

GLIEDERUNG

0	Einleitung	1
1	Grundlagen, Testaufbau und Durchführung von PADOK (J. Krause)	3
1.1	Ausgangslage	4
1.2	PADOK-Evaluierungsplan und seine Realisierung	7
1.2.1	Evaluierungsmethodologie	7
1.2.2	Evaluierungsverbund	10
1.2.3	Verlauf der Evaluierung	14
1.3	Auswahl der Testdaten	19
1.3.1	Erfaste, noch nicht bearbeitete Dokumente	19
1.3.2	Texterschließungsanalyse und organisatorisches Umfeld vs. Retrieval	20
1.3.3	Auswahlklasse	22
1.3.4	Dokumente mit und ohne Abstract	24
1.3.5	Realisierte Testdatenbank	25
1.3.6	Abfolge der Verarbeitung	25
1.4	Retrievaltests	25
1.4.1	Probanden	26
1.4.2	Aufgaben	27
1.4.2.1	Aufgabenorientierung	27
1.4.2.2	Auswahl der Aufgaben	28
1.4.3	Relevanzbewertung	29
1.4.4	Vorgespräche, Einarbeitung und Diskussion der Ergebnisse	30
1.4.5	Einteilung in Testphasen	31
1.4.6	Besondere Kontextbedingungen	33
2	Beschreibung der Systeme	35
2.1	PASSAT: Programm zur Automatischen Selektion von Stichwörtern aus Texten (V. Gräbnitz, Siemens München)	36
2.1.1	PASSAT als Programmsystem	36
2.1.2	Dokumentdemontage mit dem Programm PASSAU	37
2.1.3	Texterschließung mit dem Programm PASSAT	38

2.1.3.1	Vergleichswortliste (VWL)	39
2.1.3.2	VWL-Eintrag	40
2.1.3.3	Zeichenvorrat	47
2.1.3.4	Textworterkennung	48
2.1.4	Nachbearbeitung der Deskribiererergebnisse	51
2.1.5	Dokumentmontage mit dem Programm PASDOK	51
2.1.6	Texterschließung mit der Programmschnittstelle von PASSAT	53
2.1.7	Sortieren und Sichern der Texterschließungsergebnisse mit dem Programm PASFISH	54
2.1.8	PASSAT-Anwendung beim Satz-Rechen-Zentrum Berlin	54
2.2	Das System CTX - Syntaxanalyse zur Texterschließung (C. Schneider)	56
2.2.1	Leistungen	56
2.2.2	Verfahren	58
2.2.2.1	Textaufbereitung	60
2.2.2.2	Die syntaktische Analyse	61
2.2.2.3	Das morpho-syntaktische Wörterbuch	62
2.2.2.4	Das Derivations- und Zerlegungswörterbuch COMPLX	63
2.2.2.5	Deskriptorerzeugung	64
2.2.3	Der Einsatz von CTX in PADOK	65
2.3	Automatische Erkennung von Nominalgruppen für das Referenzretrieval mit DETECT (G.F. Schulz, GID Frankfurt)	66
2.3.1	Einleitung	66
2.3.2	Beschreibung des Verfahrens	71
2.3.2.1	Automatische Nominalgruppenerkennung	71
2.3.2.1.1	Das Programmsystem DETECT	72
2.3.2.1.1.1	Partielles Parsing und Longest-Match- Verfahren	74
2.3.2.1.1.2	Precompiler-/Interpreterkonzept	76
2.3.2.1.1.3	Tokenmengenbeschreibungen	77
2.3.2.1.1.4	Integration eines frei definierbaren Scanners	78
2.3.2.1.1.5	Iteratives Entwicklungs- und Testkonzept mit integrierten Qualitätsbewertungsverfahren	78

2.3.2.1.2	Analysegrammatiken für die Nominalgruppenerkennung	79
2.3.2.1.2.1	Beispiele für PSG-Regeln von Nominalgruppen-Teilstrukturen in DETECT-Notation	80
2.3.2.1.2.2	Entwicklungs- und Testverfahren	83
2.3.2.2	Designüberlegungen für den Datenbankaufbau	85
2.3.3	Ausblick	88
3	Analyse der Texterschließung (C. Schneider)	90
3.1	Einordnung	90
3.2	Grundlagen der Bewertung	91
3.2.1	Abweichung vs. Fehler	91
3.2.2	Token-Zählung vs. Type-Zählung	92
3.2.3	Fehlerzählung in den einzelnen Analysebereichen	93
3.3	Auswahl der Datengrundlagen	94
3.3.1	Grundformenreduktion	95
3.3.2	Kompositazerlegung	95
3.3.3	Syntaktische Relationen	96
3.3.4	Zusammenfassung und technischer Ablauf	96
3.4	Aufwandsfaktoren im Verlauf der Texterschließung	98
3.4.1	Die Textaufbereitung bei CTX	98
3.4.2	Lexikonerweiterungen in PASSAT und CTX	99
3.5	Einfachdeskriptoren	100
3.5.1	Vorgehen	101
3.5.2	Bewertung	102
3.5.3	Ergebnisse	104
3.6	Zusätzliche Deskriptoren	106
3.6.1	Bewertung bei zusätzlichen Deskriptoren	106
3.6.2	Wortbildung in PASSAT und CTX	107
3.6.2.1	Zielsetzung und Verfahren von PASSAT	108
3.6.2.2	Zielsetzung und Verfahren von CTX	108
3.6.2.3	Datengrundlagen für die Bewertung	109
3.6.2.4	Bewertung	111
3.6.2.5	Ergebnisse	111
3.6.2.5.1	Fehler	112
3.6.2.5.2	Anzahl der Zerlegungselemente	112
3.6.3	Weitere zusätzliche Deskriptoren	113

3.6.3.1	Teilelemente durch unterschiedliche Wortdefinition in PASSAT	113
3.6.3.2	Mehrfachzuordnungen durch PASSAT	114
3.6.4	Fazit	115
3.7	Syntaktische Relationen	116
3.7.1	Zur Vergleichbarkeit von CTX und DETECT	116
3.7.2	Vollständige Syntaxanalyse vs. partielles Parsing (CTX vs. DETECT)	117
3.7.3	CTX	119
3.7.3.1	Zur Definition Komplexer Deskriptoren	119
3.7.3.2	Bewertung Komplexer Deskriptoren	120
3.7.3.3	Ergebnisse	123
3.7.4	DETECT	125
3.7.4.1	Möglichkeiten der Bewertung in DETECT	125
3.7.4.2	Bewertung von DETECT im Vergleich zu CTX	125
3.8	Fazit	130
4	Aufbau der Datenbanken und Retrievalmöglichkeiten von PADOK (L. Hitzenger, G. Rasche)	132
4.1	Aufbau der Datenbanken	132
4.2	Struktureller Aufbau der Datenbanken	132
4.2.1	Generelle Strukturprinzipien	132
4.2.2	Die PASSAT-Datenbank	133
4.2.3	Die CTX-Datenbank	134
4.2.4	Die Freitextdatenbank	138
4.3	Funktionale Beschreibung der wichtigsten GRIPS-Kommandos	138
4.3.1	Vorbemerkung	138
4.3.2	Datenbankauswahl	139
4.3.3	DISPLAY	140
4.3.4	FIND	140
4.3.5	TAB	141
4.3.6	SHOW	141
4.3.7	PRINT	142
4.3.8	INFO COST	142
4.3.9	STOP	142
4.4	Festlegungen für den Ablauf des Tests	142
4.5	Technische Probleme	143

4.5.1	Print-Kommando	143
4.5.2	Abweichen vom Kodierungsstandard	143
4.5.3	Datellenverlust	144
4.5.4	Protokollierung der Tests: Echoeffekt im Screen- Mode	144
4.5.5	Fazit	145
5	Statistische Auswertung des Retrievaltests auf der Grundlage von recall und precision (G. Spettel, C. Womser-Hacker)	146
5.1	Meßtheoretisches Design des Retrievaltests	146
5.2	Statistische Bewertungsmaße	148
5.2.1	recall und precision	149
5.2.2	Einheitsmaß	150
5.2.3	Mittelwertbildung	151
5.2.4	Besonderheiten der Relevanzbestimmung	152
5.2.5	Nullantworten	153
5.2.5.1	Definition einer "Nullantwort"	153
5.2.5.2	Konsequenzen für die statistischen Bewertungs- maße	154
5.3	Technische Abwicklung der Auswertung	155
5.4	Ergebnisse der statistischen Auswertung auf der Grundlage von recall und precision	155
5.4.1	Ergebnisse der Gesamtauswertung	156
5.4.2	Blockbildung nach Prüfergruppen	158
5.4.2.1	Ergebnisse der Prüfer des Deutschen Patentamtes	159
5.4.2.2	Ergebnisse der Industrie-/FIZ-Prüfer	160
5.4.2.3	Fazit: DPA- vs. Industrie-/FIZ-Prüfer	161
5.4.3	Blockbildung nach Testphasen	162
5.4.4	Sonderprobleme bei Gesamtauswertung und Block- bildung	165
5.4.4.1	Die Behandlung der Relevanzstufe "nicht ent- scheidbar"	165
5.4.4.2	Abweichungen bei den Nullantworten in Test- phase I	167
5.4.4.3	Nullantworten	168
5.4.4.3.1	Zahlenmäßiger Anteil der Nullantworten	168

5.4.4.3.2	Bewertungsalternativen für Nullantworten vom Typ A mit precision = 1	169
5.4.4.4	Abweichungen im situativen Kontext	170
5.4.5	Bildung der Überschneidungsmengen bei den relevanten Dokumenten	172
5.4.6	Signifikanztest	173
5.4.6.1	Theoretisches zum Signifikanztest	174
5.4.6.1.1	Testauswahl	174
5.4.6.1.1.1	Die zugrundeliegende Verteilung der Stichprobe	174
5.4.6.1.1.2	Verbundene bzw. abhängige Stichproben	175
5.4.6.1.1.3	Stichprobenumfang	175
5.4.6.1.1.4	Meßskala	176
5.4.6.1.1.5	Ergebnis der Testauswahl	177
5.4.6.1.2	Ablaufbeschreibung des Vorzeichentests	177
5.4.6.2	Ergebnisse des PADOK-Signifikanztests	180
5.4.6.2.1	Signifikanzüberprüfung der Gesamtergebnisse	181
5.4.6.2.2	Signifikanzüberprüfung nach Prüfergruppen	182
5.4.7	Fazit	183
6	Analytische Auswertung der Retrievalprotokolle: Trunkierung, Komplexe Deskriptoren und Ankreuztechnik (G. Bauer, T. Engl, B. Hantscher, J. Krause, B. Kudlich, R. Schlape)	185
6.1	Aufgabenorientierung vs. Einzelanfragen	185
6.2	Trunkierung	187
6.2.1	Trunkierungsarten beim Find-Befehl	187
6.2.2	Trunkierungshäufigkeit	189
6.2.3	Fazit: Trunkierung	189
6.3	Komplexe Deskriptoren und Ankreuztechnik	189
6.3.1	Direkte Verwendung komplexer Deskriptoren	190
6.3.2	Ankreuztechnik	192
6.3.3	Komplexe Deskriptoren und recall	194
6.4	Deskriptoreinführung ohne Ankreuzen	196
6.5	Fazit	197

7	Aufwandsabschätzungen und Organisation der Inhaltserschließungsläufe (J. Krause)	198
7.1	Organisatorische Voraussetzungen	199
7.2	Inhaltserschließung	200
7.2.1	CPU-Zeiten	200
7.2.2	Zeiten für die manuellen/intellektuellen Eingriffe von PASSAT und CTX	203
7.2.3	Personaleinsatz	204
7.2.4	Speicherplatzbedarf	204
7.3	Aufbau der Datenbanken	205
7.4	Retrievalzeiten	205
7.5	Fazit	206
8	Problemfeld Patenterteilung und derzeitige Informationsbeschaffung (J. Krause)	208
8.1	Problemlösungsprozeß Patenterteilung	209
8.1.1	Das Patent	209
8.1.2	Verfahrensschritte der Patenterteilung	209
8.1.3	Patentfähigkeit	212
8.1.3.1	Neuheit	213
8.1.3.2	Erfindungshöhe	214
8.1.4	Die Patenterteilung als argumentative Kommunikationssituation	215
8.1.4.1	Monologischer Diskurs	215
8.1.4.2	Die Patentschrift als juristischer Diskurs	216
8.1.4.2.1	Grundschema	216
8.1.4.2.2	Das Argumentationsschema von Toulmin (1958)	217
8.1.4.2.3	Rechtfertigung der Prämissen	218
8.1.4.3	Argumentationsziele	219
8.1.5	Aufbau der Patentschrift	220
8.1.6	Sprachliche Ausdrucksmittel und Funktion der Zeichnungen	222
8.1.6.1	Abstraktionsebene und Konkretisierung	223
8.1.6.2	Verschleiende Wortwahl	224
8.1.6.3	Bedeutungselemente der Textgrundlage	225
8.1.6.4	Neubildung von Komposita	225
8.1.6.5	Komplexe Nominalphrasen	226
8.1.6.6	Zeichnungen	226

8.2	Informationsbeschaffung heute	227
8.2.1	IPC-Schlüssel	227
8.2.2	Materielle Grundlage der Recherche	228
8.2.3	Interne maschinelle Dokumentationssysteme	229
8.2.4	Externe Datenbanken	230
8.2.5	Vorgehen bei der Recherche	230
8.2.6	Grundparameter DPI	232
9	Zusammenfassende Bewertung der PADOK-Teilergebnisse (J. Krause)	234
9.1	Übertragbarkeit und Eigengesetzlichkeiten des Unter- suchungsbereiches	234
9.2	Varianten der Inhaltserschließung: Freitext, PASSAT, CTX	235
9.2.1	Freitext	235
9.2.2	Morphologie und Syntax: PASSAT vs. CTX	236
9.3	Alternative Ansätze ohne größere Wörterbücher und ohne umfangreiche Syntaxanalyse	238
9.4	Retrievalhilfen	239
9.5	Fazit	240
	Abkürzungsverzeichnis	241
	Literatur	242
	An PADOK beteiligte Institutionen	248