relationales Datenmodell verwendet[2]. Aus den Herstellerangaben geht hervor, daß FUL eine Vielzahl von Dokumentformaten unterstützt und umfangreiche Konfigurationsmöglichkeiten bietet. Das System erlaubt eine kombinierte Suche in strukturierten und unstrukturierten Datensätzen. Es soll dem Benutzer eine Reihe mächtiger Retrievalkonzepte und Rankingmethoden zur Verfügung stellen.

In dem vorliegenden Benutzertest wurden den Versuchspersonen (Vpn) in FUL zwei Retrievalmodelle angeboten:

---

[2] Vgl. dazu Jürgen Krause und Peter Mutschke: Indexierung und Fulcrum-Evaluierung, IZ-Arbeitsbericht Nr.17, Bonn, Mai 1999, S.27ff.

- strict Boolean: exact-match-Retrieval mit scharfer Interpretation von Bool`schen Operatoren und

- vector space: statistisches Modell, welches die Ähnlichkeit zwischen Anfrage(vektor) und Dokument(vektor) berechnet.

Die Vpn konnten wählen bzw. ausprobieren, mit welchem der beiden Retrievalmethoden sie arbeiten wollten. Es war auch möglich, beide Methoden zu kombinieren.

Beide Retrievalmodelle können grundsätzlich mit verschiedenen Ranking-Methoden verknüpft werden. Für den Test wurde die Methode critical terms ordered ausgewählt. Das bedeutet, daß die gefundenen Dokumente nach der Häufigkeit der Terme im jeweiligen Dokument in Relation zu der Häufigkeit derselben in der gesamten Kollektion geordnet werden, wobei größeres Gewicht auf diejenigen Terme gelegt wird, die in der Kollektion seltener vorkommen.

Die nach der relativen Häufigkeit der Suchterme wichtigsten Dokumente stehen somit an vorderer Stelle der Ergebnisliste. Die Vpn haben die Möglichkeit, die Ergebnisliste von einer bestimmten Stelle an abzuschneiden, wenn sie den Eindruck haben, daß die weitere Liste keine relevanten Treffer mehr enthält.

Für den Test wurde eine Oberfläche von FUL eingesetzt, die im IZ-Projekt ELVIRA programmiert worden ist. In ELVIRA wird ein prototypisches Marktinformationssystem zur integrierten Online-Recherche nach Fakten in Form von statistischen Zeitreihen und Texten entwickelt. Das Text-Suchwerkzeug entspricht der Oberfläche für den GIRT-Test[3]. Sie bietet sowohl Suchmöglichkeiten in einem Feld ”Freie Suche” als auch in einem Raster, bei dem die Bool`sche Logik verwendet wird.

---

[3] Zu Hintergründen und Inhalten von ELVIRA siehe auch: Krause,Jürgen; Stempfhuber, M.; Mandl, T.: Das Verbandsinformationssystem ELVIRA II, Projektskizze, ELVIRA Arbeitsbericht 12, Informationszentrum Sozialwissenschaften, Bonn, 1997 / Krause, Jürgen; Schaefer, A.: Textrecherche–Oberfläche in ELVIRA II, ELVIRA-Arbeitsbericht 16, Bonn,1998 / Schaefer, André: Benutzertests zur Textrecherche–Oberfläche in ELVIRA II, ELVIRA- Arbeitsbericht 20, Informationszentrum Sozialwissenschaften, Bonn, 1999.

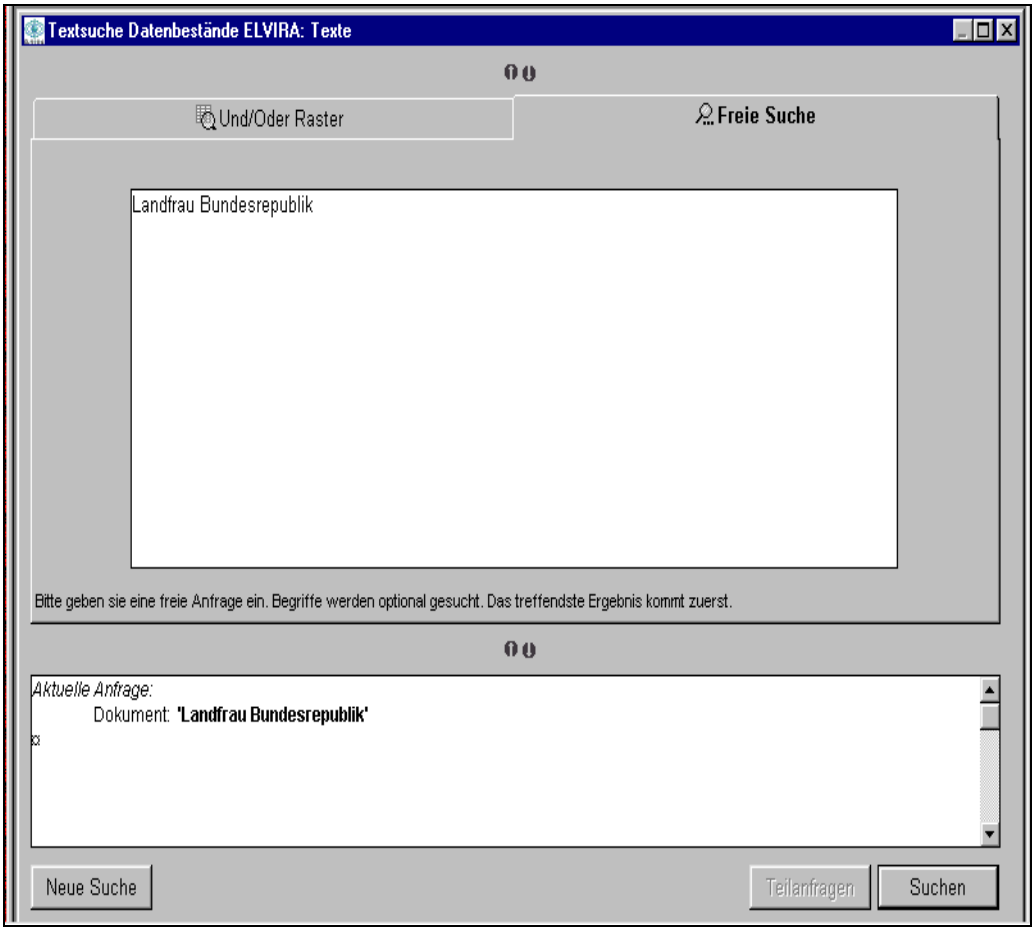

Abbildung 1: Oberfläche FULCRUM Freie Suche

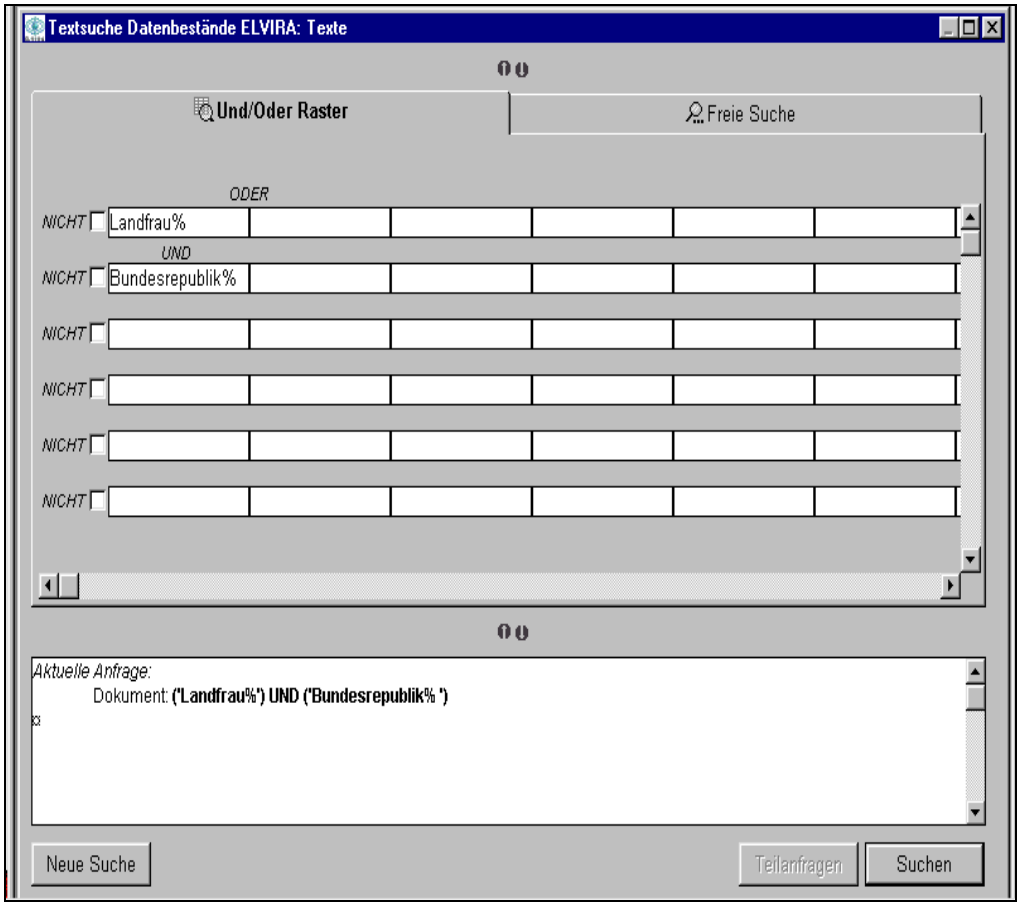

Abbildung 2: Oberfläche FULCRUM Und/Oder Raster

Bei MES wurde eine von STN entwickelte Version unter Verwendung einer maskengesteuerten Oberfläche mit vorgegebenen Suchfeldern (SOLFOR-Abfrage) eingesetzt, mit denen einzeln oder in Kombination (”Globale Suche”) recherchiert werden kann. Die Bool`schen Operatoren können innerhalb und zwischen den Suchfeldern verwendet werden. Bei jedem Feld ist die Einsicht in die jeweiligen Indexe (Basic Index =Freitext inklusive Schlagwörter, CT-Index=inhaltskennzeichnende Schlagwörter etc.) möglich. Diese Schlagwörter sind alphabetisch sortiert und mit der Häufigkeit des Vorkommens in der Datenbank versehen. Sie können per Mausklick in die Suchanfrage übernommen werden. Der Index gibt auch Auskunft über die erforderlichen Schreibvarianten (z.B. ae statt ä) und schützt vor Schreibfehlern bei der Eingabe der Suchbegriffe. Rechts- und Linkstrunkierung sind möglich. Eine Anordnung des Suchergebnisses nach Relevanz ist nicht möglich.

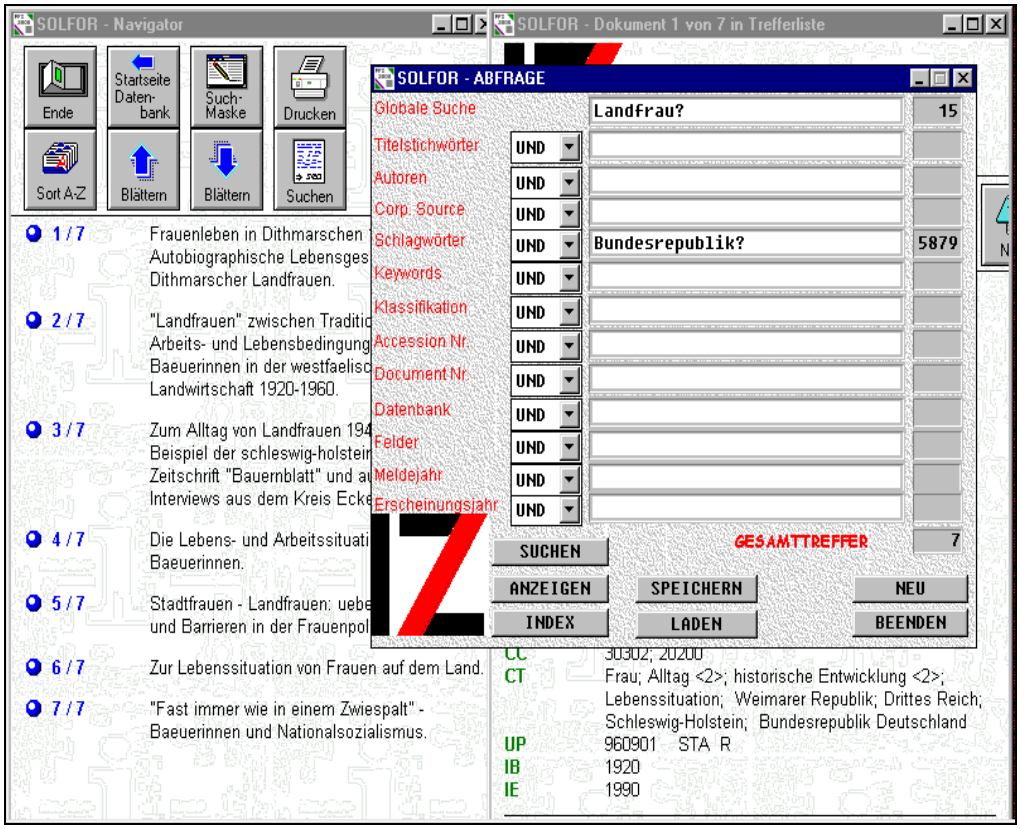

Abbildung 3: Oberfläche MESSENGER

Im Gegensatz zum Pretest wurde nicht mit der Suche im Kommando-Modus gearbeitet. Dadurch war eine Erleichterung für die Versuchspersonen gegeben, zumal sie auch in FUL maskenorientiert recherchieren sollten.

1.5 Auswahl der Recherchethemen

Die Ergebnisse des Pretests zeigen, daß neben den externen Versuchspersonen auch die internen Experten Probleme bei der Bearbeitung der Recherchethemen hatten. Offenbar waren einige der Themen schwieriger, andere leichter zu bearbeiten, unabhängig von freeWAISsf und MES. Da die Ergebnisse eines derartigen Vergleichstests demnach offenbar sehr stark von der Themenformulierung abhängen, schenkte die vorliegende Hauptuntersuchung der Ausarbeitung und der Auswahl der Themen besondere Aufmerksamkeit.

Grundlage für die Themenfindung waren 28 Vorschläge aus dem TREC-Projekt vom August 1998[4], die in einem Test mit der um Zeitungsartikel erweiterten Datenbasis GIRT/NZZ eingesetzt worden waren. Diese 28 Themen, die sich von den Themen des Pretests unterscheiden, sollen die Grundlage für alle weiteren Untersuchungen in diesem Projektbereich sein. Damit soll die Vergleichbarkeit der Ergebnisse verschiedener Tests gewährleistet werden.

Neben dem inhaltlichen sozialwissenschaftlichen Bezug der Themenformulierung (z.B. nicht das Thema "Elektrofahrzeuge") wurden bei der Auswahl der Themen folgende formale Kriterien herangezogen:

- Die Zahl der in der GIRT-Datenbank für ein Recherchethema insgesamt enthaltenen relevanten Dokumente (Anker) soll überschaubar sein, damit die Vpn auch bei Verzicht auf einen cut-off bei der Treffermenge alle gefundenen Treffer in der vorgesehenen Bearbeitungszeit (ca. 3 Stunden für 6 Themen) auf Relevanz prüfen können. Da dafür neben den Überschriften auch die Kurzbeschreibungen (zumindest oberflächlich) durchgesehen werden müssen, wurde eine Obergrenze von maximal 60 bis 70 relevanten Treffern pro Recherchethema angestrebt. Dies entspricht in etwa den Empfehlungen des Pretests[5], die für Recherchen ohne automatisches Ranking eine Obergrenze von 60 Dokumenten vorschlagen. Auf die explizite Formulierung eines cut-off gegenüber den Vpn wurde verzichtet, da dies, entsprechend den Erfahrungen des Pretests, das Rechercheverhalten der Vpn gegenüber einer realen Situation verzerrt (Orientierung der Recherche an der zahlenmäßigen Obergrenze, weniger am inhaltlichen Ergebnis). Bei den Recherchen in FUL wurde zur Begrenzung der Ergebnisliste eine Obergrenze von 150 Treffern voreingestellt, die aber praktisch keine Rolle spielte, da die Vpn von sich aus mit kleineren Treffermengen arbeiten wollten.

- Auf der anderen Seite sollte die GIRT-Datenbank eine Mindestzahl von relevanten Treffern pro Recherchethema enthalten, damit statistisch die Chance besteht, daß sich MES und FUL in der absoluten Zahl der gefundenen relevanten Dokumente unterscheiden können. Angestrebt wurde eine Mindestzahl von 10 bis 15 relevanten Dokumenten.

---

[4] Hinweise zum TREC-Projekt in Frisch/Kluck, a.a.O., S.18

[5] vgl. Frisch/Kluck, a.a.O., S.45

- Die Struktur der Recherchethemen sollte so beschaffen sein, daß die Umsetzung der Themen in eine Rechercheanfrage einerseits nicht zu einfach (nicht nur ein Suchbegriff, der bereits im Titel des Recherchethemas enthalten ist), andererseits auch nicht zu schwierig ist (keine komplexen Formulierungen mit mehreren logischen Operatoren). Dieser möglichst konstant zu haltende "mittlere Schwierigkeitsgrad" der Recherchethemen wurde dadurch definiert, daß jeweils ein Sachgebiet benannt wird, das mit einem Suchbegriff aus der sozialwissenschaftlichen Terminologie beschrieben werden kann (z.B. Kriminalität). Dieses inhaltliche Gebiet wird durch einen weiteren Suchbegriff spezifiziert bzw. eingegrenzt (z.B. Frau). Bei Verwendung der Bool`schen Logik kann durch eine Verknüpfung dieser beiden Suchbegriffe mit "und" das Grundgerüst einer durchaus angemessenen Anfrage formuliert werden. Je nach Resultat können Verfeinerungen (z.B. Ergänzung des Suchbegriffs "Kriminalität" durch "Delinquenz" mit einer Oder-Verknüpfung) vorgenommen werden, die das Rechercheergebnis verbessern. Bei Verwendung des vector-space-Modells in FUL (Freie Suche) können die beiden grundlegenden Suchbegriffe ohne logischen Operator gemeinsam eingegeben werden. Das System müßte dann (so der Anspruch) durch das Rankingverfahren in der Lage sein, Dokumente, die beide Begriffe (evtl. mehrfach) enthalten, als relevante Treffer an vorderer Stelle der Ergebnisliste auszuweisen.

Folgende sechs Themen wurden für die Hauptuntersuchung ausgewählt (in Klammer: Zahl der jeweils relevanten Treffer bzw. des Ankers):

1. Kriminalität bei Frauen (18)
2. Antisemitismus in Deutschland nach 1945 (63)
3. Lean production in Japan (21)
4. Computer im Alltag (11)
5. Gewaltbereitschaft von Jugendlichen (67)
6. Armut und Obdachlosigkeit in Städten (13)

Diese Titel der Recherchethemen wurden ergänzt durch eine Erläuterung der Themenstellung und ggf. eine Spezifizierung, die das inhaltliche Verständnis und die Auswahl der Suchbegriffe erleichtern sollen. Bei dem Thema "Kriminalität bei Frauen" beispielsweise wurden folgende Erläuterungen vorgenommen: "Welche Berichte, Fälle, empirische Untersuchungen und Analysen gibt es zur Kriminalität und Delinquenz bei Frauen? Relevante Dokumente befassen sich mit den speziellen Problemen der Frauenkriminalität einschließlich der Probleme der Resozialisierung und des Strafvollzuges bei

Frauen. Nicht relevant sind historische Untersuchungen (vor 1945), Jugendliche und Kinder (vor allem Mädchen), allgemeine Statistiken, rechtsphilosophische Betrachtungen, Terrorismus" (vgl. Anlage).

Diese Texte standen den Vpn bei der Recherche schriftlich zur Verfügung.

Um systematisch prüfen zu können, ob die Qualität der Recherchen mit der Zahl der relevanten Dokumente pro Thema in der Datenbank zusammenhängt, wurden die sechs Fragen drei Gruppen zugeordnet:

- geringe Anzahl relevanter Dokumente (Computer, Armut)
- mittlere Anzahl relevanter Dokumente (Kriminalität, Lean production)
- hohe Anzahl relevanter Dokumente (Antisemitismus, Gewaltbereitschaft).

1.6 Bestimmung der relevanten Treffer (Anker)

Um die Qualität der Rechercheergebnisse berechnen und zwischen MES und FUL vergleichen zu können, benötigt man pro Frage einen Anker, der die Gesamtmenge der möglichen relevanten Treffer definiert. Aufgrund der Kenntnis dieses Ankers ist es möglich, Aussagen über die Ausschöpfungsquote (Recall) und den Ballast (Precision) einer Recherche zu treffen. Die Gesamtmenge aller möglichen relevanten Treffer einer Recherche ist ein hypothetisches Konstrukt: Sie ist abhängig vom Verständnis der Rechercheфrage und von der Interpretation der von den Retrievalinstrumenten (je nach Queryformulierung) ermittelten Treffer. Die Vorstellung, man könnte durch eine wie auch immer geartete Bestimmungsmethode "alle möglichen relevanten Treffer" für eine Rechercheфrage objektiv korrekt ermitteln, läßt sich nicht realisieren, da es diese Größe nicht gibt. Bei dem Vergleich der Leistungsfähigkeit von zwei Rechercheinstrumenten ist das Postulat eines objektiv "richtigen" Ankers zudem überflüssig. Es genügt, wenn bei der Ermittlung der möglichen relevanten Treffer mit beiden Instrumenten gearbeitet wird und mehrere erfahrene Rechercheure zu einer gemeinsam erzielten Ergebnismenge kommen, bei der die als relevant bezeichneten Treffer in einem inhaltlich plausiblen Zusammenhang zum Suchthema stehen. Bei der Ermittlung des Gesamtergebnisses sollten deshalb Vorkehrungen getroffen werden, daß Ausreißer quantitativ nicht zu Buche schlagen.

Im Pretest wurde die Definition des Ankers von einem einzelnen Fachwissenschaftler durchgeführt, von dessen persönlicher Relevanzbewertung die Ein-

beziehung eines Dokuments in den Anker abhängig war[6]. Diese Vorgehensweise führt dazu, daß Idiosynkrasien unkontrolliert in die Definition des Ankers eingehen können. Um dieses Problem zu vermeiden, wurden die Recherchen zur Ermittlung der Anker von drei erfahrenen RechercheurInnen des IZ durchgeführt. Die Ergebnisse der Recherchen wurden verglichen. Es wurden nur solche Dokumente in die Gesamtmenge der relevanten Dokumente aufgenommen, die in mindestens zwei der drei Expertenrecherchen vorkamen. Dadurch wurde die Menge der relevanten DE in einer überschaubaren Größenordnung gehalten und die Anker einer intersubjektiven Ergebnisvalidierung unterzogen, die der Letztentscheidung durch einen einzelnen Juror vorzuziehen ist[7].

Empirisch zeigt sich, daß die Ergebnisse zumindest von zwei der drei Expertenrecherchen im allgemeinen nicht sehr weit auseinanderliegen: Bei dem Thema ”Kriminalität bei Frauen” beispielsweise wurden vom ersten Rechercheur 15 relevante Dokumente genannt, von denen 10 in der zweiten und 13 in der dritten Recherche übereinstimmend enthalten sind. Zu diesen 13 Dokumenten des Ankers kommen weitere Dokumente aus der zweiten Recherche hinzu, von denen 5 nicht in der ersten, wohl aber in der dritten enthalten sind. Dies ergibt in diesem Fall zusammengenommen einen Anker von 18 Treffern.

1.7 Auswahl und Schulung der Versuchspersonen

Ein zentrales Ergebnis des Pretests besteht darin, daß sich die intellektuelle Leistung der Probanden bei der Formulierung der Problemlösungsstrategie nur schwer von der Leistung der benutzten Software trennen läßt[8] . Deshalb muß der Versuchsaufbau der Vermeidung dieses Interaktionseffekts beider Einflußgrößen besondere Aufmerksamkeit widmen. Die Versuchspersonen sollten durch Schulungsmaßnahmen in die Lage versetzt werden, beide Systeme ”einigermaßen” kompetent zu bedienen und die Anfragen jeweils mit dem einen und dann mit dem anderen System zu bearbeiten. M.a.W.: die Recherchekompetenz sollte soweit wie möglich konstant gehalten werden.

---

6 vgl. Frisch/Kluck, a.a.O.,S.23

7 zur Frage der Zuverlässigkeit und Gültigkeit bei Inhaltsanalysen vgl. bspw. Helmut Kromrey: Empirische Sozialforschung, 7. Auflage, Opladen 1995, S.251ff.

8 vgl. Frisch/Kluck, a.a.O., S.46

Um den Schulungsaufwand in vertretbaren Grenzen zu halten, sollten Personen ausfindig gemacht werden, die über eine grundständige Erfahrung im Umgang mit einer Windows-Oberfläche oder dem Internet oder mit Datenbanken auf CD-ROM mitbringen. Auf der anderen Seite sollten es jedoch keine Recherche-Profis sein, da die Resultate einen Aufschluß über ”durchschnittlich” kompetente Nutzer geben sollen. Durch die Schulung der Versuchspersonen fällt der (im Pretest durchgeführte) wirklichkeitsfremde Einsatz einer Mittlerperson, die die Suchanfragen am PC eingibt, weg.

1.8 Soziodemographische Merkmale der Vpn

Um eine ausreichende Zahl von auswertbaren Recherchen zu erreichen und um jede der im Versuchsplan vorgesehenen Abfolgevarianten der Fragen und Rechercheinstrumente einmal realisieren zu können, wurden 24 Versuchspersonen gewonnen, die als Studenten der Sozialwissenschaft an der Fernuniversität Hagen (vornehmlich Soziologie im Hauptfach) auch ein inhaltliches Verständnis bei der Arbeit mit einer sozialwissenschaftlichen Datenbank mitbringen.

Den Vpn wurde ein Fragebogen vorgelegt, aus dem folgende Daten zur sozialstrukturellen Zusammensetzung der Auswahl und zu den spezifischen Vorkenntnissen hervorgehen: 9 der 24 Personen sind männlich (38%), 15 weiblich (62%). Sie sind zwischen 23 und 50 Jahren alt, bei einem Mittelwert von 38 Jahren. Die überwiegende Zahl ist berufstätig, vor allem in Dienstleistungsberufen. Sie studieren berufsbegleitend die Fächer Soziologie/Sozialwissenschaft (ca. 71% im Hauptfach), Psychologie bzw. soziale Verhaltenswissenschaft (17%), Erziehungswissenschaft (8%) und Literaturwissenschaft (4%). 30% befinden sich im Grundstudium, weitere 20% im 5. und 6. Fachsemester. Die andere Hälfte studiert sozialwissenschaftliche Fächer in höheren Semestern.

Im Zusammenhang mit ihrer Berufstätigkeit und ihrem Studium haben ca.71% der Vpn Rechercheerfahrungen mit Datenbanken sammeln können, 67% haben bereits im Internet recherchiert. 42% haben Erfahrungen mit Datenbankanfragen unter Verwendung der Bool´schen Operatoren.

Es handelt sich damit um eine Personengruppe, die in ihrer Mehrzahl EDV-Anwendererfahrungen mitbringt. Die Vpn waren zudem an den getesteten Rechercheinstrumenten sehr interessiert, da sie im Rahmen ihres Studiums des öfteren Literaturrecherchen (z.B. für Hausarbeiten) durchführen müssen und deshalb einschlägige Schulungen sehr gut gebrauchen können. Als Fernstudenten sind sie zudem ganz besonders auf das Internetangebot der Fernuniversität angewiesen.

Bei der Durchführung der Tests zeigt sich das hohe Maß an Motivation: Keine der 144 Recherchen wurde vorzeitig abgebrochen, die Frageformulierungen und -neuformulierungen sowie die subjektive Relevanzeinschätzung erfolgten mit hoher Konzentration. Im Mittel dauerte eine Sitzung 3 Stunden (incl. Schulung).

1.9 Durchführung des Tests

Der Versuchsplan soll sicherstellen, daß der Vergleich von FUL und MES anhand der bearbeiteten Fragen ohne systematischen Fehler in der Reihenfolge der Bearbeitung erfolgen kann. Jede Vp erhielt alle 6 Fragen zur Bearbeitung, jeweils drei mit FUL und drei mit MES, wobei bei der einen Hälfte mit FUL, bei der anderen mit MES begonnen wurde. Die Reihenfolge der Fragen wurde systematisch variiert, so daß folgender Versuchsplan zustande kam:

- 72 Recherchen in FUL, 72 in MES,
- 24 Recherchen für jedes der 6 Themen, davon 12 in FUL und 12 in MES,
- 4 Recherchen pro Thema an der x-ten Stelle in der Abfolge der Themen pro Vp, davon 2 in FUL und 2 in MES.

VP 1	**VP 2**	**VP 3**	**VP 4**	**VP 5**	**VP 6**	**VP 7**	**VP 8**	**VP 9**	**VP 10**	**VP 11**	**VP 12**	**VP 13**	**VP 14**	**VP 15**	**VP 16**	**VP 17**	**VP 18**	**VP 19**	**VP 20**	**VP 21**	**VP 22**	**VP 23**	**VP 24**
F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
1	1	4	4	1	1	4	4	2	2	5	5	2	2	5	5	3	3	6	6	3	3	6	6
2	2	5	5	3	3	6	6	1	1	4	4	3	3	6	6	2	2	5	5	1	1	4	4
3	3	6	6	2	2	5	5	3	3	6	6	1	1	4	4	1	1	4	4	2	2	5	5
M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F
4	4	1	1	4	4	1	1	5	5	2	2	5	5	2	2	6	6	3	3	6	6	3	3
5	5	2	2	6	6	3	3	4	4	1	1	6	6	3	3	5	5	2	2	4	4	1	1
6	6	3	3	5	5	2	2	6	6	3	3	4	4	1	1	4	4	1	1	5	5	2	2

1= Kriminalität
2= Antisemitismus
3=Lean Production
4= Computer
5= Gewalt
6=Armut
VP=Versuchsperson
F=FULCRUM
M=MESSENGER

Tabelle 1:Versuchsplan

Die Vpn wurden zu Beginn des Tests mit FUL bzw. MES (je nach Einstieg) anhand eines Übungsbeispiels vertraut gemacht. Die Schulung bezog sich auf den technischen Umgang mit dem Rechercheinstrument insgesamt, die Auswahl der Suchbegriffe (incl. Truncierung), die Formulierung einer Suchanfrage (incl. der Bool`schen Logik) sowie die Bearbeitung des Suchergebnisses. Die Aufgabenstellung bestand darin, die Anfragen innerhalb jedes Suchsystems so zu formulieren, daß möglichst alle in der Datenbank vorhandenen relevanten DE und gleichzeitig jedoch möglichst wenige irrelevante DE im Rechercheergebnis angezeigt werden. Die Vpn sollten das Rechercheergebnis nach relevanten und irrelevanten Treffern bewerten. Die Teilnehmer wurden in einem Anschreiben auf die Zielsetzung und den Ablauf des Tests vorbereitet (vgl. Anlage).

Auf eine genaue Zeitvorgabe wurde verzichtet. Bei der Terminabsprache wurde darauf hingewiesen, daß die Vpn mit einer Gesamtdauer von ca. 3 Stunden rechnen könnten. Davon sollten jeweils ca. 20 Minuten auf die Schulung in FUL und MES entfallen. Die tatsächliche Dauer der einzelnen Recherche richtete sich nach dem individuellen Zeitbedarf, der für ein angemessenes Ergebnis samt Bewertung erforderlich war. Die Vpn sollten ihren Zeiteinsatz und ihr Vorgehen an einer realen Recherchesituation ausrichten. Deshalb gab es auch keine Vorgaben zur Höchst- oder Mindestzahl der Treffer.

Mit Einverständnis der Vpn wurde die gesamte Sitzung auf Video aufgezeichnet. Alle Aktionen mit FUL wurden in logfiles festgehalten.

Am Ende der Sitzung wurde den Vpn zudem ein Fragebogen vorgelegt, in dem neben Fragen zur Person (z.B. Alter, Geschlecht, Studienfach, EDV-Vorkenntnisse) die Beurteilung des Testablaufs und die Einschätzung der Rechercheinstrumente erhoben wurden (vgl. Anlage).

1.10 Datenaufbereitung und Datenanalyse

Aus den 144 durchgeführten Recherchen stehen für die Auswertung folgende Informationen zur Verfügung:

- die von den Vpn ausgefüllten Fragebögen,
- Angaben zum Rechercheinstrument und zum Thema der Recherche, zur Position der Recherche in der Versuchsreihe (entsprechend dem Versuchsplan) sowie zur Zeitdauer der Sitzung,
- die Formulierung der endgültigen Suchanfrage,
- die Liste der gefundenen Treffer,

- die subjektive Einschätzung der gefundenen relevanten Treffer durch die Vpn.

Aus den logfiles der Recherchen mit FUL lassen sich zudem die Zahl der Query-Reformulierungen, die Zahl der Anfrageversuche im Freitextfeld bzw. im Raster sowie die Zeitdauer der einzelnen Anfragen ermitteln. Entsprechende Auswertungen liegen vor[9].

Für die statistische Ergebnisanalyse sind die 144 Trefferlisten von besonderer Bedeutung, da sich aus ihnen durch Auszählung bzw. Berechnung die Werte für die folgenden Variablen ermitteln lassen:

- Gesamtzahl der gefundenen Treffer pro Recherche (relevante und irrelevante).

- Gesamtzahl und Dokumentnummern der gefundenen relevanten Treffer pro Recherche. Die gefundenen relevanten Treffer werden durch einen Abgleich der Trefferliste mit dem Anker (Liste der möglichen relevanten Treffer pro Frage) ermittelt.

- Gesamtzahl der von den Vpn als relevant eingestuften Treffer.

Die ersten beiden Kriteriumsvariablen können zusammen mit dem Anker zur Berechnung der Standardmaßzahlen Recall und Precision herangezogen werden.

Mit Recall bezeichnet man den Anteil der in einer Recherche gefundenen relevanten Dokumente, bezogen auf alle für diese Fragestellung relevanten Dokumente in der Datenbank (Anker). Unter Precision versteht man den Anteil der in einer Recherche gefundenen relevanten Dokumente an allen gefundenen Dokumenten dieser Recherche[10].

Recall und Precision können als Indikatoren für die Qualität einer Recherche interpretiert werden: Recall ist ein Ausdruck für die Vollständigkeit bzw. den Abdeckungsgrad eines Rechercheergebnisses, Precision verweist auf den Ballast, den ein Rechercheergebnis aufweisen kann. Beide Maßzahlen können die Werte von 0% bis 100% annehmen.

---

[9] André Schaefer: Benutzertests zur Textretrievalkomponente für ELVIRA II, ELVIRA-Arbeitsbericht 20, IZ-Bonn Juni 1999, S. 20/21

[10] vgl. Frisch/Kluck, a.a.O., S.10-14

Einfache Überlegungen zeigen, daß bei der Interpretation der Qualität einer Recherche Recall und Precision nicht einzeln, sondern kombiniert betrachtet werden sollten: Bei einem numerisch kleinen Anker und einer großen Gesamttreffermenge bei einer Recherche kann leicht ein Recall von 100% herauskommen. Allerdings wird dann dieser ausgezeichnete Recall-Wert durch einen schlechten Precision-Wert relativiert. Umgekehrt kann bei einer sehr niedrigen Gesamttreffermenge der Precision-Wert leicht 100% erreichen (z.B. alle beiden gefundenen Treffer sind relevant), der Recall-Wert ist bei einem numerisch größeren Anker dann jedoch sehr schlecht[11].

2 Retrieval mit vector space gegenüber strict Boolean in FULCRUM

Wie in Kapitel 1 bereits angedeutet, hatten die Vpn bei FUL die Möglichkeit, ihre Recherchen als Freitext-Recherchen mit dem Vektorraummodell in Kombination mit der Rating-Methode „critical terms ordered“ oder als Recherchen mit dem exact-match-retrieval (Bool) durchzuführen. Sie konnten auch mit einem der beiden Verfahren beginnen und dann bei Bedarf auf das andere überwechseln.

Beide Verfahren wurden bei der ca. 20 minütigen Schulung an einem Beispiel (”Berufstätigkeit von Soziologen”) vorgestellt; die Vpn führten selbst unter Anleitung die Proberecherche durch. Anhand des Beispiels konnten sie erkennen, in welcher Weise sich beide Verfahren unterscheiden und was jeweils beachtet werden muß, wenn man ein wie auch immer definiertes zufriedenstellendes Rechercheergebnis erzielen möchte. An Ende der Schulung waren sie in der Lage, mit beiden Verfahren Recherchen zu Themen mit einem ”mittleren Schwierigkeitsgrad”durchzuführen.

**Die 24 Vpn entschieden sich unabhängig voneinander, alle 72 Recherchen in FUL mit dem exact-match-retrieval, d.h. unter Verwendung der Bool`schen Logik, durchzuführen.** Nur in sehr wenigen Fällen wurde das Vektorraummodell überhaupt in Erwägung gezogen, d.h. bei Zwischenschritten einer Testrecherche ausprobiert. Die Entscheidung für das Bool`sche Retrieval war offenbar bereits während der Schulung aufgrund der erzielten Ergebnisse bei der Proberecherche gefallen.

---

[11] zu Fragen der Verknüpfung von Recall und Precision vgl. Christa Womser-Hacker: Der PADOK-Retrievaltest, Hildesheim/Zürich/New York, 1989, S.46ff.

Die Diskussionen mit den Vpn am Ende der Testreihe sowie eigene Erfahrungen mit der ”Freien Suche” in FUL ergaben, daß dieses Rechercheinstrument in der gegenwärtigen Entwicklungsphase noch einen deutlichen Verbesserungsbedarf aufweist, und zwar sowohl bei der Formulierung der Suchanfrage als auch bei der Darstellung und der Ausgabe der Ergebnisliste.

2.1 Suchanfrage mit dem Vektorraummodell

Das Hauptproblem besteht darin, daß das System häufig schon bei der Eingabe von ein oder zwei Suchbegriffen eine sehr große Treffermenge aufweist, die, um die Ergebnisliste überhaupt handhabbar zu machen, in unserem Fall auf 150 Treffer begrenzt wurde. Da die Vpn die Rechercheanfrage so gestalten sollten, daß das Ergebnis möglichst viel relevante Treffer und möglichst wenig Ballast enthalten sollte, würde dies bedeuten, daß die Vpn die 150 Treffer einzeln im Titel und oft auch in der Kurzbeschreibung durchsehen müßten, um entscheiden zu können, ob sie die Anfrage modifizieren oder das Ergebnis akzeptieren und ausdrucken möchten. Da dies u.U. mehrmals nacheinander geschehen muß, bis ein zufriedenstellendes Resultat vorliegt, wäre ein Zeit- und Konzentrationsaufwand erforderlich, der weder in der Test- noch in einer Echtsituation getrieben werden kann.

Dieses Problem würde entscheidend entschärft, wenn das Ranking tatsächlich die relevanten Dokumente an den Anfang der Ergebnisliste setzen würde. In diesem Fall könnte man sich auf die ersten x Dokumente konzentrieren und die Prüfung nach der x-ten Stelle abbrechen. Die Vpn konnten allerdings schnell erkennen, daß die Relevanzbestimmung von FUL nicht ohne weiteres akzeptabel ist und daß viele vorn stehenden Dokumente z.T. schon vom Titel her klar erkennbar irrelevant sind.

Dies soll an einem Beispiel demonstriert werden:

Bei dem Thema ”Computer im Alltag” liegt es nahe, mit dem sinntragenden Suchbegriff ”Computer” zu beginnen, um an dem Ergebnis zu sehen, ob Erweiterungen (z.B. mit ”PC”) sinnvoll sind oder ob man gleich mit der Einschränkung ”Alltag” fortfahren sollte. Die Freie Suche mit ”Computer” ergibt 117 Treffer. Von den ersten 30 Treffern bezieht sich lediglich einer (der 8.) im Titel erkennbar auch auf ”Alltag”. Da der Ballast extrem hoch ist, macht es Sinn, die Recherche einzuschränken und als zweiten Suchbegriff ”Alltag” einzugeben. Da das Vektorraummodell keine logischen Verknüpfungen von Suchbegriffen kennt, wird der Umfang der Suche durch diesen zweiten Begriff jedoch tatsächlich erweitert (beide Begriffe werden getrennt untersucht) und es entsteht ein Suchergebnis von 272 Treffern. Die ersten 30 dieser Treffer enthalten wohl Titel zum Thema ”Computer”, jedoch nur in 4 Fällen zu ”Computer im Alltag” (Treffer 7, 8, 9 und 16, überprüft an dem Anker zu dieser Frage). Um feststellen

zu können, ob die Recherche weitere relevante Treffer enthält, müßte man auch die nächsten Treffer der Ergebnisliste analysieren (evtl. alle 272). Schließlich könnte man ausprobieren, ob eine neue Recherche mit einem weiterer Suchbegriff (z.B. ”Freizeit”) mehr relevante Treffer erbringen würde. Allerdings müßte man wegen der abermals verlängerten Ergebnisliste mit einem noch größeren Prüfungsaufwand rechnen. Dies war den Vpn schon aus Zeitgründen nicht zuzumuten.

Im übrigen ist die Qualität dieser Recherche mit 272 Treffern relativ schlecht: Geht man von den ersten 30 Dokumenten aus, so ist der Recall (4 von 11) = 36%, die Precision (4 von 30) = 13%. Ein Vergleich mit den durchschnittlichen Recall- und Precisionwerten der 72 FUL-Recherchen, die in dem Test MES-FUL durchgeführt worden sind, zeigt, daß der Recall leicht unter dem Durchschnitt von 41%, die Precision jedoch sehr stark unter dem Durchschnitt von 57% liegt (Tabelle 2: Hauptergebnisse im Überblick).

Die Alternative innerhalb von FUL liegt beim Wechsel zur Bool`schen Logik, d.h. zur Suche im Raster mit der Möglichkeit, die Suche durch ”oder” zu erweitern bzw. durch ”und” oder ”nicht” eine große Ergebnismenge handhabbar einzuschränken. Im vorliegenden Beispielsfall erbringt die Einschränkung durch die schlichte Suchformulierung ”Computer% und Alltag%”[12] 18 Treffer, 8 davon sind relevant. Die Qualität dieser einfachen Anfrage ist erheblich besser als beim Vektorraummodell: Recall (8 von 11) = 73%, Precision (8 von 18) = 44%.

Das Ranking gibt auch bei der Suche mit Bool keine besonders zuverlässigen Hinweise auf relevante Treffer: Die 8 relevanten Treffer verteilen sich beinahe gleichmäßig auf die 18 Positionen der Ergebnisliste (Position 1, 2, 5, 6, 11, 13, 17 und 18), d.h. neben den beiden ersten und einigen mittleren sind auch die beiden letzten Treffer der Liste relevant. Hätte man die Ergebnisliste z.B. bei der Hälfte, d.h. bei der 9. Stelle, abgeschnitten, hätte man nur die Hälfte der relevanten Treffer identifiziert.

Die Crux bei der Arbeit mit dem Vektorraummodell ist die Reduktion von großen Ergebnismengen, die bereits mit ein oder zwei Suchbegriffen erzeugt werden. Das gegenwärtig eingesetzte Ranking liefert unzuverlässige Ergebnisse, d.h. die höchstgerankten Treffer sind nur zum Teil oder gar nicht relevant. Das Ranking ersetzt nicht die Durchsicht langer Ergebnislisten. Die Struktur der in dem Test MES -FUL eingesetzten Fragen (vgl. Kapitel 1) und der Inhalt der Datenbasis GIRT (jeweils große Trefferzahl mit den inhaltskennzeichnenden Suchbegriffen in den ausgewählten Sachgebieten) erfordern die Kombination von mindestens zwei Suchbegriffen, die in Form einer Einschränkung aufeinan-

[12] Das Zeichen „%“ bedeutet Truncierung.

der bezogen sein müssen (z.B. ”Computer” durch die Spezifikation ”Alltag”). Bool leistet dies durch die Verknüpfung mit ”und” (und ggf. ”nicht”). Ein funktionierendes Äquivalent ist beim Vektorraummodell gegenwärtig nicht erkennbar.

Von uns durchgeführte Vergleichstests zeigen, daß das Vektorraummodell bei Recherchethemen, die mit einem einzigen Suchbegriff oder zwei bis drei Synonymen formuliert werden können, die jeweils nur selten (z.B. maximal bei 30 Dokumenten) in der Datenbank vorkommen oder die so spezifisch sind, daß automatisch praktisch alle Treffer relevant sind (etwa bei der Suche nach Dokumenten zu einem Straftatbestand wie Korruption), durchaus akzeptable Ergebnisse erzielen kann. In diesen Fällen rückt die Ranking-Problematik in den Hintergrund.

Neben dieser Grundproblematik existieren bei der gegenwärtigen Installation der Freien Suche in FUL weitere Unzulänglichkeiten, die bei den Vorarbeiten für die Benutzertests zum Vorschein kamen. Die Vpn selbst konnten auf diese Probleme nicht stoßen, da sie bei den Tests die Freie Suche praktisch nicht eingesetzt haben.

1. Das System erkennt keine Suchbegriffe, die aus mehreren Einzelwörtern bestehen. Statt dessen wird die Anfrage so abgearbeitet, als ob es sich um mehrere Suchbegriffe handelte, deren Häufigkeiten getrennt ermittelt werden müßten. Bei diesem Vorgehen kann es durchaus zu unsinnigen Ergebnissen kommen.
   Beispiel: Eine Anfrage aus dem Forschungsgebiet zur sozialen Ungleichheit mit dem Thema ”Soziale Schicht”, bei der ”soziale Schicht” als Suchbegriff eingegeben wird, erbringt mit der Freien Suche 1.247 Treffer. Die im Ranking an vorderer Stelle stehenden 5 Treffer beziehen sich auf 12-Stunden-Schicht, Nachtschicht, 3-Schicht-Betrieb etc. Auch die weiteren Treffer enthalten zum erheblichen Teil ”Schicht” im Sinne von Arbeitszeit, nicht sozialer Schicht. Das System ist nicht in der Lage, den sozialwissenschaftlichen Grundbegriff ”soziale Schicht” als Mehrwortbegriff angemessen zu recherchieren.
   In der Anzeige wird das Suchwort allerdings mit ‘soziale Schicht’ angezeigt, so daß der Nutzer den Eindruck gewinnt, die beiden Hochkommata wiesen auf einen einzelnen, aus zwei Wörtern bestehenden Suchbegriff hin.

2. Die Freie Suche erfaßt den gesamten in GIRT vorliegenden Dokumenttext. Es besteht keine Möglichkeit, einzelne Textarten (z.B. die Namen von Autoren) aus der Recherche auszuschließen. Auch dies kann zu sehr unerwünschten Resultaten führen.

Beispiel: Eine stadtsoziologische Recherche verwendet den Suchbegriff ”urban”. Gefunden werden 13 Treffer. Zwei davon beziehen sich auf Urbanität, die anderen 11 werden aufgenommen, weil die Autoren Dieter bzw. Franz Urban heißen. Die beiden schreiben allerdings nichts zum Thema Urbanität.

3. Eingegebene Suchbegriffe mit Sonderzeichen werden in unterschiedlicher, teilweise überraschender Weise behandelt.
   Ein Bindestrich führt dazu, daß der Suchbegriff insgesamt ignoriert wird. Die Suche mit ”lean-production” beispielsweise führt zu der Auskunft: Gefundene Treffer Null. Es erfolgt keine Fehlermeldung, so daß man davon ausgehen muß, die Datenbank enthalte keine einschlägigen Dokumente. Eine Kontrolluntersuchung mit ”lean production” zeigt jedoch, daß dem nicht so ist. Allerdings wird ”lean” und ”production” (wie soziale Schicht) nicht als Mehrwortbegriff erkannt.
   Die Suche nach ”medizinisch-technisches Personal” führt wohl zu Treffern. Es handelt sich allerdings um dieselben, die eine einfache Suche nach ”Personal” ebenfalls ergeben hätte. Hier führt die Verwendung des Bindestrichs zum Ausschluß eines Teils eines Mehrwort-Suchbegriffs.
   Die Suche nach der Stadt Frankfurt/Oder wird so vorgenommen, daß diesmal das Zeichen ”/” und nicht der gesamte Suchbegriff ignoriert wird. Die Folge ist, daß die Ergebnisliste eben auch Dokumente zu Frankfurt/Main enthält, die durch den Zusatz ”Oder” ausgeschlossen werden sollten. Zudem wird der Fluß Oder mit dem logischen Partikel ”oder” gleichgesetzt, so daß die Ergebnismenge 2.158 Treffer umfaßt.

4. Die Einbeziehung nicht-sinntragender Wörter wie ”oder” bzw. ”von” führt zu weiteren unerwünschten Ergebnissen. Bei der Suche nach dem Autor mit dem Namen ”von Alemann” erhält man immerhin 9.978 Treffer, das sind rund 75% der gesamten Datenbank. Bei der Suche nach den Ursachen dieser exorbitanten wissenschaftlichen Produktivität dieses Sozialwissenschaftlers stößt man auf die Tatsache, daß die Freie Suche in FUL den Namenszusatz ”von” als eigenständiges Suchwort interpretiert und alle Dokumente nachweist, in denen dieses Wort vorkommt.

5. Es besteht keine Möglichkeit, Suchbegriffe zu truncieren. Dies führt bei Recherchen, die mit umfangreichem Vokabular zu einem bestimmten Suchgebiet arbeiten müssen (z.B. ”Arbeits...”), dazu, daß sehr viele Begriffe mit demselben Wortstamm getrennt eingegeben werden müssen. Diese Arbeit ist zeitraubend und sollte überflüssig sein. Es wurde getestet, ob bei der Freien Suche Wortvarianten (z.B. Pluralformen) automatisch berücksichtigt werden, so daß die fehlende Truncierungsmöglichkeit wenigstens teilweise kompensiert würde:

Eine Recherche mit dem Suchbegriff ”Soziologe” erbringt mit der Freien Suche 7 Treffer. Eine Recherche mit dem Plural ”Soziologen” erbringt 33 Treffer. Die beiden Treffermengen überschneiden sich nicht.
Eine weitere Recherche mit dem Suchbegriff ”Soziologin” erbringt 1 Treffer, eine andere mit dem Plural ”Soziologinnen” 7 Treffer. Wiederum keine Überschneidung.
Weitere Tests mit den Begriffen ”Bauer-n” und ”Antisemit-en” bestätigten, daß die automatische Truncierung bei Pluralbildungen nicht funktioniert. Bei einem System, daß die Freitextsuche verwendet, ist die Truncierungfunktion unverzichtbar.

6. Die Freie Suche unterstützt den Nutzer nicht bei der Auswahl, Formulierung und Eingabe der Suchbegriffe. Zumindest bis zur Testdurchführung lag kein Index vor. Schreibfehler bei der Eingabe führen zu der mißverständlichen Meldung: Gefundene Dokumente Null. Schreibvarianten werden nicht angeglichen, z.B. Photographie: 1 Treffer, Fotographie: Null. Wird nach ”lean production” gesucht, werden Dokumente mit dem Wort ”Produktion” nicht einbezogen.

2.2 Darstellung und Ausgabe der Ergebnisliste

Die Ergebnisausgabe weist bei der gegenwärtigen Installation von FUL noch etliche Unzulänglichkeiten auf:

1. Die Ergebnisliste, auch wenn sie 150 Treffer enthält, kann als Liste nur komplett ausgedruckt werden. Eine Markierung und Auswahl der relevanten Dokumente ist nicht möglich.
   Beim Ausdruck der Treffer im Volltext ist eine Auswahl oder Unterdrückung einzelner Textelemente (z.B. des Kurzreferats) nicht möglich. Beides führt zu unnötigen Druckzeiten und -kosten.
2. Werden bei einer Suche große Treffermengen erzielt, wird dies zunächst als Summenergebnis angezeigt (z.B. ”2.387 Treffer gefunden”). Ist für einen Nutzer diese Menge zu groß, besteht danach keine Möglichkeit mehr, den Anzeigevorgang abzubrechen. Es erscheint ein Bildschirm mit einer Liste der ersten Treffer. Die Anzeige der restlichen wird weiter vorbereitet und nimmt Zeit in Anspruch. Erst danach ist ein Wechsel in den Hauptbildschirm bzw. eine neue Suche möglich. Dies ist ein für den Nutzer u.U. sinnloser Zeitaufwand.

Ein Teil dieser Unzulänglichkeiten bei der Suche und der Ergebnisausgabe wurde bei der Entwicklung der FUL-Retrievaloberfläche, die mit der Bool`schen Logik und dem Suchraster arbeitet, behoben: Das Raster enthält Eingabefelder, die auch Mehrwortbegriffe als spezifische Suchbegriffe zulassen (incl. Mehrwort-Autoren-namen). Die Eingabe von Binde- und Schrägstrichen wird in der

vom Nutzer vorgesehenen Weise akzeptiert; Links- und Rechtstruncierung ist möglich. Allerdings wird auch in dieser Version von FUL der gesamte Dokumenttext durchsucht, so daß das genannte Problem bei ”urban” wieder auftritt. Die Suche in einzelnen Textsegmenten bzw. der Ausschluß einzelner Textteile bei der Suche ist nicht vorgesehen.

Zum Zeitpunkt des Tests wurden die Nutzer auch bei dieser Version weder durch einen Index oder Thesaurus, noch durch Rechtschreibhilfen noch durch eine Hilfefunktion unterstützt. Die genannten Desiderate bei der Darstellung und Ausgabe der Ergebnisliste harren ebenfalls noch der Realisierung.

3 Die Hauptergebnisse im Überblick

Nach den Ausführungen zu den Spezifika von FUL soll nun in diesem Kapitel die Frage beantwortet werden, welche Ergebnisse dieses FUL-Retrievalsystem (Und/Oder Raster) im Vergleich zu MES bei einem Testeinsatz geliefert hat. Die Ergebnisse der Untersuchung werden im Hinblick auf die gefundenen relevanten Treffer sowie auf Recall und Precision dargestellt. Die Resultate werden in erster Linie nach den Faktoren Rechercheinstrument (MES vs. FUL) und Recherchethema differenziert. Darüber hinaus wird der Einfluß von Merkmalen der Vpn (z.B. Alter, Geschlecht, Rechercheerfahrung) überprüft. Bei den hauptsächlich verwendeten statistischen Verfahren handelt es sich um einfache Varianzanalysen (ANOVA) und Kovarianzanalysen. Darüber hinaus wurde eine Regressionsanalyse durchgeführt.

Die von den Vpn formulierten Suchanfragen werden beispielhaft qualitativ analysiert. Die Auswertung der subjektiv relevanten Treffer sowie der Bewertungsfragen aus dem Fragebogen erfolgt in einem letzten Analyseabschnitt.

Tabelle2 informiert über die Hauptergebnisse der Studie: Insgesamt werden in den 144 Recherchen 2.206 relevante Treffer aufgefunden. Das sind 47,6% aller relevanten Treffer, die maximal hätten aufgefunden werden können. Bezogen auf die insgesamt gefundenen 4.132 Treffer (relevante und irrelevante) sind dies 53,3%.

Mit anderen Worten: Knapp die Hälfte aller möglichen Treffer ist gefunden worden und ungefähr jeder zweite Treffer ist relevant.

Die 72 MES-Recherchen enthalten 1.299 relevante Treffer, die 72 FUL-Recherchen nur 907 Treffer. Dies ist ein Unterschied von 392 Treffern. **Bezogen auf die Zahl von 907 relevanten Treffern bei FUL weist MES damit einen Vorsprung von 43,2% auf.**

	Messenger	Fulcrum	insgesamt	Vorsprung Mes		Mittelwerte			Eta	Sig.
	N=72	N=72	N=144	N	% von Ful	insges.	Mes	Ful		
maximale. Zahl relevanter Treffer (Anker)	2.316	2.316	4.632	-	-	32,2	32,2	32,2	-	-
Summe und Mittelwert aller Treffer	2.214	1.918	4.132	296	15,4	28,7	30,8	26,7	.07	-
Summe und Mittelwert der gefundenen relevanten Treffer	1.299	907	2.206	392	43,2	15,3	18,0	12,6	.17	**
Recall 1)	-	-	-	-	-	.49	.56	.41	.30	**
Precision 1)	-	-	-	-	-	.58	.60	.57	.07	-
Recall 2)				-	-	.48	.56	.39		
Precision 2)				-	-	.53	.59	.47		

1) Berechnung nach **Makro-Methode**
2) Berechnung nach **Mikro-Methode**
) Signifikanz: **5%-Niveau

Tabelle 2: Hauptergebnisse im Überblick

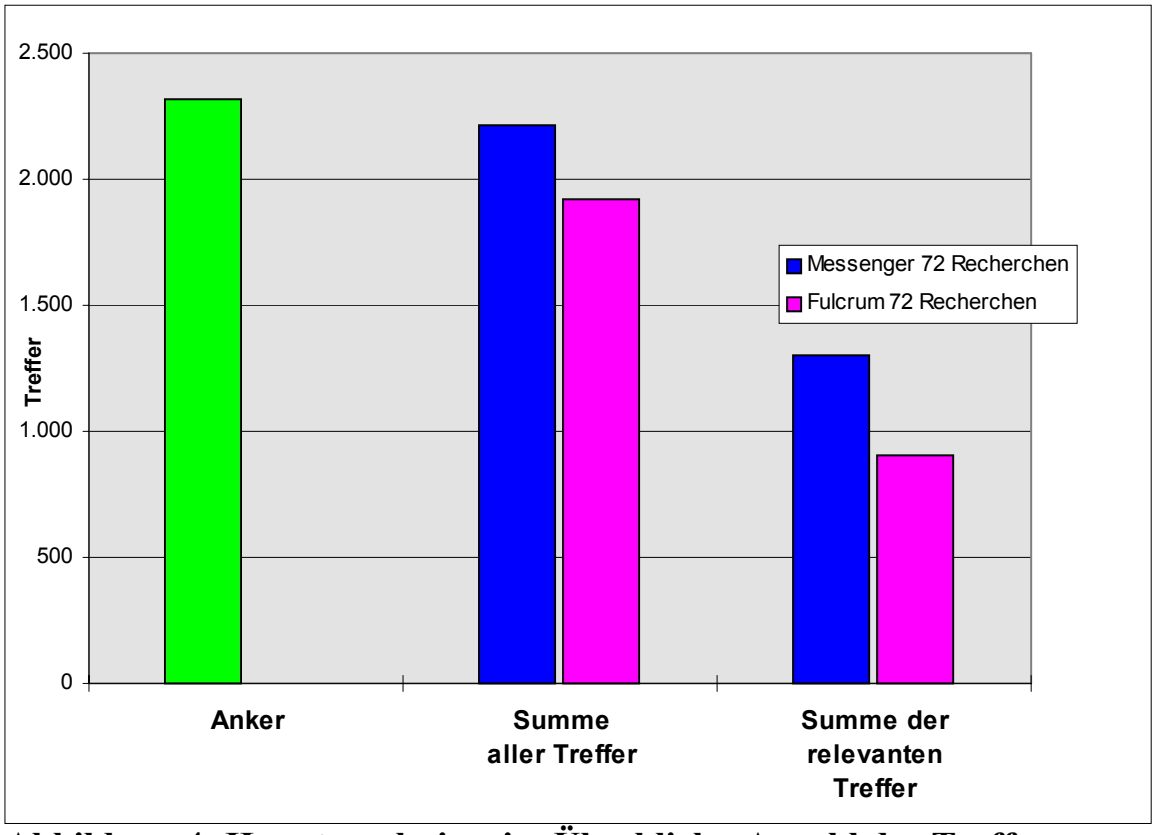

Abbildung 4: Hauptergebnisse im Überblick - Anzahl der Treffer

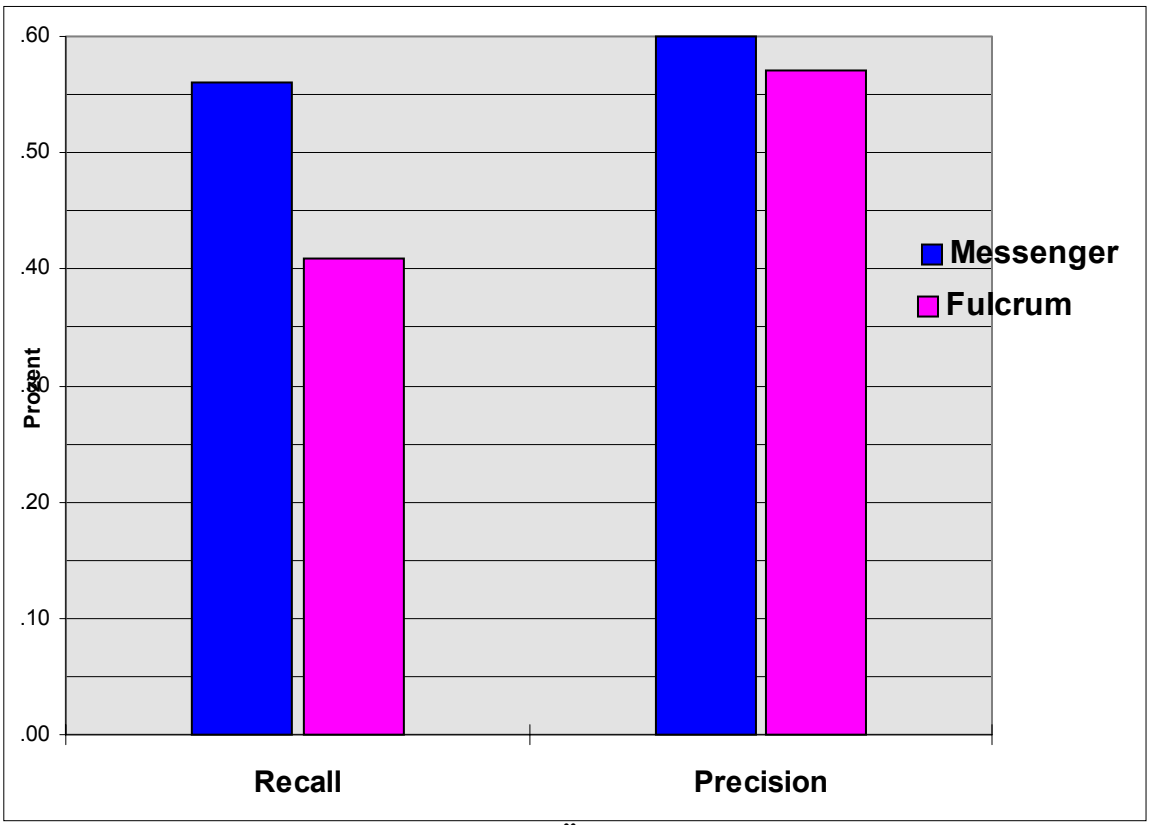

Abbildung 5: Hauptergebnisse im Überblick - Recall und Precision

Im Durchschnitt werden in einer Recherche 15,3 relevante Treffer identifiziert. Mit MES werden durchschnittlich 18 relevante Treffer gefunden, mit FUL 12,6. Dies ist ein Vorsprung von MES gegenüber FUL von 5,4 relevanten Treffern pro Recherche. Der Mittelwertsunterschied zwischen MES und FUL ist signifikant; der Eta-Wert beträgt .17[13]

Dieser Vorsprung von MES gegenüber FUL gilt auch für den Recall: Die Werte von MES liegen durchschnittlich bei 56%, von FUL bei 41% (Makro-Methode). Dieser Unterschied ist signifikant; der Eta-Wert beträgt .30. Der Vorsprung von MES liegt bei 15 Prozentpunkten. Bezogen auf den Wert von FUL mit 41% bedeutet dies einen Vorsprung im Recall von 36,6%.

Für die Precision kann ein signifikanter Unterschied zwischen MES und FUL (mit der Makromethode) nicht nachgewiesen werden.

Die genannten Werte für Recall und Precision beruhen auf Mittelwertsberechnungen nach der Makro-Methode. Bei dieser Berechnungsweise werden die Kennwerte zunächst für jede Recherche getrennt berechnet, addiert und dann durch die Gesamtzahl der Recherchen (hier: 144) dividiert. Bei der Berechnung von Precision entsteht dabei insofern eine Gewichtung der einzelnen relevanten Treffer, als ihr Beitrag zum jeweiligen Precision-Wert einer Recherche von der Gesamttreffermenge bei dieser Recherche abhängt.

Beispiel: 4 relevante Treffer bei 10 Gesamttreffern einer Recherche ergibt p=.40; 4 relevante Treffer bei 20 Gesamttreffern einer Recherche: p=.20.

Das bedeutet: Bei Recherchen mit niedriger Gesamttrefferzahl führt eine bestimmte Zahl von relevanten Treffern zu einem höheren Precision-Wert als bei Recherchen mit höheren Gesamttrefferzahl. Wenn es in einer Teilgruppe der untersuchten 144 Recherchen überdurchschnittlich viele Recherchen mit niedriger Gesamttrefferzahl gibt, führt dies im Mittel zu relativ höheren Precision-Werten als bei der Vergleichsgruppe. Genau dies ist bei den FUL-Recherchen der Fall: Die Untersuchung der Gesamttreffermengen aller 144 Recherchen zeigt, daß der Anteil der FUL-Recherchen an allen Recherchen mit maximal 8 Treffern 67% beträgt, bei allen Recherchen mit maximal 10 Treffern sind es noch 64%. Die Betrachtung der Mittelwertsunterschiede bei der Gesamttreffermenge führt zu demselben Resultat: Die FUL-Recherchen haben im Mittel 26,7

[13] Der Korrelationskoeffizient Eta kann Werte zwischen .00 und 1.00 annehmen. Je höher der Wert, desto größer ist der Unterschied zwischen MES und FUL in bezug auf die betrachtete Variable.

Treffer, die MES-Recherchen 30,8. Durch diesen überproportionalen Anteil an "kleinen" Recherchen bei FUL werden die Precision-Werte von FUL relativ erhöht. Die verbleibenden Unterschiede zwischen MES und FUL sind statistisch nicht signifikant.

Verwendet man die Mikro-Methode, erhält man deutliche Unterschiede zwischen MES und FUL. Bei der Mikro-Methode werden bei der Berechnung von Precision jeweils alle erzielten Treffer und alle erzielten relevanten Treffer über alle 144 Recherchen addiert und dann der Quotient gebildet. Die Berechnung eines Mittelwerts von Quotienten ist nicht erforderlich. Auf diese Weise gehen alle gefundenen relevanten Treffer mit gleichem Gewicht in die Berechnung ein. Im vorliegenden Fall sind dies bei allen 144 Recherchen insgesamt 2.206 gefundene relevante Treffer bezogen auf 4132 Gesamttreffer, das sind 53%. Bei MES sind es 1.299 von 2.214 Treffern, das sind 59%. Bei FUL sind es 907 von 1.918 Treffern, das sind 47%. Damit ergibt sich nach der Mikro-Methode auch für Precision ein erheblicher Vorsprung von MES gegenüber FUL von 12 Prozentpunkten. Bezogen auf den Wert von FUL mit 47% ist dies ein Vorsprung von MES von 26%.

Die Anlage der Untersuchung mit konstant gehaltenen Recherchemengen für MES und FUL bzw. für die 6 Recherchethemen verlangt eine statistische Auswertung auf der Basis der Recherchen und nicht der einzelnen Treffer, da nur auf diese Weise Gewichtungsprobleme bei dem Vergleich von MES und FUL, die aus den unterschiedlichen Trefferzahlen bei beiden Systemen entstehen würden, vermieden werden können. Das bedeutet, daß bei den Datenanalysen, über die im folgenden berichtet wird, die Werte für die Kriteriumsvariablen Recall und Precision jeweils mit der Makromethode ermittelt wurden[14].

Insgesamt zeigt sich, daß MES bei den Variablen Recall und gefundene relevante Treffer deutliche und statistisch signifikante Vorsprünge gegenüber FUL aufweist. Bei der Berechnung mit der Makro-Methode zeigen sich bei Precision keine signifikanten Unterschiede. Bei Anwendung der Mikro-Methode ergeben sich allerdings auch bei dieser Variablen deutliche Vorteile zugunsten von MES.

---

[14] Zur Mittelwert-Berechnung nach der Makro- und der Mikromethode vgl. Womser-Hacker, a.a.O., S.67f. Sie zitiert zustimmend eine Arbeit von Roccio, der der Makromethode den Vorzug gibt, „da diese Vorgehensweise dem Benutzerstandpunkt entspricht" (S.68).

4 Nach Recherchethemen differenzierte Ergebnisse

Die 2.206 gefundenen relevanten Treffer verteilen sich höchst ungleich auf die sechs Recherchethemen. Tabelle 3 zeigt, daß das Thema ”Jugend und Gewalt” mit durchschnittlich knapp 31 relevanten Treffern am ergiebigsten gewesen ist, direkt gefolgt von dem Thema ”Antisemitismus in Deutschland nach 1945” mit 29,5 relevanten Treffern. Die beiden Themen ”Lean Production in Japan” und ”Kriminalität von Frauen” liegen mit jeweils ca. 10 relevanten Treffern im Mittelfeld. Bei den Themen ”Computer im Alltag” wurden im Schnitt ca. 6 relevante Treffer, bei dem Thema ”Armut und Obdachlosigkeit in Städten” ca. 5 relevante Treffer gefunden. Der Vergleich der Reihenfolge der sechs Themen bei den gefundenen relevanten Treffern und den Ankern zeigt, daß die beiden Themen mit den höchsten Ankern auch die höchste Zahl an gefundenen relevanten Treffern aufweisen. Dasselbe gilt für die beiden Fragen im Mittelfeld. Die beiden Fragen mit den geringsten Ankerwerten liegen auch bei den gefundenen relevanten Treffern am Ende der Liste, allerdings rutscht das Thema ”Armut” vom 5. auf den letzten Platz ab.

Thema	**Relevante Treffer**					
	Mittelwerte			Differenz **		
	Mes	Ful	insges.	Mes $\overline{x}$ - Ful $\overline{x}$	Eta	Sig.
1. Kriminalität	11,8	8,3	10,0	+ 3,5 (+41)	.54	*
2. Antisemitismus	41,2	17,8	29,5	+ 23,4 (+281)	.56	*
3. Lean Production	11,8	8,7	10,3	+ 3,1(+38)	.37	-
4. Computer	7,5	4,8	6,2	+ 2,7 (+32)	.52	*
5. Gewalt	31,4	30,3	30,9	+ 1,1 (+13)	.03	-
6. Armut	4,6	5,7	5,1	- 1,1 (-13)	.20	-
Summe	18,1	12,6	15,3	+ 5,5 (+392)		

* Signifanz: 5% - Niveau
** Die erste Zahl in dieser Spalte ist die Differenz der Mittelwerte der relevanten Treffer in Prozentpunkten, die zweite Zahl in Klammern ist die Differenz der relevanten Treffer in absoluten Zahlen.

Tabelle 3: Relevante Treffer nach Recherchethemen

Thema	Recall					
	Mittelwerte			Differenz		
	Mes	Ful	insges.	Mes $\bar{x}$ - Ful $\bar{x}$	Eta	Sig.
1. Kriminalität	.65	.46	.56	+ .19	.54	*
2. Antisemitismus	.65	.28	.47	+ .37	.56	*
3. Lean Production	.56	.41	.49	+ .15	.37	-
4. Computer	.68	.44	.56	+ .24	.52	*
5. Gewalt	.47	.45	.46	+ .02	.03	-
6. Armut	.35	.44	.39	- .09	.20	-
Summe	.56	.41	.49	+ .15		

*Signifanz: 5% - Niveau

Tabelle 4: Recall nach Recherchethemen

Thema	Precision					
	Mittelwerte			Differenz		
	Mes	Ful	insges.	Mes $\bar{x}$ - Ful $\bar{x}$	Eta	Sig.
1. Kriminalität	.57	.54	.56	+ .03	.09	-
2. Antisemitismus	.53	.48	.50	+ .05	.09	-
3. Lean Production	.71	.70	.71	+ .01	.02	-
4. Computer	.59	.43	.51	+ .16	.31	-
5. Gewalt	.84	.76	.80	+ .12	.21	-
6. Armut	.39	.48	.44	- .09	.16	-
Summe	.61	.57	.59	+ .04		

*Signifanz: 5% - Niveau

Tabelle 5: Precision nach Recherchethemen

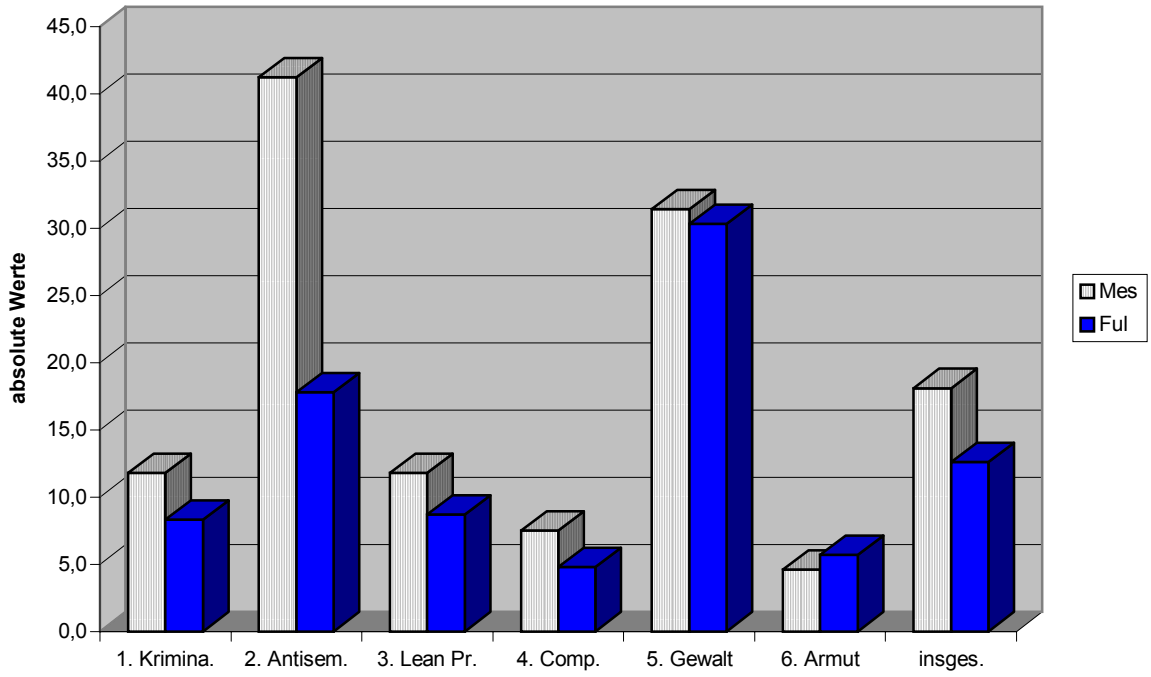

Abbildung 6: Relevante Treffer nach Recherchethemen

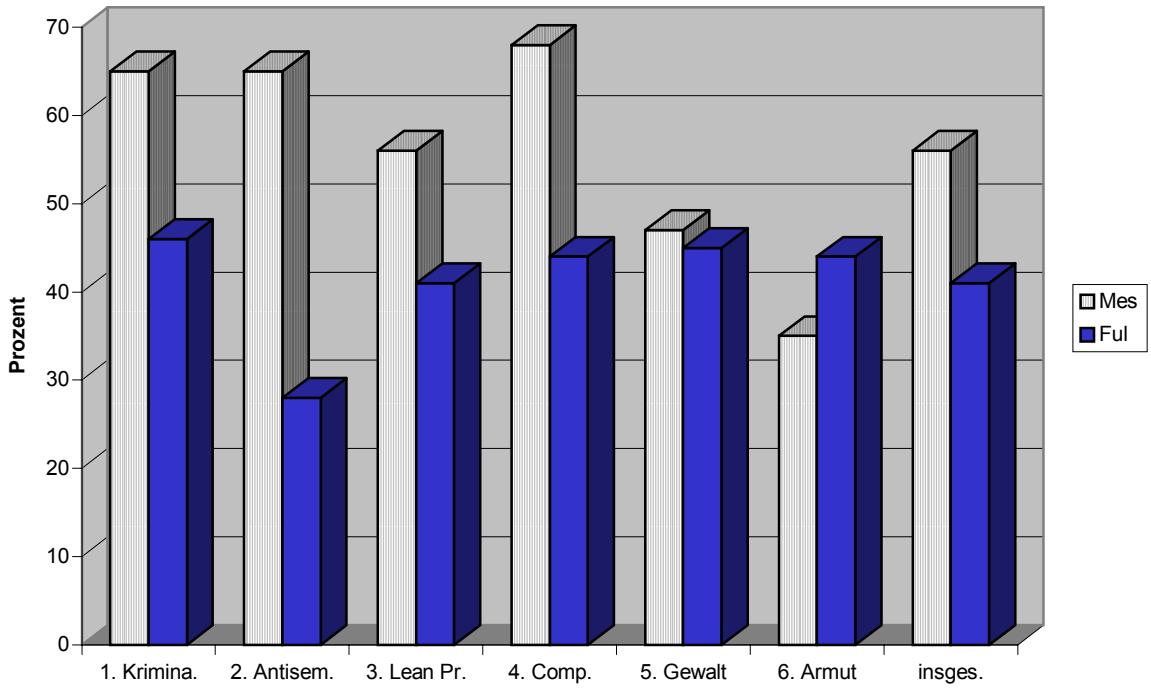

Abbildung 7: Recall nach Recherchethemen

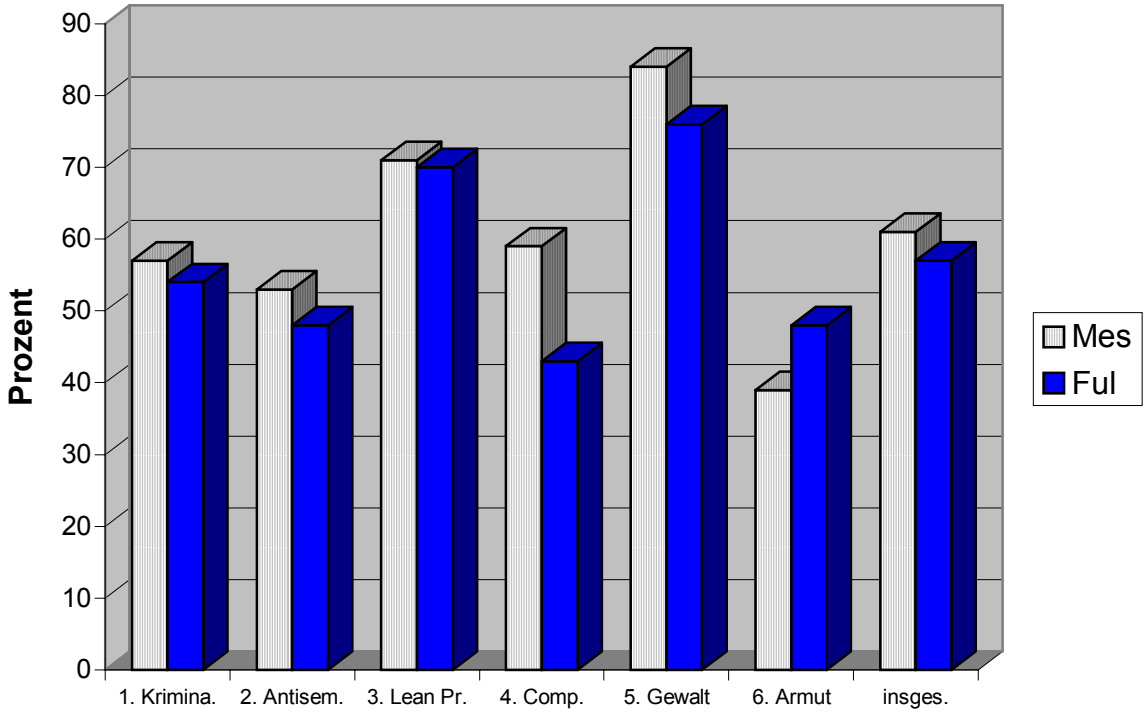

Abbildung 8: Precision nach Recherchethemen

Berechnet man die Anteile der absoluten Trefferzahlen pro Recherchethema an der Gesamtzahl der gefundenen relevanten Treffer, so entfallen auf die beiden (nach den Ankerzahlen) ”großen” Themen ”Gewalt” und ”Antisemitismus” knapp 34% bzw. 32%, d.h. diese beiden Themen enthalten zusammen zwei Drittel aller gefundenen relevanten Treffer. Auf ”Kriminalität” und ”Lean Production” entfallen jeweils ca. 11%, während die beiden ”kleinen” Themen ”Computer” und ”Armut” mit jeweils ca. 6% beteiligt sind.

Für die Fragestellung der Studie besonders relevant ist die Betrachtung der Unterschiede zwischen MES und FUL hinsichtlich der gefundenen relevanten Treffer. Im Hinblick auf die Reihenfolge der Themen nach dem Mittelwert bzw. der Summe der Treffer ergeben sich keine besonders bemerkenswerten Differenzen: Sowohl bei MES wie auch bei FUL stehen die ”großen” Themen am Anfang (allerdings mit getauschter Reihenfolge der Plätze 1 und 2), die ”mittleren” in der Mitte und die ”kleinen” am Ende (ebenfalls mit getauschter Reihenfolge). Vergleicht man jedoch die erzielten Durchschnittswerte der gefundenen relevanten Treffer im einzelnen, ergeben sich bemerkenswerte Differenzen: Bei dem Thema ”Antisemitismus” kommen die MES-Recherchen auf durchschnittlich ca. 41 Treffer, die FUL-Recherchen auf lediglich knapp 18 Treffer. Dies bedeutet, daß bei diesem Thema die MES-Recherchen mehr als doppelt so ergiebig waren wie die FUL-Recherchen. Insgesamt weist MES einen Vorsprung

von 281 relevanten Treffern auf. Damit gehen 72% der insgesamt 392 relevanten Treffer, die MES im Vorteil ist, auf das Konto dieses Recherchethemas. Die Differenz ist mit einem Eta von .56 auch statistisch hoch signifikant.

Ebenso hochsignifikante Differenzen zugunsten von MES findet man bei den Themen ”Kriminalität” und ”Computer”, allerdings sind die absoluten Zahlen mit 41 bzw. 32 Treffern weit weniger spektakuär als bei dem Thema ”Antisemitismus”. Allerdings verweisen die Eta-Werte mit über .50 auf gravierende (relative) Unterschiede.

Bei dem Thema ”Lean Production” liegt MES mit 11,8 Treffern ebenfalls deutlich besser als FUL mit 8,7 Treffern, auf Grund der relativ geringen Fallzahlen wird die Differenz jedoch nicht als statistisch signifikant ausgewiesen. Bei dem Thema ”Gewalt” ist der Vorsprung von MES nicht erheblich.

Eine Ausnahme von der allgemeinen Tendenz bildet das Thema ”Armut”: Hier wurde mit den FUL-Recherchen durchschnittlich ein relevanter Treffer mehr aufgefunden als mit den MES-Recherchen. Der Eta-Wert beträgt jedoch nur .20 und die Differenz ist statistisch nicht signifikant, so daß sie nicht interpretiert werden sollte.

Insgesamt zeigt sich bei der Betrachtung der gefundenen relevanten Treffer, daß die Vpn mit beiden Rechercheinstrumenten übereinstimmend bei denjenigen Themen viele Treffer gefunden haben, die auch von den ”Experten” als ertragreich (d.h. hohe Ankerzahlen) eingeschätzt worden sind. Ebenso besteht Übereinstimmung bei der Reihenfolge der”mittleren” und die ”kleinen” Themen. Diese grundsätzliche Übereinstimmung soll jedoch den Blick dafür nicht verstellen, daß es bei einzelnen Recherchethemen deutliche Unterschiede zwischen MES und FUL gibt: Bei drei Themen erzielt MES hochsignifikante Vorteile, bei dem Thema ”Antisemitismus” ist der Vorsprung auch in absoluten Zahlen spektakuklär. Daraus und aus den kleineren Vorsprüngen bei zwei weiteren Themen ergibt sich insgesamt ein Vorsprung von MES von 392 mehr gefundenen relevanten Treffern gegenüber FUL.

Betrachtet man nicht die absolute Zahl der gefundenen relevanten Treffer, sondern den relativen Anteil bezogen auf den Anker (d.h. also den Recall), ergibt sich folgendes Ergebnis: Die Themen ”Kriminalität” und ”Computer” haben überdurchschnittliche Werte erzielt, das Thema ”Armut” mit Abstand das schlechteste. Die übrigen drei liegen in der Nähe des Gesamtdurchschnitts von .49.

Bei der besonders interessierenden Betrachtung der Unterschiede zwischen MES und FUL ist aufgrund der Berechnungsformel für den Recall zu erwarten, daß die für die absolute Zahl der erzielten relevanten Treffer ermittelten Differenzen auch hier gelten (der Nenner der Formel ist im Vergleich eine Konstante). Die Tab.3 zeigt folgerichtig identische Eta-Werte und Signifikanzen. Besonders hinzuweisen ist nochmals auf den geradezu katastrophalen Einbruch der FUL-Recherchen bei dem Thema ”Antisemitismus” mit .28 bei einem Gesamtdurchschnitt von .41 und ansonsten sehr ähnlichen Werten bei den fünf anderen Themen von ca. .45. Es stellt sich die Frage nach den Ursachen für diesen schlechten Recall-Wert bei diesem Thema, auch im Vergleich mit dem überdurchschnittlichen Wert von MES von .65. Diesen Ursachen soll im folgenden Kapitel nachgegangen werden. Darüber hinaus sollen die signifikanten Recall-Unterschiede zwischen MES und FUL bei den Themen ”Kriminalität” und ”Computer” analysiert werden.

Die Precision-Werte liegen im Mittel etwas höher als die Recall-Werte, dies gilt insbesonders für das Thema ”Gewalt”, bei dem ein weit überdurchschnittlicher Wert von .80 erzielt worden ist. Das bedeutet, daß nur 20% der erzielten Treffer bei diesem Thema zum Ballast gezählt werden müssen. Allerdings wird dieses hervorragende Ergebnis relativiert durch einen leicht unterdurchschnittlichen Recall-Wert, d.h. die Vpn haben eigentlich zuwenig gefunden. Auch das Thema ”Lean Production” weist deutlich überdurchschnittliche Precision-Werte auf, der Recall ist aber nur durchschnittlich. Das Thema ”Armut” fällt auch bei der Precision mit einem weit unterdurchschnittlichen Wert aus dem Rahmen.

Der Vergleich von MES und FUL erbringt mit der hier verwendeten Makromethode bei keiner der sechs Fragen eine signifikante Differenz. Allerdings fällt auf, daß MES bei denselben drei Fragen leichte Vorsprünge aufweist, bei denen beim Recall signifikante Vorteile für MES bestehen. Das Thema ”Armut” fällt mit einem (nichtsignifikanten) Vorsprung für FUL wiederum aus dem Rahmen.

5 Analyse der Recherchethemen „Antisemitismus", „Kriminalität" und „Computer"

In Kap.4 wurde gezeigt, daß sich MES und FUL bei dem Thema "Antisemitismus in Deutschland nach 1945" hinsichtlich der gefundenen relevanten Treffer und des Recall hochsignifikant unterscheiden. Im MES wurden 281 relevante Treffer mehr gefunden als in FUL. Das sind 72% aller 392 relevanten Treffer, die mit MES in der Summe über alle sechs Themen hinweg mehr gefunden wurden. Insofern erscheint es lohnend, dieses Recherchethema genauer zu analysieren um herauszufinden, wie der Vorsprung von MES gegenüber FUL bei den gefundenen relevanten Treffern bzw. dem Recall zu erklären ist.

Aus den Tabellen 6 und 7 kann man die Auflistungen der jeweils 12 Query-Formulierungen zu dem Thema "Antisemitismus" in MES und in FUL entnehmen. Die Tabellen sind jeweils nach der Höhe des Recalls (bzw. der absoluten Zahl der gefundenen relevanten Treffer) geordnet.

Versuchs-person	Suche: Global / Schlagwörter **und ...**	**Rechercheformulierung**	Ergebnis: Treffer(T), relevante Treffer (relT), Recall(r), Precision(p)	Rangfolge nach Recall
VP14	Global	Antisemitismus? **und** (Bundesrepublik? **oder** DDR)	T: 95	.94
			relT= 59, r= .94, p= .62	
VP19	Global	Antisemit? **und** (Bundesrepublik **oder** DDR)	T: 96	.92
			relT= 58, r= .92, p= .60	
VP6	Global	(Antisemitismus **oder** Judendiskriminierung **oder** judenfeindlich? **oder** Judenfeindschaft **oder** Judenhass? **oder** Judenmord? **oder** Judenprogrom? **oder** Judenverfolgung **oder** Judenvernichtung) **und** (Bundesre? **oder** BRD **oder** DDR)	T: 130	.92
			relT= 58, r= .92, p= .45	
VP18	Global	Antisemitismus? **und** (Bundesrepublik **oder** DDR)	T: 90	.87
	Schlagwörter **und ...**	Antisemitismus	relT= 55, r= .87, p= .61	
VP10	Global	Antisemitismus? **oder** 1950er **oder** 1960er **oder** 1970ff	T: 94	.87
	Schlagwörter **und ...**	Bundesrepublik? **oder** DDR?	relT= 55, r= .87, p= .59	
VP7	Global	((Antisemitismus?) **oder** (Antisemitismus? **und** Ideologie?) **oder** (Antisemitismus **und** Aktion?))	T: 88	.86
	Schlagwörter **und ...**	(Bundesrepublik? **oder** DDR?)	relT= 54, r= .86, p= .61	
VP3	Global	Antisemitismus	T: 88	
	Schlagwörter **und ...**	Bundesrepublik? **oder** DDR	relT= 53, r= .84, p= .60	.84
VP23	Global	Antisemitismus? **und** Bundesrepublik?	T :91	.83
			relT= 52, r= .83, p= .57	
VP15	Global	Antisemitismus **nicht** Geschichte	T: 52	.57
	Schlagwörter **und ...**	Bundesrepublik? **nicht** historische Entwicklung?	relT= 36, r= .57, p= .69	
VP11	Global	Antisemitismus **und** Bundesrepublik **und** DDR	T: 12	.14
			relT= 9, r= .14, p= .75	
VP22	Global	Antisemitismus	T: 36	.05
	Schlagwörter **und ...**	Deutschland	relT= 3, r= .05, p= .08	
VP2	Global	Antisemitismus	T: 16	
	Schlagwörter **und ...**	Deutschland **nicht** zweiter Weltkrieg **nicht** Deutsches Reich **nicht** Mittelalter **nicht** Osteuropa **nicht** französische Revolution **nicht** 18. Jh. **nicht** 19. Jh. **nicht** Arbeiterbewegung	relT=2, r= .03, p= .13	.03
		Auswertung über die 12 Einzelrecherchen:		
		$\overline{x}$ r = .65, $\overline{x}$ p = .53		

Tabelle 6: Rechercheformulierungen zum Thema "Antisemitismus" in MES

Versuchs-person	Rechercheformulierung	Ergebnis: Treffer(T), relevante Treffer (relT), Recall(r), Precision(p)	Rangfolge nach Recall
VP21	(Antisemit% **oder** Juden) **und** (Deutschland% **oder** Bundesrepublik% **oder** DDR **oder** BRD)	T: 150 relT= 42, r= .67, p= .28	.67
VP1	Antisemitismus **und** (Bundesrepublik **oder** DDR **oder** Ideologie)	T: 48 relT= 31, r= .49, p= .65	.49
VP16	Antisemitismus **und** (Bundesrepublik **oder** DDR)	T: 31 relT= 28, r= .44, p= .90	.44
VP20	(BRD **oder** DDR **oder** Bundesrepublik **oder** deutsche demok%) **und** Antisemi% **nicht** Nationalstaat	T: 31 relT= 26, r= .41, p= .84	.41
VP8	(Antisemitismus **oder** Neo-Nazismus **oder** Judenfeindlichkeit) **und** (Deutschland **oder** DDR **oder** BRD) **nicht** (historisch **oder** geschichtlich **oder** Historiker **oder** Geschichte **oder** fremdenfeindlich) **nicht** (europäisch **oder** international **oder** Europa **oder** Rechtsextremismus **oder** Rechtsradikalismus)	T: 47 relT= 21, r= .33, p= .45	.33
VP17	(Antisemit% **oder** Judenfeind%) **und** (%1945 **oder** %1946 **oder** %1947 **oder** %1948 **oder** %1949 **oder** %195% **oder** %196% **oder** %197% **oder** %198%)	T: 43 relT= 17, r= .27, p= .40	.27
VP4	(Antisemitismus **oder** Ideologie) **und** Deutschland	T: 52 relT= 14, r= .22, p= .27	.22
VP9	(Antisemitismus **oder** Rassismus%) **und** (Bundesrepublik **oder** Faschis% **oder** Presseberichte **oder** Empir%) **nicht** Rechtsextrem%	T :88 relT= 13, r= .21, p= .15	.21
VP5	Antisemitismus% **und** (Deutsch% **oder** DDR **oder** BRD) **und** 194%	T: 20 relT= 8, r= .13, p= .40	.13
VP12	Antisemitismus% **und** (Ideologie% **oder** Aktionen% **oder** Aktivitäten%)	T: 20 relT= 5, r= .08, p= .25	.08
VP13	(Antisemit% **oder** Jude) **und** (Deutschland **oder** Bundesrepublik **oder** DDR) **und** (1945 **oder** Nachkriegszeit) **nicht** (1933-1945 **oder** 1933 bis 1945 **oder** 1870-1945 **oder** Alliier%)	T: 18 relT= 4, r= .06, p= .22	.06
VP24	Antisemitismus% **und** Deutschland% **und** nach 1945	T: 4 relT=4, r= .06, p= 1.00	.06
	Auswertung über die 12 Einzelrecherchen:		
	$\overline{x}$ r = .28, $\overline{x}$ p = .48		

Tabelle 7: Rechercheformulierungen zum Thema "Antisemitismus" in FUL

Betrachtet man die Rechercheformulierungen in MES, so kann man erkennen, daß 9 der 12 Recherchen überdurchschnittliche Recall-Werte aufweisen: Der Mittelwert des Recall bei allen 72 Recherchen in MES liegt bei .56, die neun besten MES-Recherchen zum Thema ”Antisemitismus” haben Recall-Werte zwischen .94 und .57. Die drei restlichen Recherchen fallen im Vergleich dazu deutlich ab und erreichen Recall-Werte von lediglich .14 bis .03.

Die hohen Recall-Werte kommen dadurch zustande, daß 9 Vpn bei MES durch eine Und- Kombination der Suchbegriffe ”Antisemitismus” sowie ”Bundesrepublik oder DDR” (zwei mit Weglassung von ”DDR”) entweder innerhalb der Globalen Suche oder in der Kombination zwischen der Globalen Suche und dem Schlagwörterfeld zwischen 59 und 36 relevante Treffer gefunden haben. Die drei weniger erfolgreichen Recherchen zeichnen sich dadurch aus, daß in zwei Fällen mit dem (historisch gemeinten) Schlagwort ”Deutschland” (statt ”Bundesrepublik”) gesucht wurde. In einem Fall wurde ein falscher logischer Operator (”Bundesrepublik **und** DDR”) verwendet. In diesen Fällen wurden lediglich 9 bis 2 relevante Treffer identifiziert.

Betrachtet man dagegen die Rechercheergebnisse in FUL, erkennt man, daß nur vier der 12 Recherchen für FUL überdurchschnittliche oder durchschnittliche Recall-Werte aufweisen (bezogen auf den durchschnittlichen Recall-Wert von .41 aller 72 FUL-Recherchen). Die Werte der vier besten Recherchen schwanken zwischen .67 und .41. Die Zahl der dabei gefundenen relevanten Treffer liegt zwischen 42 und 26. Auf der anderen Seite gibt es drei Recherchen, die mit Recall-Werten von weniger als .10 als mißglückt bezeichnet werden müssen. Dabei werden 5 bzw. 4 relevante Treffer identifiziert.

Im Vergleich mit MES erreicht nur die beste FUL-Recherche mit .67 den Durchschnittswert aller MES-Recherchen von .65. 8 von 12 MES-Recherchen erreichen jeweils bessere Recall-Werte und höhere relevante Trefferzahlen als die beste FUL-Recherche.

Zur Erklärung dieses beträchtlichen Unterschiedes kann man die Query-Formulierungen von FUL heranziehen. Geht man von der oben dargestellten erfolgreichen Recherchestrategie in MES aus und vergleicht sie mit der besten Rechercheformulierung in FUL, kann man sehen, daß auch in FUL die Verbindung von ”Antisemitismus”und ”Bundesrepublik oder DDR” (mit Zusätzen) bei den vier besten Recherchen zu akzeptablen Ergebnissen führt. Die Zahl der relevanten Treffer liegt allerdings mit maximal 42 (in diesem Fall sogar erkauft durch eine schlechte Precision aufgrund der maximalen Ergebnisliste mit 150 Treffern) deutlich unter der Trefferzahl der besten MES-Recherchen, die im Prinzip mit derselben Strategie arbeiten. Die 8 anderen Recherchen mit FUL

verwenden z.T. andere (inhaltlich weniger zutreffende) Suchbegriffe oder restriktive Ausschlußbedingungen mit Und- bzw. Nicht-Kombinationen (z.B. Antisemitismus? **und** Deutschland? **und** nach 1945), die zu einer geringen Zahl relevanter Treffer führen.

Damit ist festzuhalten, daß in FUL nur vier, in MES dagegen 9 Vpn mit einer erfolgreichen Recherchestrategie arbeiten. Ein Teil des Vorsprungs von MES dürfte demnach aus der Wahl der Suchbegriffe bzw. ihrer Kombination zu erklären sein. Ein anderer Teil liegt jedoch darin begründet, daß bei FUL auch die Verwendung einer erfolgversprechenden Recherchestrategie weit weniger relevante Treffer erbringt als bei MES. Dieser Unterschied zwischen MES und FUL in der Zahl der gefundenen relevanten Treffer bei vergleichbarer erfolgversprechender Recherchestrategie könnte darauf zurückzuführen sein, daß FUL bei der experimentellen Testsituation auf die Suche im Text der Dokumente eingeschränkt ist, während MES zusätzlich auf vergebene Schlagwörter (entweder im Rahmen der Globalen Suche oder im Suchfeld ”Schlagwörter”) zurückgreifen kann. Die erfolgreichen Vpn bei MES haben durchweg von dieser Möglichkeit Gebrauch gemacht und neben ”Antisemitismus” gleichzeitig mit den Schlagwörtern ”Bundesrepublik” und/oder ”DDR” gesucht. Sie können dann auch Dokumente finden, in denen diese Suchbegriffe nur als Schlagwörter und nicht als Textbestandteil vorhanden sind. Bei der Freitextsuche in FUL müssen diese Dokumente notwendigerweise (bei der Verwendung von Und-Kombinationen in der Query) systematisch verpaßt werden, wenn sie nicht durch die Ausweitung der Recherche um zusätzliche Textbegriffe wie ”BRD” oder ”Deutschland” erfaßt werden können. Im Falle von ”Deutschland” ist zudem zu bedenken, daß dieser Suchbegriff im Freitext zu einer sehr großen Treffermenge mit erheblichem Ballast (Dokumente, die sich auf die Zeit vor 1945 beziehen) führt und sich deshalb nicht gut als Ersatz für das fehlende (präzisere) Schlagwort ”Bundes- republik” eignet. Bei dem Versuch, diese große Treffermenge wiederum einzuschränken, kann es dann leicht zu einem unbeabsichtigten Kollaps beim Rechercheergebnis oder zumindest zum unbeabsichtigten Ausschluß vieler relevanter Treffer kommen. Insofern sind etliche der weniger erfolgreichen Recherchestrategien bei FUL aus der Nicht-Verfügbarkeit treffender Schlagwörter zu erklären.

Die Annahme, daß der Vorsprung von MES gegenüber FUL vor allem mit der Verwendung von Suchbegriffen aus dem Feld ”Schlagwörter” zu erklären ist, wurde anhand einer inhaltlichen Analyse aller 63 relevanten Dokumente (Anker) des Recherchethemas ”Antisemitismus” überprüft. Dabei wurden die Dokumente danach unterschieden, ob sie die häufig verwendeten Suchbegriffe ”Antisemitismus”, ”Bundesrepublik Deutschland” und ”DDR” im Text (z.B. im

Abstract) enthalten oder ob ”Antisemitismus” und/oder ”Bundesrepublik Deutschland oder DDR” nur als Schlagwörter vorkommen.

Insgesamt wurden 31 Anker-Dokumente identifiziert, bei denen der letztgenannte Sachverhalt zutrifft. Von diesen 31 Dokumenten enthalten 24 die Suchbegriffe ”Bundesrepublik Deutschland” oder ”DDR” nur im Schlagwortfeld, 5 den Suchbegriff ”Antisemitismus” nur im Schlagwortfeld und 2 die beiden Suchbegriffe ”Bundesrepublik Deutschland” und ”Antisemitismus” nur im Schlagwortfeld. Bei der Suche mit MES entfielen auf diese 31 Anker-Dokumente insgesamt 268 Treffer, mit FUL wurden bei diesen Dokumenten 39 Treffer erzielt. Das bedeutet, daß MES bei diesen 31 Dokumenten einen Vorsprung von 229 Treffern aufweist. Das sind 81,5% der 281 Treffer, die MES gegenüber FUL bei diesem Recherchethema im Vorteil ist.

Danach kann man als bestätigt festhalten, daß ein großer Teil des Vorsprungs von MES gegenüber FUL bei dem Recherchethema ”Antisemitismus” auf die Existenz und die Verwendung von Schlagwörtern zurückzuführen ist, und zwar hauptsächlich auf das Schlagwort ”Bundesrepublik Deutschland”.

Im folgenden wird überprüft, ob die signifikanten Recall-Vorteile von MES gegenüber FUL bei den Themen ”Kriminalität” und ”Computer” ebenfalls im wesentlichen auf die Existenz und Verwendung von Schlagwörtern zurückzuführen ist.

Die 12 Recherchen in MES zum Thema ”Kriminalität bei Frauen” erbringen durchschnittlich 11,8 relevante Treffer. Bei den 12 Recherchen in FUL zu demselben Thema wurden im Durchschnitt lediglich 8,3 relevante Treffer gefunden. Insgesamt beträgt der Vorsprung von MES gegenüber FUL bei diesem Thema 41 relevante Treffer (vgl. Tab. 3). Dies sind 10,5% aller 392 relevanten Treffer, mit denen MES gegenüber FUL im Vorteil ist. In absoluten Zahlen erscheint der durchschnittliche Vorsprung von 3,5 relevanten Treffern pro Recherche im Vergleich zum Thema ”Antisemitismus” (mit über 23 relevanten Treffern) nicht besonders hoch. Angesichts der Tatsache, daß der Anker bei diesem Thema jedoch nur 18 Dokumente beträgt, ist der Vorsprung von MES allerdings beachtlich (42% bezogen auf den Wert von FUL) und auch statistisch hoch signifikant. Der Eta-Wert liegt mit .54 ganz in der Nähe des Eta-Wertes beim Antisemitismus-Thema (.56).

Der Recall der MES-Recherchen beträgt .65, derjenige der FUL-Recherchen .46. Diese Werte liegen in beiden Fällen etwas über dem Gesamtdurchschnitt jeweils von MES und von FUL. Bei der Precision dagegen bestehen diese Un-

terschiede zwischen MES und FUL und jeweils bei MES und FUL zu den übrigen Themen nicht. Die genauere Analyse der Query-Formulierungen kann zeigen, worin die Besonderheiten dieses Recherchethemas bestehen und wie die Recall-Unterschiede zwischen MES und FUL zustandekommen. Aus den Tabellen 8 und 9 kann man die von den Vpn gefundenen 12 Query-Formulierungen zu dem Thema ”Kriminalität bei Frauen” in MES und in FUL entnehmen. Die Tabellen sind wiederum nach der Höhe des Recalls geordnet. Bei den MES-Recherchen fällt zunächst die Homogenität der Ergebnisse auf: Bei 11 von 12 Recherchen werden zwischen 11 und 13 relevante Treffer gefunden. Das ergibt einen Recall von .61 bis .72, also bezogen auf alle MES-Recherchen überdurchschnittlich gute Ergebnisse. Die Vpn erzielen diese guten Ergebnisse entweder mit Suchen ausschließlich innerhalb der Globalen Suche (8 Fälle), in einer Kombination von Globaler Suche und dem Schlagwortfeld (3 Fälle) oder ausschließlich innerhalb des Schlagwortfelds (1 Fall). Alle drei Strategien sind ähnlich erfolgreich.

Versuchs-person	Suche: Global / Schlagwörter **und ...**	**Rechercheformulierung**	Ergebnis: Treffer(T), relevante Treffer (relT), Recall(r), Precision(p) [Anker: 18]	Rangfolge nach Recall
VP10	Global	(Kriminalitaet? oder kriminelle?) und Frauen?	T: 22	.72
	Schlagwörter **und nicht**	historische Entwicklung	relT= 13, r= .72, p= .59	
VP14	Global	(Kriminalitaet oder Delinquenz oder Strafvollzug oder Resozialisierung) und Frauen	T: 24	.72
			relT= 13, r= .72, p= .54	
VP 23	Global	Frauen? und Kriminalitaet?	T: 22	.67
			relT= 12, r= .67, p= .55	
VP 7	Global	(Kriminalitaet und Frauen) oder Resozialisierung oder (Strafvollzug und Frauen) nicht (Maedchen oder Statistik)	T: 24	.67
			relT= 12, r= .67, p= .50	
VP 19	Global	Kriminali? und Frau? nicht historisch	T: 24	.67
			relT= 12, r= .67, p= .50	
VP 6	Schlagwörter **und ...**	Kriminalitaet? und (Frau? oder Maedchen? oder jugendlich?)	T: 37	.67
			relT= 12, r= .67, p= .32	
VP 2	Global	Kriminalitaet	T: 15	
	Schlagwörter **und ...**	Frau nicht Geschichte nicht Frauenbewegung nicht Mittelalter nicht historische Entwicklung nicht Kind nicht Opfer	relT= 11, r= .61, p= .73	.61
VP 11	Global	Kriminalitaet und Frauen	T :20	.61
			relT= 11, r= .61, p= .55	
VP 22	Global	Frauen und Kriminalitaet	T: 20	.61
			relT= 11, r= .61, p= .55	
VP 18	Global	(Kriminalitaet? oder Delinquenz?)	T: 23	.61
	Schlagwörter **und ...**	Frau?	relT=11, r= .61, p= .48	
VP 3	Global	Frau? und (Kriminalitaet? oder Delinquen?)	T: 25	.61
			relT= 11, r= .61, p= .44	
VP 15	Global	Frauenkriminalitaet oder (Strafvollzug und Frauen)	T: 9	
			relT=9, r= .50, p= 1.00	.50
		Auswertung über die 12 Einzelrecherchen:		
		$\bar{x}$ r = .65 , $\bar{x}$ p = .57		

Tabelle 8: Rechercheformulierungen zum Thema "Kriminalität" in MES

Versuchs-person	Rechercheformulierung	Ergebnis: Treffer(T), relevante Treffer (relT), Recall(r), Precision(p) [Anker: 18]	Rangfolge nach Recall
VP17	(%Kriminalität% oder Verbrechen% %Straf%) und (%Frauen% oder %weib%)	T: 48 relT= 17, r= .94, p= .35	.94
VP24	Frau% und (%Kriminalität% oder Delinquenz% oder Strafvollzug%)	T: 18 relT= 12, r= .67, p= .67	.67
VP20	(Kriminalität% oder Straf% oder Resozial%) und Faru% nicht (Einwanderungsland oder Jedefrau oder Auseinandersetzung)	T: 29 relT= 11, r= .61, p= .38	.61
VP21	(Kriminalität% oder Delinquenz% oder Verbrechen%) und (Frau % oder weibl%)	T: 19 relT= 10, r= .56, p= .53	.56
VP1	(Kriminalität oder Strafvollzug) und Frauen nicht Terrorismus	T: 11 relT= 8, r= .44, p= .73	.44
VP4	(Kriminalität oder Delinquenz oder Resosialisierung oder Strafvollzug) und Frauen	T: 11 relT= 8, r= .44, p= .73	.44
VP13	(Frauen% oder weib% oder Frauenkriminalität) und (straffällig% oder kriminell%)	T: 10 relT= 7, r= .39, p= .70	.44
VP9	(Frauen% oder Kriminalität) und (Straf% oder Resozi%) nicht männl%	T :27 relT= 7, r= .39, p= .26	.33
VP8	(Kriminalität oder Delinquenz oder Straffälligkeit) und (Frau% oder weib%) nicht (Hexen oder Mittelalter oder Jedefrau oder Mensch als Ware oder Fremdenfeindlichkeit oder junger Ausländer)	T: 8 relT= 6, r= .33, p= .75	.33
VP5	Frau% und Kriminalität% nicht Mädchen% nicht Ausländer	T: 8 relT= 5, r= .28, p= .63	.28
VP16	(Kriminalität oder Resozialisierung) und Frauen	T: 8 relT= 5, r= .28, p= .63	.28
VP12	(Delinquenz% oder Resozialisierung% oder Strafvollzug% oder Wiedereingliederung%) und Frauen	T: 31 relT=4, r= .22, p= .13	.22
	Auswertung über die 12 Einzelrecherchen:		
	$\bar{x}$ r = .46 , $\bar{x}$ p = .54		

Tabelle 9: Rechercheformulierungen zum Thema "Kriminalität" in FUL

Die FUL-Recherchen sind heterogener und im Schnitt deutlich schlechter. Lediglich die 4 besten Recherchen liegen zwischen .94 und .56 und übertreffen damit den durchschnittlichen Recall-Wert aller 144 Recherchen von .49. Auch FUL-intern liegen nur 5 der 12 Recherchen über dem Durchschnittswert aller 72 FUL-Recherchen von .41. 4 der 12 FUL-Recherchen müssen mit Recall-Werten zwischen .22 und .33 als weitgehend gescheitert bezeichnet werden.

Die gewählten erfolgreichen Query-Formulierungen in MES zeigen, daß bereits eine einfache Und-Verbindung zwischen den beiden Suchbegriffen ”Kriminalität” und ”Frau” (evt. mit Truncierungen) zu guten Ergebnissen führt (z.B. bei VP 23). Eine geringfügige Verbesserung kann durch eine Oder-Erweiterung (z.B. mit ”kriminelle”, ”Delinquenz”, ”Strafvollzug”) erreicht werden. Die Einschränkung ”nicht historisch” oder ”nicht historische Entwicklung” kann zu einer Verbesserung der Precision führen und ist für den Recall zumindest unschädlich.

Eine ähnlich einfache Vorgehensweise führt bei FUL nicht zum Erfolg, beispielsweise bei VP 16. Die (wenigen) erfolgreichen Recherchen zeichnen sich dadurch aus, daß die beiden zentralen Suchbegriffe ”Kriminalität” und ”Frau” in jedem Fall trunciert werden, evtl. sogar mit Linkstruncierung. Gleichzeitig wird mit Oder- Erweiterungen gearbeitet. VP 24 beispielsweise kommt mit dieser Strategie auf 12 relevante Treffer.

Dies bedeutet, daß auch in FUL durchaus akzeptable Ergebnisse erzielt werden können, die den MES-Werten entsprechen. Offensichtlich fällt es den Vpn allerdings häufig viel schwerer, eine erfolgreiche Query zu formulieren (beispielsweise wird die Truncierung vergessen).

Dieses Ergebnis bedeutet, daß MES bei dem Thema ”Kriminalität bei Frauen” vor allem deshalb einen signifikant höheren Recall aufweist, weil auch einfache Rechercheformulierungen, d.h. die schlichte Und-Verbindung von ”Kriminalität und Frau”, bessere Werte als bei FUL erbringen. Das liegt wiederum daran, daß auch bei der Globalen Suche (die von den meisten Vpn gewählt wurde) die Schlagwörter ”Kriminalität” und ”Frau” automatisch einbezogen werden, so daß auch ohne Truncierung bzw. Oder-Erweiterungen die Zieldokumente aufgrund der Verschlagwortung gefunden werden. Die Analyse der 18 Anker-Dokumente zeigt, daß in der Tat bei 5 Dokumenten die Begriffe ”Kriminalität” oder ”Frau” nicht im Text, sondern nur als Schlagwörter vorhanden sind. In FUL können diese Dokumente nur über Oder-Erweiterungen (z.B. Einbeziehung von Synonymen) bzw. über Truncierungen gefunden werden. Beide Strategien haben jedoch den Nachteil, daß u.U. die Precision empfindlich verschlechtet wird. Das wiederum führt dazu, daß durch eine Spezifikation der

Query mit ”und nicht” einzelne unzutreffende Dokumente ausselektiert werden sollen (z.B. VP 8). Dadurch entsteht die Tendenz, daß die Queries schwierig, fehleranfällig und unübersichtlich werden. Evtl. werden sogar unbeabsichtigt relevante Treffer ausgeschlossen. Diesen Effekt kann man vor allem bei dem folgenden Recherchethema studieren.

Das Thema ”Computer im Alltag” ähnelt in mehrfacher Hinsicht dem Thema ”Kriminalität bei Frauen”: Mit MES werden im Durchschnitt deutlich mehr relevante Treffer pro Recherche erzielt als mit FUL (7,5 gegenüber 4,8, vgl. Tab.3). Dementsprechend ist der Recall besser (.68 gegenüber .44). Der Vorsprung von MES gegenüber FUL in absoluten Zahlen ist mit insgesamt 32 relevanten Treffern im Vergleich zum Thema ”Antisemitismus” relativ gering (8,2% des Vorsprungs von 392 relevanten Treffern). Der Anker ist mit 11 Dokumenten jedoch recht niedrig, so daß die durchschnittliche Differenz von 2,7 relevanten Treffern erheblich (56,3% auf der Basis des FUL-Wertes) und auch statistisch hochsignifikant ist. Der Eta-Wert liegt mit .52 nahe bei den Werten der Themen ”Antisemitismus” und ”Kriminalität bei Frauen”.

Der Recall liegt bei MES mit .68 deutlich über dem Mittelwert aller 72 MES-Recherchen von .56. Der Recall bei FUL liegt mit .44 nur unerheblich über dem gesamten FUL-Wert von .41. Bei der Precision ergeben sich bei den MES-Recherchen deutlich bessere Werte als bei den FUL-Recherchen. Der Eta-Wert ist mit .31 mit Abstand der höchste im Vergleich der 6 Recherchethemen. Dennoch ist die Differenz statistisch nicht signifikant.

Betrachtet man die 12 MES-Recherchen im einzelnen, zeigt sich in bezug auf den Recall wie bei dem Thema ”Kriminalität” eine beachtenswerte Homogenität auf hohem Qualitätsniveau: 11 von 12 Recherchen erzielen 7 bis 9 relevante Treffer (von 11 möglichen). Sie übertreffen mit einem Recall von .82 bis .64 den MES-Durch-schnitt von .56. Lediglich bei einer Recherche wird mit 5 relevanten Treffern ein Ergebnis von .45 erzielt.

Versuchs-person	Suche: Global / Schlagwörter **und ...**	**Rechercheformulierung**	Ergebnis: Treffer(T), relevante Treffer (relT), Recall(r), Precision(p) [Anker: 11]	Rangfolge nach Recall
VP12	Global	Computer und (Freizeit oder Hobby oder Alltag oder Haushalt oder Privat?)	T: 22	.82
			relT= 9, r= .82, p= .41	
VP17	Global	PC oder Computer oder Technikeinsatz? oder Informationstechnik	T: 23	.82
	Schlagwörter **und ...**	Alltag? oder ?Privat? oder ?Spiel?	relT= 9, r= .82, p= .39	
VP 21	Global	Computer? und (Allt? oder Freizeit?)	T: 30	.82
			relT= 9, r= .82, p= .30	
VP 5	Global	Computer? und (Alltag? oder alltaeglich?)	T: 22	.73
	Schlagwörter **und nicht**	Beruf? und Unterricht? oder Ausbildung?	relT= 8, r= .73, p= .36	
VP 1	Global	Computer	T: 7	.64
	Schlagwörter **und ...**	Alltag	relT= 7, r= .64, p= 1.00	
VP 8	Global	Computer	T: 7	.64
	Schlagwörter **und ...**	Alltag	relT= 7, r= .64, p= 1.00	
VP 4	Global	Computer oder PC	T: 9	
	Schlagwörter **und ...**	Alltag oder Nutzung	relT= 7, r= .64, p= .78	.64
VP 16	Global	Computer und Alltag	T :10	.64
			relT= 7, r= .64, p= .70	
VP 20	Global	(Computer oder PC) und Alltag	T: 10	.64
	Schlagwörter **und nicht**	Beruf	relT= 7, r= .64, p= .70	
VP 13	Global	(Computer oder Computeralltag oder EDV oder technischer Wandel) und Alltag	T: 15	.64
	Schlagwörter **und nicht**	Unterricht oder ?Bildung	relT= 7, r= .64, p= .47	
VP 9	Global	(Alltag oder alltaegl?) und (Computer oder Computerarbeit oder Computerarbeitsplaetze?) oder Technikangst oder Technikauseinandersetzung oder Technikbeziehung oder Computerbenutzer	T: 17	.64
			relT= 7, r= .64, p= .41	
VP 24	Global	Computer? und Freizeit?	T: 10	
	Schlagwörter **und nicht**	Beruf?	relT=5, r= .45, p= .50	.45
		Auswertung über die 12 Einzelrecherchen:		
		$\bar{x}$ r = .68 , $\bar{x}$ p = .59		

Tabelle 10: Rechercheformulierungen zum Thema "Computer" in MES

Versuchs-person	**Rechercheformulierung**	Ergebnis: Treffer(T), relevante Treffer (relT), Recall(r), Precision(p) [Anker: 11]	Rangfolge nach Recall
VP11	Computer% und (Alltag% oder Nutzung%) nicht ("Beruf und Computer" oder Unterricht% oder Ausbildung% oder Technikeinsatz% oder Lernprogramme% oder Didaktik% oder Automatisierung%)	T: 38 relT= 9, r= .82, p= .24	.82
VP19	(Computer% oder EDV% oder Terminal%) und (Alltag% oder Haushalt%) nicht (Beruf% oder Unterricht oder Ausbildung)	T: 21 relT= 8, r= .73, p= .38	.73
VP14	(Computer% oder Datenverarbeitung oder PC) und (Alltag oder alltäglich%) nicht (Beruf% oder Ausbildung)	T: 14 relT= 7, r= .64, p= .50	.64
VP23	Computer% und Alltag% nicht (Beruf% oder Unterricht% oder Ausbildung%)	T: 17 relT= 7, r= .64, p= .41	.64
VP7	Computer% und Alltag nicht Beruf%	T: 9 relT= 6, r= .55, p= .67	.55
VP3	Computer% und Alltag% nicht (Beruf% oder Ausbildung% oder Unterricht% oder Produktion%)	T: 11 relT= 6, r= .55, p= .55	.55
VP22	Computer und (Alltag oder Freizeit oder Hobby)	T: 7 relT= 5, r= .45, p= .71	.45
VP2	(Computer oder PC) und Alltag nicht (Beruf oder Unterricht oder ausbilden)	T :5 relT= 4, r= .36, p= .80	.36
VP6	(Computer oder PC oder Rechner) und Alltag nicht (tägliche Leben oder privat oder zuhause oder Hobby)	T: 6 relT= 4, r= .36, p= .67	.36
VP10	Computer% und Alltag% nicht (Beruf% oder Unterricht% oder Technik%)	T: 7 relT= 1, r= .09, p= .14	.09
VP18	(Computer% oder EDV% oder Internet% oder PC%) und %Alltag% nicht (Ausbildung oder Technik oder Unterricht)	T: 10 relT= 1, r= .09, p= .10	.09
VP15	(Computer oder PC oder Internet) und (Alltag oder Freizeit oder Spiel) nicht (Beruf% oder Unterricht% oder Ausbildung% oder Technik% oder Betrieb% oder Arbeit)	T: 2 relT=0, r= .00, p= .00	.00
	Auswertung über die 12 Einzelrecherchen:		
	$\overline{x}$ r = .44, $\overline{x}$ p = .43		

Tabelle 11: Rechercheformulierungen zum Thema "Computer" in FUL

Die FUL-Recherchen sind wiederum wesentlich heterogener und im Recall teilweise drastisch schlechter: Die Recall-Werte reichen von .82 bis .00; drei Recherchen sind mit 1 bzw. 0 relevanten Treffern völlig mißlungen. Auf der anderen Seite erreichen 4 Recherchen in FUL mit 7 bis 9 relevanten Treffern die in MES üblichen guten Werte.

Betrachtet man die Query-Formulierungen in MES, fällt die häufige Kombination der Globalen Suche mit dem Schlagwortfeld auf (8 Fälle). In der Hälfte dieser Fälle wird das Schlagwortfeld zum Ausschluß von Dokumenten mit ”und nicht” verwendet. In 4 Fällen wird nur mit der Globalen Suche recherchiert. Insgesamt ist jedoch nicht erkennbar, daß eine der beiden Strategien eindeutig bessere Recall-Ergebnisse als die andere zeitigen würde.

Der Erfolg der MES-Recherchen besteht darin, daß grundsätzlich die Suchbegriffe ”Computer” und ”Alltag” (teilweise trunciert) durch eine Und-Verbindung verknüpft werden, wobei zusätzlich teilweise Oder-Erweiterungen vorgenommen werden. Die einfachste Rechercheformulierung ”Computer und Alltag” (ohne Truncierung) erbringt mit MES immerhin 7 relevante Treffer und damit einen Recall von .64. Die Oder-Erweiterungen erbringen im besten Fall zwei zusätzliche relevante Treffer. Dabei müssen allerdings teilweise relativ schlechte Precision-Werte in Kauf genommen werden (zwischen .30 und .41 bei den 4 nach dem Recall besten Recherchen). Ähnlich wie bei dem Thema ”Kriminalität” kann man auch hier feststellen, daß bei MES mit einfachen Rechercheformulierungen recht akzeptable Recall-Werte erreicht werden (können).

Die einfachste Rechercheformulierung ”Computer und Alltag” (teilweise mit Truncierung) erbringt mit FUL 6 bis 7 relevante Treffer und liegt damit praktisch mit MES gleichauf. Die Oder-Erweiterungen können wie bei MES zwei zusätzliche relevante Treffer erbringen. Die Freitextrecherche bei FUL ist insofern der Globalen oder der kombinierten Suche unter Einbeziehung von Schlagwörtern bei MES bezüglich des Recall nicht grundsätzlich unterlegen. Es stellt sich die Frage, wie die signifikante Differenz zwischen MES und FUL dann zu erklären ist.

Betrachtet man die 8 FUL-Recherchen, die den 11 besten MES-Recherchen im Recall unterlegen sind (Recall von .55 und weniger), fällt auf, daß bei 7 dieser 8 Recherchen mit einer ”und nicht”-Verknüpfung gearbeitet wird, d.h. es werden alle Dokumente ausgeschlossen, die im Text einen derartigen Such- bzw. Ausschlußbegriff enthalten. Bei Freitextrecherchen ist die großzügige Vergabe von Ausschlußbegriffen besonders riskant, da auch relevante Dokumente von der Tilgung aus der Ergebnisliste betroffen sein können. Genau dies ist offenbar bei den drei schlechtesten Recherchen geschehen: Durch den Auschluß aller Doku-

mente, die den Begriff ”Technik” enthalten, wird das Rechercheergebnis zerstört. Allein dieser Recherchefehler führt zum Verlust von 15 bis 20 relevanten Dokumenten bei FUL, also dem größeren Teil der Differenz zwischen MES und FUL. Dazu kommt noch der vollständige oder teilweise Verzicht auf die Truncierung bei den zentralen Suchbegriffen ”Computer” und ”Alltag” bei 4 FUL-Recherchen. Dies macht einen zusätzlichen Verlust von 7 bis 9 relevanten Treffern aus.

Bei den 4 gelungenen FUL-Recherchen, die mit 7 bis 9 relevanten Treffern an die MES-Werte heranreichen, fällt dagegen auf, daß fast durchgängig mit Truncierungen und mit (truncierten) Oder-Erweiterungen gearbeitet wird. Die dadurch entstehenden Verluste bei der Precision sollen offenbar durch "und nicht”- Einschränkungen aufgefangen werden, was im Falle der 4 besten Recherchen insofern einigermaßen gut gelingt, als diese Einschränkungen durch die richtige Auswahl der Ausschlußbegriffe nicht zu Lasten des Recalls gehen.

Der Vergleich der 12 MES- mit den 12 FUL-Recherchen zeigt, daß bei MES in 4 Fällen mit einer ”und nicht”-Verknüpfung gearbeitet wird, bei FUL dagegen in 11 Fällen. Die Summe der durch ”und nicht” ausgeschlossenen Begriffe beträgt bei MES 7, bei FUL dagegen 39. Es liegt nahe zu vermuten, daß die bei diesem Thema zunächst erforderliche Truncierung der zentralen Suchbegriffe ”Computer” und ”Alltag” bei der reinen Freitextsuche zu unerwünschten Zwischenergebnissen führt (zu lange Trefferliste mit zuviel Ballast). Die dann naheliegende Einschränkung der Liste durch eine ”und nicht”-Bedingung kann bei relativ unerfahrenen Rechercheuren dazu führen, daß die in der Ausgangsrecherche bereits gefundenen relevanten Treffer durch einen einschlägigen Ausschlußbegrff (hier:”Technik”) wieder verlorengehen.

6 Einflußgrößen auf die Kriteriumsvariable „Recall“

Die bisherigen Analysen haben gezeigt, daß die Höhe des Recall bei den 144 Recherchen sowohl vom verwendeten Rechercheinstrument als auch vom Recherchethema abhängt. Es soll nun die Frage beantwortet werden, welcher dieser beiden Einflußfaktoren der wichtigere ist.

Ausgangspunkt ist Tabelle 4 mit den nach Recherchethemen geordneten Recall-Mittelwerten der MES- und der FUL-Recherchen. Ordnet man diese 12 Mittelwerte nach ihrer Größe, erhält man folgendes Resultat:

Rang	Recall-Wert	Recherche-Instrument	Thema
1	.6818	Mes	Computer
2	.6534	Mes	Antisemitismus
3	.6528	Mes	Kriminalität
4	.5635	Mes	Lean Production
5	.4689	Mes	Gewalt
6	.4630	Ful	Kriminalität
7	.4527	Ful	Gewalt
8	.4394	Ful	Computer
9	.4359	Ful	Armut
10	.4127	Ful	Lean Production
11	.3526	Mes	Armut
12	.2817	Ful	Antisemitismus
insges.	.4882		

Tabelle 12: Recall-Werte nach Rechercheinstrument und -thema

5 der 6 Recherchethemen, in denen MES verwendet wurde, nehmen die ersten 5 Plätze ein. Lediglich das MES-Thema ”Armut” liegt als Ausreißer auf dem vorletzten Platz. Alle 6 FUL-Recherchethemen belegen mittlere bzw. hintere Plätze. Die Tabelle vermittelt den Eindruck, daß die Höhe des Recall in erster Linie vom Rechercheinstrument abhängt, d.h. die MES-Recherchen weisen mit einer Ausnahme höhere Recall-Werte als die FUL-Recherchen auf. Daneben spielt aber auch das Recherchethema eine gewisse unabhängige Rolle, da z.B. Recherchen zum Thema ”Kriminalität” sowohl bei MES als auch bei FUL höhere Recall-Werte liefern als beispielsweise das Thema ”Lean production”. Allerdings ist die Reihenfolge der Themen bei MES und FUL etwas unterschiedlich, so daß die erzielten Recall-Werte zum Teil auch aus der Wechselwirkung von Rechercheinstrument und Recherchethema zu erklären sind.

Mit Hilfe einer Kovarianzanalyse soll dieser Eindruck statistisch überprüft und quantifiziert werden.

Univariate Varianz- und Kovarianzanalysen untersuchen den Einfluß von einer oder mehreren unabhängigen Variablen auf eine abhängige Variable. Die unabhängigen Variablen sind im Normalfall nominal- oder ordinalskaliert. Sie werden dann als Faktoren bezeichnet. Intervallskalierte unabhängige Variablen bezeichnet man als Kovariaten. Werden sie einbezogen, spricht man von einer Kovarianzanalyse.

Bei der (hier verwendeten) Methode nach Fisher geht es um eine Zerlegung der Gesamtvarianz der abhängigen Variable in eine Varianz innerhalb der durch die Faktoren gebildeten Gruppen und eine Varianz zwischen diesen Gruppen. Die F-Werte der Faktoren und der Interaktionseffekte der Faktoren sind Ausdruck für die relative Bedeutsamkeit dieser Faktoren für die Werte der abhängigen Variable.

Im vorliegenden Fall wird ein Modell mit den beiden Faktoren ”Rechercheinstrument” (MES und FUL) und ”Recherchethema” (6 Ausprägungen) sowie ”Recall” als (metrische) abhängige Variable gebildet. Als Kontrollvariable wird die Kovariate ”Alter” einbezogen.

Haupteffekte	Mittel der Quadrate	F	Signifikanz
Recherche-Instrument MES/FUL	.788	15,06	*
Recherche-Thema	.096	1,85	-
Kombination	.212	4,05	*
Wechselwirkungen			
Recherche-Instrument mit -thema	.158	3,02	*
Kovariate Alter	.015	.28	-
N=144			

***** Signifikanz: 5%-Niveau

Tabelle 13: Kovarianzanalyse auf Recall

Die F-Werte zeigen, daß das Rechercheinstrument einen erheblich höheren Einfluß auf die Größe des Recall hat als das Recherchethema. Der Einfluß des Recherchethemas ist für sich genommen nicht signifikant. Das Alter der Vpn hat keinen Effekt. Es gibt signifikante Wechselwirkungen zwischen Rechercheinstrument und Recherchethema, d.h. ein Teil der Varianz in der Recall-

Verteilung ist nicht den beiden Haupteffekten einzeln, sondern ihrem Zusammenwirken zuzuschreiben.

Dieser Befund ist insofern bedeutsam, als er Hinweise auf die mögliche Generalisierbarkeit der Resultate auf Recherchethemen mit unterschiedlich großem Anker gibt. Wie in Kap.1.5 dargestellt, wurden die sechs Themen auch unter dem Gesichtspunkt unterschiedlicher maximaler relevanter Trefferzahlen ausgewählt. Es zeigt sich nun, daß dieser Aspekt für den erzielten Recall offenbar nur eine untergeordnete Rolle spielt, jedenfalls dann, wenn man die direkten Effekte (Haupteffekte) betrachtet.

Der Einfluß von einzelnen Variablen auf die Höhe des Recall kann auch mithilfe einer (multiplen) Regressionsanalyse überprüft werden, in die neben den beiden Haupteinflußfaktoren Rechercheinstrument und Recherchethema auch weitere unabhängige Variable einbezogen werden.

Die Regressionsanalyse[15] gehört zu der Kategorie der sogenannten "multivariaten Analysemethoden". Multivariate Verfahren stellen ein Bündel verschiedener Methoden dar (z.B. Cluster-, Faktoren-, Diskriminanzanalyse), denen gemeinsam ist, daß sie die gegenseitigen Beziehungen zwischen mehreren Variablen untersuchen und die Komplexität des Datenmaterials reduzieren. In unserem Fall ist die multiple Regressionsanalyse die geeignete Methode aus der Klasse der multivariaten Verfahren. Sie dient der Analyse von Beziehungen zwischen einer abhängigen und mehreren unabhängigen Variablen. Die der Regressionsanalyse zugrundeliegende Frage lautet: Wieviel von der Abweichung der Beobachtungswerte vom Mittelwert der Stichprobe ist auf den Einfluß der unabhängigen Variablen zurückzuführen und wieviel bleibt "unerklärt"? Bei der Regressionsanalyse wird demnach üblicherweise eine lineare Funktion gesucht, die möglichst viel von der gesamten Streuung durch die unabhängigen Variablen erklärt und möglichst wenig Residuen übrig läßt.

Für die Berechnung wurde das Verfahren der schrittweisen Regression gewählt. Bei der schrittweisen Regression werden die unabhängigen Variablen einzeln nacheinander in die Regressionsgleichung einbezogen, wobei jeweils diejenige Variable ausgewählt wird, die ein bestimmtes Gütekriterium maximiert. Zunächst wird eine einfache Regression mit derjenigen Variablen durchgeführt, die die höchste Korrelation mit der abhängigen Variablen aufweist. Danach wird dann jeweils die Variable mit der höchsten partiellen Korrelation ausgewählt.

---

[15] Die folgenden Ausführungen zur Regressionsanalyse beziehen sich auf K.Backhaus/B.Erichson/W.Plinke/Chr.Schuchard-Ficher/R.Weiber: Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. Berlin, Heidelberg, 1989

Aus der Rangfolge der Aufnahme läßt sich die statistische Wichtigkeit der Variablen erkennen.

In die Auswahl der in das Modell aufgenommenen Variablen soll auch die Höhe des ”Ankers” aufgenommen werden, so daß das Ergebnis der Kovarianzanalyse mithilfe eines anderen statistischen Verfahrens überprüft werden kann.

Neben den genannten Variablen sollen zusätzlich Merkmale der Versuchspersonen (u.a. Geschlecht, Recherche-Erfahrung) in das Modell aufgenommen werden, um zu überprüfen, ob neben den beiden kontrollierten Faktoren Rechercheinstrument und Recherchethema noch weitere unkontrollierte Faktoren die Höhe der Recall-Werte beeinflußt haben.

Es geht bei dieser Fragestellung also nicht, wie sonst üblich, darum, durch die Formulierung eines Modells mit möglichst wenigen unabhängigen Variablen einen möglichst hohen Varianzanteil (Determinationskoeffizient R-Quadrat) zu erklären, sondern es geht darum zu zeigen, welche Bedeutung eine Reihe von möglichen Störeinflüssen (unkontrollierten Variablen) für das Versuchsergebnis (Recall) hatten. Es handelt sich um ein explorativ es Vorgehen, bei dem sich die schrittweise Einbeziehung von unabhängigen Variablen anbietet.

Die sechs Ausprägungen der (nominalskalierten) Variable ”Recherchethema” wurden in je eine dichotome Dummy-Variable transformiert, die Variablen ”Geschlecht”, ”Rechercheerfahrung”, ”Interneterfahrung”, ”Boolerfahrung” und ”Rechercheinstrument” blieben unverändert einfache dichotome Variable. ”Alter”, ”Anker” und die Kriteriumsvariable ”Recall” sind metrisch skaliert.

Die (direkten) Pearson-Korrelationen ergeben einen signifikanten Effekt der Variable ”Recherche-Instrument MES/FUL” auf die Höhe des Recall; es handelt sich mit ca. -.30 um den höchsten Wert bei den einbezogenen unabhängigen Variablen. Bei den sechs Recherchethemen korreliert das Thema ”Armut” (negativ) am höchsten mit dem Recall. Dies ist das Ergebnis des bereits in Tab. 4 dargestellten Effekts, daß dieses Thema die mit Abstand geringsten Recall-Werte aufweist.

Unabhängige Variable	Korrelationen mit Recall	Signifikanz
Recherche-Instrument MES/FUL	-.297	*
Thema Armut	-.169	*
Thema Computer	.130	-
Thema Kriminalität	.125	-
Thema Gewalt	-.049	-
Thema Antisemitismus	-.037	-
Thema Lean Production	.000	-
Recherche-Erfahrung	.160	*
Interneterfahrung	.150	*
Bool-Erfahrung	.052	-
Geschlecht	-.126	-
Alter	-.041	-
Anker	-.066	-
N=144		

*Signifikanz: 5%-Niveau

Tabelle 14: Korrelationen mit Recall

Von den persönlichen Merkmalen der Vpn sind die Recherche- und die Interneterfahrung von Belang. Die direkte Korrelation von Anker und Recall erbringt das aus der Kovarianzanalyse erwartete Ergebnis: Der Anker weist keinen relevanten Zusammenhang mit dem Recall auf.

Setzt man die genannten unabhängigen Variablen in einer schrittweisen Regression mit der abhängigen Variablen Recall in Beziehung, erhält man folgendes Modell:

Unabhängige Variable	Beta	t	Signifikanz
Recherche-Instrument MES/FUL	-.297	-3,83	*
Recherchethema Armut	-.244	-2,48	*
Rechercheerfahrung	.100	1,19	-
Anker	-.168	-1,67	-
Geschlecht	-.152	-1,72	-
Interneterfahrung	-.146	1,71	-
Recherchethema Lean Production	-.082	-0,89	-
Bool-Erfahrung	-.032	-0,37	-
Alter	-.016	-0,20	-
Recherchethema Kriminalität	.015	-0,16	-

Konstante	.867
R	.445
R - Quadrat	.198
korrigiertes R - Quadrat	.138
N=144	

* Signifinanz: 5%-Niveau

Tabelle 15: Multiple Regression auf Recall

Das Modell enthält 10 der 13 in der Korrelationstabelle enthaltenen unabhängigen Variablen. Die drei Recherchethemen ”Computer”, ”Gewalt” und ”Antisemitismus” wurden nicht aufgenommen, da die Beta-Werte die vorgesehene Untergrenze für die Einbeziehung unterschreiten.

Es wird bestätigt, daß das Recherche-Instrument MES/FUL für Recall die relativ wichtigste Einflußgröße ist. An zweiter Stelle steht das Recherchethema ”Armut”, dessen Gewicht im Vergleich mit der direkten Korrelation noch etwas ansteigt. Beide Beta-Werte tragen signifikant zur Verbesserung der Varianzer-

klärung beim schrittweisen Einbezug einer weiteren Variable in das Regressionsmodell bei.

Dieses Ergebnis entspricht der Kovarianzanalyse, derzufolge die Bedeutung des Rechercheinstruments für den Recall größer ist als das Recherchethema. Die Regressionsanalyse zeigt nun, daß die verbleibende relativ geringe Bedeutung des Faktors ”Thema” praktisch ausschließlich auf den Ausreißer ”Armut” zurückzuführen ist.

Die Bedeutung dieses Ausreißerthemas für das Ergebnis kann auch anhand der Variable ”Anker” aufgezeigt werden, bei der dieses Thema mit dem Wert 13 (also dem zweitniedrigsten) vertreten ist. Die einfache Korrelation zwischen Anker und Recall ergibt praktisch ein Null-Ergebnis, d.h. die plausible Annahme, der Recall wäre bei einem numerisch kleinen Anker normalerweise größer als bei einem numerisch großen Anker, (bei großen Ankern muß man mehr relevante Treffer finden um einen guten Recall-Wert zu erzielen) ist, über alle sechs Fragen gesehen, falsch. Die Korrelationstabelle zeigt nun allerdings, daß bei zwei Themen mit ”kleinem” oder ”mittlerem” Anker, nämlich ”Computer” und ”Kriminalität” der erwartete Zusammenhang besteht: beide korrelieren positiv mit Recall. Bei dem Thema ”Lean Production”, ein Thema mit ”mittlerem” Anker, findet man eine Null-Korrelation. Ebenso bei den beiden ”großen” Themen ”Gewalt” und ”Antisemitismus”. Lediglich das Thema ”Armut” weist ein Ergebnis entgegen dem Trend aus: Es ist ein ”kleines” Thema, aber der Recall ist relativ schlecht.

Im Ergebnis ergibt sich dann bei der einfachen Korrelation zwischen Anker und Recall das erwähnte Null-Ergebnis. Bei der partiellen Korrelation in dem Regressionsmodell ergibt sich dagegen immerhin ein Beta-Wert von ca. .-17 für den Anker. Das heißt, der vermutete Zusammenhang (je kleiner der Anker, desto größer der Recall) tritt in Erscheinung, wenn auch nur andeutungsweise. Das hängt damit zusammen, daß in dem schrittweise gerechneten Modell das Thema ”Armut” vor der Variablen ”Anker” steht und die partielle Korrelation zwischen Anker und Recall damit quasi um den Einfluß des Themas ”Armut” bereinigt ist.

Es zeigt sich damit auch an dieser Stelle, daß es sich bei dem Thema ”Armut” um einen Sonderfall handelt, der das Gesamtergebnis beeinflußt, allerdings nicht in entscheidender Weise. Im Hinblick auf die Größe des Ankers kann auch bei Berücksichtigung dieses Sonderfalls gesagt werden, daß dieses Merkmal keinen entscheidenden Einfluß auf die Höhe des Recall hat. Oder mit anderen Worten: Der Vorteil von MES gegenüber FUL beim Recall besteht unabhängig von der Zahl der maximal erreichbaren Treffer in einer Recherche.

Der Beitrag der restlichen sieben Variablen zur Erklärung der Höhe der Recall-Werte ist unerheblich. **Das bedeutet, daß es gelungen ist, mögliche Störfaktoren wie unterschiedliche Recherche-, Internet- oder Bool-Erfahrung in ihrer Bedeutung für den Recall zu minimieren.** Dies dürfte ein Ergebnis der rund 20minütigen gründlichen Schulung sein, durch die die unterschiedlichen Recherche-Vorkenntnisse offenbar zum größten Teil ausgeglichen werden konnten.

Ebenso spielen das Lebensalter der Vpn gar keine und die Geschlechtszugehörigkeit nur eine (nichtsignifikante) geringe Rolle für den Recall.

7 Subjektive Bewertungen durch die Vpn

Nach Beendigung einer Recherche wurden die Vpn gebeten, die ausgedruckte Trefferliste nach relevanten und irrelevanten Treffern zu sortieren. Grundlage für die Entscheidung waren in der Regel die Titel der Treffer. Vor allem bei längeren Trefferlisten war die Berücksichtigung der Abstracts aus Zeitgründen praktisch unmöglich.

Treffer	MES	FUL	Summe
Gefundene relevante Treffer	1.299	907	2.206
Treffer nach subjektiver Einschätzung	1.524	1.361	2.885
Differenz gefundener relevanter und subjektiver Treffer	225	454	679
Differenz in % der gefundenen relevanten Treffer	17,3	50,1	

Tabelle 16: Gefundene relevante Treffer und Treffer nach subjektiver Einschätzung

Die summarische Auswertung der subjektiv als relevant eingeschätzten Treffer im Vergleich mit den tatsächlich gefundenen relevanten Treffern zeigt, daß insgesamt weniger relevante Treffer gefunden worden sind als angenommen. MES und FUL unterscheiden sich dabei deutlich: Bei MES beträgt die Differenz zwischen den tatsächlich gefundenen relevanten Treffern und der subjektiven Einschätzung insgesamt 225 Treffer oder 17,3%. (bezogen auf 1.299 relevante Treffer). Bei FUL beträgt diese Differenz dagegen 454 Treffer oder 50,1%.

Das bedeutet, daß die Vpn bei FUL ihr Ergebnis hinsichtlich der tatsächlich gefundenen relevanten Treffer bei FUL wesentlich deutlicher überschätzt haben als bei MES. Offensichtlich hat die vom Anspruch her nach Relevanz geordnete Ergebnisliste bei FUL den Vpn keine Hilfe bei der Relevanzbewertung geben können.

Es stellt sich die Frage, ob die subjektive Überbewertung der FUL-Ergebnisse hinsichtlich der gefundenen relevanten Treffer ihren Niederschlag in der Bewertung von FUL insgesamt findet. Die Vpn wurden am Ende des Tests gebeten, MES und FUL bezüglich der 6 Suchthemen zunächst im einzelnen und dann summarisch zu bewerten.

Die Frage zu den einzelnen Suchthemen lautete: ”Wie gut kamen Sie bei der Formulierung der Suchthemen mit den verschiedenen Systemen zurecht?”

Bewertung	MES		FUL		insgesamt	
	%	N	%	N	%	N
sehr gut	12,5	9	15,3	11	13,9	20
gut	52,8	38	55,5	40	54,1	78
unzureichend	29,2	21	27,8	20	28,5	41
gar nicht	5,5	4	1,4	1	3,5	5
N	100,0	72	100,0	72	100,0	144

Tabelle 17: Subjektive Bewertung von MES und FUL bei der Formulierung der Anfrage

Tabelle 17 zeigt, daß insgesamt ca. zwei Drittel der Bewertungen im Bereich ”sehr gut” und ”gut” liegen. Bei einem Drittel der Recherchen kamen die Vpn nur unzureichend oder gar nicht zurecht. Die Unterschiede zwischen MES und FUL sind bei dieser Frage nicht bedeutend. Lediglich bei der Antwort ”gar nicht” werden 4 MES-Recherchen gegenüber nur 1 FUL-Recherche genannt. Es handelt sich bei dieser Kategorie mit 3,5% aller Recherchen jedoch um Ausnahmefälle.

Dieses Ergebnis bedeutet, daß die Vpn in ihrer persönliche Bewertung der beiden Systeme dem tatsächlichen signifikanten Recall-Vorsprung von MES gegenüber FUL nicht Rechnung getragen haben. Dies kann man auch auf der Ebene der einzelnen Suchthemen nachweisen.

Suchthema	Bewertung			Recall		
	MES	FUL	insges.	MES	FUL	insges.
Kriminalität	3	3	**3**	2	1	**1**
Antisemitismus	4	2	**2**	2	6	**4**
Lean Production	4	1	**1**	4	5	**3**
Computer	1	5	**5**	1	3	**1**
Gewalt	2	4	**3**	5	2	**5**
Armut	6	5	**6**	6	3	**6**

Rang 1: Bester Wert

Tabelle 18: Subjektive Bewertung von MES und FUL im Vergleich mit dem Recall

In der Tabelle 18 werden zwei Rangfolgen der Suchthemen gebildet und nach MES und FUL differenziert: Zum einen werden die Themen im Hinblick auf ihre subjektive Bearbeitbarkeit geordnet. Danach kamen die Vpn insgesamt am besten mit dem Thema ”Lean Production” zurecht, am schlechtesten mit dem Thema ”Armut”. Dabei gibt es Unterschiede zwischen den FUL- und den MES-Recherchen. Vergleicht man die Rangreihe der subjektiven Bearbeitbarkeit mit der Rangreihe des tatsächlich erzielten Recall, zeigen sich in der Summe etliche Differenzen: Bei der Bewertung liegt ”Lean Production” an erster Stelle, beim Recall nur an dritter, während beispielsweise das als recht schwierig eingeschätzte Thema ”Computer” (5.Stelle) den insgesamt besten Recall aufweist. Diese Unterschiede sind bei den FUL-Recherchen noch gravierender: Das am besten eingeschätzte Thema ”Lean Production” weist den zweitschlechtesten Recall auf und das am zweitbesten eingeschätzte Thema ”Antisemitismus” mit Abstand den schlechtesten Recall.

Das bedeutet, daß vor allem bei den FUL-Recherchen die subjektive Einschätzung der Bearbeitbarkeit und der tatsächlich erzielte Recall weit auseinanderliegen.

Die subjektiven Präferenzen für FUL kommen schließlich in der summarischen Abschlußfrage noch eimal zum Ausdruck. Die Vpn wurden gefragt: ”Mit welchem System sind Sie am besten zurechtgekommen?”

17 Personen (70,8%) votierten für FUL, 7 Personen (29,2%) für MES.

8 Schlußbemerkung

Krause/Mutschke gehen in ihrem Papier ”Indexierung und Fulcrum-Evaluierung” davon aus, daß zukünftige leistungsfähige Systeme zur Inhaltserschließung Mischformen von intellektueller und automatischer Indexierung sein könnten. ”Die als erster Schritt zu testende Alternative ist somit die Mischform aus Freitextrecherche und intellektueller Indexierung im Kontext eines Bool`schen Retrievalsystems, d.h. Bool`sche Suche mit Deskriptoren und Freitextbegriffen, versus der automatischen Indexierung im Kontext eines statistischen Retrievalsystems.”[16] Als Modell für eine automatische Indexierung schlagen sie einen quantitativ-statistischen Ansatz vor, bei dem die Einbettung in das Bool`sche Retrieval durch Best-Match-Verfahren ersetzt wird.

Die vorliegende Vergleichsuntersuchung MES-FUL gehört in den Kontext dieses Untersuchungsprogramms. Die Ergebnisse zeigen, daß im Rahmen des quantitativ-statistischen Ansatzes, wie er mit FUL realisiert worden ist, das Bool`sche exact-match-Retrieval dem vector-space-Modell (Best-Match-Verfahren) vorgezogen worden ist. Die in MES realisierte Mischform aus intellektueller und automatischer Indexierung erwies sich gegenüber dem quantitativ-statistischen Ansatz mit Bool`schem exact-match-Retrieval in FUL beim Recall als überlegen.

9 Literatur

Backhaus, K. u.a.: Multivariate Analysemethoden. Eine anwendungsorientierte Einführung, Berlin/ Heidelberg, 1989

Frisch, Elisabeth; Kluck, Michael: Pretest zum Projekt German Indexing and Retrieval Testdatabase (GIRT) unter Anwendung der Retrievalsysteme Messenger und freeWAISsf, IZ-Arbeitsbericht Nr. 10, Bonn, 2.Auflage, Oktober 1997.

Krause, Jürgen; Stempfhuber, Max; Mandl, Thomas: Das Verbandsinformationssystem ELVIRA II, Projektskizze, ELVIRA Arbeitsbericht 12, Informationszentrum Sozialwissenschaften, Bonn, 1997

Krause, Jürgen; Schaefer, André: Textrecherche–Oberfläche in ELVIRA II, ELVIRA-Arbeitsbericht 16, Bonn, 1998

Krause, Jürgen; Mutschke, Peter: Indexierung und Fulcrum-Evaluierung, IZ-Arbeitsbericht Nr.17, Bonn, Mai 1999

[16] Vgl. Krause/Mutschke, a.a.O., S.16

Kromrey, Helmut: Empirische Sozialforschung, 7.Auflage, Opladen, 1995

Schaefer, André: Benutzertests zur Textretrievalkomponente für ELVIRA II, ELVIRA-Arbeitsbericht 20, IZ-Bonn, Juni 1999

Womser-Hacker, Christa: Der PADOK-Retrievaltest, Hildesheim/Zürich/New York, 1989

10 Anhang

Anlage 1: Anschreiben an die Vpn

GESIS

InformationsZentrum Sozialwissenschaften

Projekt MESSENGER - FULCRUM; Benutzertest zu Datenbankrecherchen

Sehr geehrte

vielen Dank für Ihre Bereitschaft, an unserem Benutzertest zu Datenbankrecherchen mit den Suchsystemen MESSENGER und FULCRUM teilzunehmen. Wie telefonisch vereinbart, kommen Sie zur Durchführung des Tests am xxx um xxx Uhr in das Informationszentrum Sozialwissenschaften (IZ) Bonn, Lennéstr.30.

Da das IZ nur 6 Gehminuten vom Hauptbahnhof Bonn entfernt ist und in der Lennéstraße kaum Parkmöglichkeiten vorhanden sind, empfehlen wir die Anreise mit öffentlichen Verkehrsmitteln. Zu Ihrer Orientierung fügen wir die Kopie eines Stadtplanausschnitts bei.

Wir gehen davon aus, daß Sie mit Datenbankrecherchen im einzelnen noch nicht vertraut sind. Wir haben deshalb am Tag der Testdurchführung für Sie eine Einzelschulung vorgesehen, die Sie in die Lage versetzen soll, den Test kompetent durchzuführen. Zur Einstimmung auf diese Schulung fügen wir diesem Schreiben Unterlagen bei, aus denen Sie Einzelheiten zur Anlage und zur Durchführung des Tests entnehmen können.

Für Rückfragen stehen wir gerne telefonisch unter 0228/2281-174 (Matthias Stahl) oder 0228/2281-166 (Gisbert Binder) zur Verfügung.

Wir freuen uns auf Ihre Teilnahme und verbleiben

mit freundlichen Grüßen

Anlage

Informationen zum Benutzertest: Datenbankrecherchen mit den Retrieval- bzw. Suchsystemen MESSENGER und FULCRUM

Zielsetzung

Will man in umfangreichen Datenbanken Informationen gezielt und schnell auffinden, ist man auf den Einsatz spezieller Such- bzw. Retrievalsoftware angewiesen. Die Suchsoftware vergleicht die eingegebenen Begriffe mit den Textinhalten bzw. den vergebenen Schlagwörtern der Datenbanken. Sie ist somit in der Lage, die gesuchten Dokumente (DE) bzw. Informationen in den Datenbanken zu identifizieren und als Rechercheergebnis auszuweisen.

Solange nur ein einzelner Suchbegriff eingegeben werden muß, ist die Suche einfach. Bei der Verwendung von mehreren Suchbegriffen muß man jedoch bei vielen Retrievalsystemen entscheiden, mit welchen logischen Operatoren (sog. Bool´sche Logik: und, oder, nicht) die Begriffe verbunden werden sollen.

Der im IZ geplante Benutzertest hat die Aufgabe, die Leistungsfähigkeit von zwei unterschiedlichen Retrievalsystemen zu untersuchen:

Das System **MESSENGER** verwendet die Bool´sche Logik und greift auch auf intellektuell vergebene Deskriptoren bzw. Schlagwörter zurück.

Das System **FULCRUM** verzichtet dagegen auf Deskriptoren und bietet **zwei Alternativen**: Suche im gesamten Text der DE ohne logische Operatoren und Suche im gesamten Text unter Verwendung der Bool´schen Logik.

Datenbasis und Suchmöglichkeiten

Der Benutzertest wird auf der Grundlage einer eigens für Testzwecke eingerichteten Datenbank durchgeführt. Dabei handelt es sich um Auszüge aus den sozialwissenschaftlichen Datenbanken SOLIS (Sozialwissenschaftliches Literaturinformationssystem) und FORIS (Forschungsinformationssystem Sozialwissenschaften) mit Schwerpunktsetzung auf den Themenbereichen Industrie- und Betriebssoziologie, Frauenforschung sowie Migration und ethnische Minderheiten aus den Jahren 1990 bis 1996. Die Testdatenbank enthält knapp 13.000 Dokumentationseinheiten (DE).

Die Datenbank enthält grundsätzlich DE mit folgenden Suchfeldern, aus denen das Thema der Publikation (SOLIS) bzw. des Forschungsprojekts (FORIS) entnommen werden kann: Titel und Untertitel sowie Kurzreferat bzw. Projektbeschreibung, normalerweise in deutsch, bei SOLIS-DE teilweise zusätzlich auch in englisch.

Wird beispielsweise der Suchbegriff „Landfrau“ bei **FULCRUM** eingegeben, werden alle DE aus der Datenbank angezeigt, bei denen dieser Begriff im Titel oder im Untertitel oder im Kurzreferat bzw. der Projektbeschreibung zumindest einmal vorkommt (sog. „Treffer“). Findet FULCRUM mehrere Treffer, was in der Regel der Fall ist, werden die gefundenen DE nach ihrer Relevanz für das Thema geordnet, d.h. bei der Ergebnisanzeige stehen die wichtigsten DE an vorderer Stelle.

Eine typische DE, die **mit FULCRUM** gefunden werden kann, haben wir Ihnen als

Anlage 1 beigefügt.

MESSENGER greift auf dieselben Suchfelder und **zusätzlich** auf intellektuell vergebene Deskriptoren zurück. Dabei sind die beim Datenbankaufbau vergebenen inhaltskennzeichnenden Schlagwörter von besonderer Bedeutung: Beispielsweise findet man bei der Suche mit der Wortkombination „Landbevölkerung **und** Frau“ weitere DE, bei denen in den Textfeldern das Wort „Landfrau“ wohl nicht explizit vorkommt, die sich jedoch inhaltlich auf die weibliche Landbevölkerung beziehen. Es ist allerdings erforderlich, daß bei der Suche die vergebenen bzw. zugelassenen Schlagwörter und keine anderen verwendet werden. Dazu liegt ein bei der Suche aufrufbares maschinenlesbares Wörterbuch vor.

Ob in den Textfeldern, in den Deskriptorenfeldern oder in beiden recherchiert werden soll, hängt von der Fragestellung ab und bedarf in jedem Einzelfall der Erprobung.

Eine typische DE, die **mit MESSENGER** gefunden werden kann, haben wir Ihnen als Anlage 2 beigefügt.

MESSENGER gibt alle DE als Treffer aus, die in den Textfeldern oder in den Deskriptorenfeldern oder in beiden den eingegebenen Suchbegriff bzw. die Kombination von Suchbegriffen zumindest einmal enthalten. Eine Anordnung des Suchergebnisses nach Relevanz ist nicht vorgesehen.

Durchführung des Benutzertests

Es stellt sich nun die Frage, welches Suchsystem bzw. welche Recherchealternative zu besseren Ergebnissen führt. Dazu wurde im IZ eine experimentelle Versuchsanordnung entwickelt, die den direkten Vergleich beider Systeme gestatten soll. Dabei werden Testpersonen gebeten, insgesamt sechs Recherchethemen zu bearbeiten, zu einer Hälfte mit FULCRUM, zur anderen mit MESSENGER. Es geht dabei darum, die Anfragen innerhalb jedes Suchsystems so zu formulieren, daß möglichst alle in der Datenbank vorhandenen relevanten DE und gleichzeitig jedoch möglichst wenige

irrelevante DE im Rechercheergebnis angezeigt werden. Die Resultate der Testpersonen werden am Ende der Versuchsreihe zusammengefaßt und statistisch analysiert. Neben der Gesamtbewertung der Suchsysteme geht es auch um die Frage, ob und ggf. wie die vorhandenen Rechercheoberflächen im Sinne einer besseren Benutzerfreundlichkeit verändert werden sollen.

Anlage 2: Fragen im Rahmen des GIRT - Nachfolgeprojekts

Title: Kriminalität bei Frauen

Description: Welche Berichte, Fälle, empirische Untersuchungen und Analysen gibt es zur Kriminalität und Delinquenz bei Frauen?

Narrative: Relevante Dokumente befassen sich mit speziellen Problemen der Frauenkriminalität einschließlich der Probleme der Resozialisierung und des Strafvollzugs bei Frauen. Nicht relevant sind historische Untersuchungen (vor 1945), Jugendliche und Kinder (vor allem Mädchen), allgemeine Statistiken, rechtsphilosophische Betrachtungen, Terrorismus.

Title: Antisemitismus in Deutschland nach 1945

Description: Welche Berichte, Fälle, empirische Untersuchungen und Analysen gibt es zum Antisemitismus in Deutschland nach 1945?

Narrative: Relevante Dokumente befassen sich mit dem Antisemitismus in Deutschland (BRD und DDR) nach 1945. Dazu gehören sowohl antisemitische Aktionen und Aktivitäten als auch ideologische Aspekte des Antisemitismus. Nicht relevant sind historische Untersuchungen (vor 1945).

Title: Lean - production in Japan

Description: Welche Berichte, Fälle, empirische Untersuchungen und Analysen gibt es zur Einführung und Anwendung von Lean - production - Konzepten in Japan?

Narrative: Relevante Dokumente befassen sich mit den speziellen Problemen der Lean - production in Japan. Dazu gehören sowohl allgemeine Aussagen zum Managementkonzept der Lean - production als auch konkrete Anwendungsfälle und Untersuchungen. Nicht relevant sind generelle Berichte zur Lean - production oder spezifische Studien über andere Länder.

Titel: Computer im Alltag

Description: Welche Berichte und Analysen gibt es zur Verwendung von Computern im Alltag?

Narrative: Relevante Dokumente befassen sich mit der alltäglichen Nutzung von Computern. Nicht relevant sind Dokumente zur Nutzung von Computern im Beruf, zu Unterricht und Ausbildung an Computern und zu allgemeinen Fragen des Technikeinsatzes.

Title: Gewaltbereitschaft von Jugendlichen

Description: Welche Berichte, Fälle, empirische Untersuchungen und Analysen gibt es zur Gewaltbereitschaft von Jugendlichen?

Narrative: Relevante Dokumente befassen sich mit den speziellen Problemen der Gewaltbereitschaft von Jugendlichen (Jungen und Mädchen), d.h. mit Gewalt und Gewalttätigkeit, die von Jugendlichen ausgeht. Nicht relevant sind Berichte über Jugendliche als Opfer von Gewalt, über Gewalt und Sexualität und über allgemeine Probleme der Gewalt und Gewaltanwendung.

Title: Armut und Obdachlosigkeit in Städten

Description: Welche Berichte und Analysen gibt es zur Armut, Verelendung und Obdachlosigkeit in Städten?

Narrative: Relevante Dokumente befassen sich mit den Berichten zur Armut und zur Obdachlosigkeit in Städten und Großstädten. Dazu gehören das alltägliche Leben und die generelle soziale Lage in Städten und in bestimmten Stadtvierteln oder Quartieren (z.B. Slums). Nicht relevant sind allgemeine Untersuchungen zur Stadtstruktur.

Anlage 3

GESIS

IZ InformationsZentrum Sozialwissenschaften

Fragebogen zum Recherchetest

Testverlauf

Nr.:	Datum:	Uhrzeit: von bis

Studium/Arbeitsgebiet

Studienfach:	Semester:
berufl. Tätigkeit:	

Vorkenntnisse

Haben Sie Erfahrungen mit Recherchen in elektronischen Datenbanken?	Ja ☐ Nein ☐
Wenn ja, mit welchem System?	. . . . .
Haben Sie schon im Internet recherchiert?	Ja ☐ Nein ☐
Haben Sie schon ***Boolesche Operatoren*** (und/oder/nicht) bei vorherigen Recherchen verwendet?	Ja ☐ Nein ☐

Zum Ablauf des Tests

Waren der einführende Text und die personelle Unterstützung verständlich?

☐ sehr gut verständlich ☐ gut verständlich ☐ wenig verständlich ☐ ganz unverständlich

Ist es Ihnen gelungen, die vorgegebenen Fragen in **Rechercheanfragen** umzusetzen?

☐ sehr gut gelungen ☐ gut gelungen ☐ unzureichend gelungen ☐ gar nicht gelungen

Wie gut kamen Sie bei der Formulierung der Suchthemen mit den verschiedenen Systemen zurecht?

Bitte verwenden Sie in der nachfolgenden Tabelle die Ziffern (1 - 4) um Ihre jeweiligen Einschätzungen wiederzugeben (*sehr gut(1), gut(2), unzureichend(3), gar nicht(4)):*

System:	**FULCRUM:**	**MESSENGER:**
Frage:	(Einschätzung 1 - 4)	(Einschätzung 1 - 4)
Frauenkriminalität		
Antisemitismus		
Lean-Production		
Computer im Alltag		
Gewalt bei Jugendlichen		
Armut		

(bitte füllen Sie die für Sie zutreffenden Recherchefelder aus)

Mit welchem System sind Sie am besten zurechtgekommen?

☐ MESSENGER ☐ FULCRUM

Welchen Eindruck hatten Sie insgesamt vom Recherchetest und den getesteten Systemen?