

**PETRUS-Workshop "Automatische
Erschließungsverfahren" 21./22.03.2011**

Dipl.-Psych. Michael Gerards

Semiautomatische Erschließung von Psychologie-Information

Die Literaturdatenbank PSYINDEX:

- Erschließt die im Bereich der Psychologie veröffentlichte Literatur aus deutschsprachigen Ländern
- Umfasst z. Z. fast 249.000 Dokumente
- Enthält neben den bibliographischen Angaben für fast alle Dokumente Abstracts und eine umfangreiche inhaltliche Erschließung
- Teil der inhaltlichen Erschließung: Vergabe von Deskriptoren auf der Grundlage des "Thesaurus of Psychological Index Terms" der "American Psychological Association"

Aufgabe für die automatische Indexierung:

Generierung von Deskriptorvorschlägen aus den Dokumenten
(Titel, Abstract, Autorenschlagworte)

Ziel der automatischen Indexierung:

Unterstützung der intellektuellen Verschlagwortung -> semi-
automatische Indexierung

dazu seit 2006 eingesetztes Verfahren:

AUTINDEX (Automatic Indexing) vom "Institut der Gesellschaft
zur Förderung der Angewandten Informationsforschung e.V." (IAI)

Vorgehen von AUTINDEX:

Einsatz linguistischer Intelligenz bei der wörterbuch- und regelbasierten Textanalyse kombiniert mit statistischen Elementen

- AUTINDEX arbeitet "verstehensbasiert" und nicht "string-basiert"
- "Ausreizung" der linguistischen Möglichkeiten, um statistische Verfahren für die Indexierung optimal zu unterstützen
- Die Analyse erfolgt in mehreren Schritten

Satz- und Wortformerkennung, Erkennung von morphologischen, syntaktischen und semantischen Varianten, Kompositazerlegung, Ausschluss von Stoppwörtern

Sammlung der bedeutungstragenden Elemente und Ermittlung der am häufigsten auftretenden semantischen Klassen

Identifikation von Nominalphrasen, Bestimmung des Subjekts und finiten Verb eines Satzes

Linguistische Analyse
MPro



Evaluierung der Textelemente



Oberflächen-Parsing



Thesaurus



Abgleich



Ausgabe der Deskriptorvorschläge

Schwellenwert überschritten?



Ergebnisausgabe

Desambiguierte syntaktische Repräsentation des Textes

+

Gewichtung der Terme anhand von Termfrequenz, Stellung des Terms im Text, Häufigkeit der semantischen Klasse des Terms

Voraussetzung für die linguistische Analyse:

Fachsprache muss in das AUTINDEX-Wörterbuch integriert sein

- Einbau der ca. 6.000 Deskriptorsterme und der im Thesaurus enthaltenen ca. 1.340 Synonyme
da unzureichend
- Entwicklung des Indikatorkonzeptes: Einbau von Begriffen, die in enger Beziehung zu den Deskriptoren stehen, ohne als direkte Synonyme zu gelten (ca. 18.600 Begriffe)

1. Erweiterung des Thesaurus durch Indikatoren

Bearbeitungsmaske für Indikatoren; Beispiel: Rechenschwäche (Acalculia)

Input Form: Deskriptoren bearbeiten - Record No. = 41

Öffnen Weiter Zurück Eintrag Einf. Eintrag Löschen Abbrechen Speichern Hilfe

Indikatoren für die automatische Indexierung

Indikatoren	TI	AB	UT
Rechenschwäche	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Rechenunfähigkeit	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Akalkulie	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Akalkulia	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Dyskalkulie	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Dyskalkulia	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Rechenstörung	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Mathematikschwäche	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
rechenschwach	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
verzögerter Rechenerwerb	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
gestörte Rechenfähigkeit	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Übern. ? (j=Ja; n=Nein)

Neue Indikatorvorschläge

Bearbeitungsstatus: unbearbeitet unvollständig noch validieren erledigt

Indikatorvorschläge (Archiv)

Deskriptor (deutsch): Rechenschwäche Deskriptor (englisch): Acalculia

Thesaurus Synonyme: Rechenunfähigkeit Eingeführt: 1973

Scope Note: Form of aphasia involving impaired ability to perform simple arithmetic calculations.

Broader Terms: Aphasie

Narrower Terms:

Related Terms: Lernbehinderungen

Bemerkungen:

Neue Übersetzungsvorschläge:

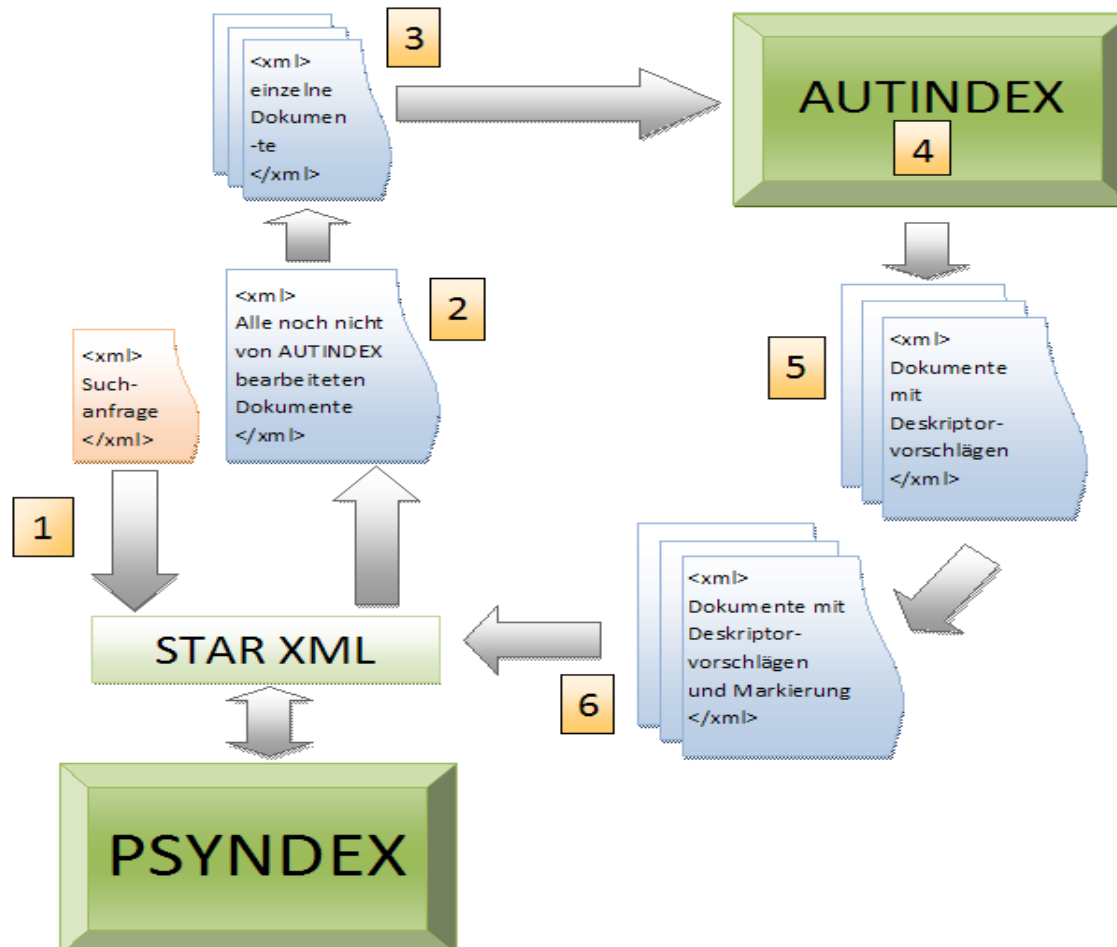
2. Überarbeitung des Thesaurus

- Einschränkung auf Felder, in denen der Thesaurus-/Indikatorbegriff vorkommen muss

Indikatoren	TI	AB	UT
Fragebögen	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Fragebogen	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

- Ausschluss ungeeigneter Thesaurusbegriffe (diese Begriffe kommen auch in der Alltagssprache oder zu häufig in der Fachsprache vor)
- Bei nicht trennscharfen Thesaurusbegriffen:
Normierung
oder
Zuweisung der Indikatoren zu beiden Begriffen

3. Schaffung der Indexierungsperipherie



4. Integration in die PSYNDEX-Erfassung

PSYNDEX Erfassungsmaske: Deskriptorvorschläge übernehmen

STAR Client - zpidu9/gerardsa

File Edit Connection Form Fields Options Window Help

Input Form: Inhaltserfassung - Record No. = 1966

1 Abstract 2 Inhalt 3 Controlled Term 4 Titelübers. 5 Nebenabstr. 6 Tests 7 Fehler

D196924: Krämer, Kai, 2006, 195-221, Resignation im Rahmen der Erwerbsarbeit - Bestandsaufnahme zu einem psychologischen Konzept

Controlled Term (enql oder qerm)	Controlled Term (Übersetzung)	Gewichtet?
Employee Attitudes	Arbeitnehmereinstellungen	<input checked="" type="checkbox"/> gew.
Job Satisfaction	Arbeitszufriedenheit	<input checked="" type="checkbox"/> gew.
Employee Motivation	Arbeitnehmermotivation	<input checked="" type="checkbox"/> gew.
Occupational Stress	Beruflicher Stress	<input checked="" type="checkbox"/> gew.
Job Involvement	Berufliches Engagement	<input type="checkbox"/> gew.
Health Complaints	Gesundheitliche Beschwerden	<input type="checkbox"/> gew.
Somatoform Disorders	Somatoforme Störungen	<input type="checkbox"/> gew.
Prevention	Prävention	<input type="checkbox"/> gew.

AUTINDEX-CT-Vorschläge	übern?	
Working Conditions	Arbeitsbedingungen	<input type="checkbox"/>
Health Complaints	Gesundheitliche Beschwerde	4
Somatoform Disorders	Somatoforme Störungen	3
Employee Motivation	Arbeitnehmermotivation	1
Job Involvement	Berufliches Engagement	2
Prevention	Prävention	5
Psychological Terminology	Psychologische Terminolo	<input type="checkbox"/>
Stress	Stress	<input type="checkbox"/>

Zusatzdeskript. deutsch

Phrase

resignation at worksite, development & health-related consequences & prevention & intervention, occupational stress & control of working conditions & job satisfaction & burnout & psychosomatic complaints, model of resignation development & suggestion of prevention &

AUTINDEX durchführen bzw. wiederholen
 nein ja Inhaltliche Erfassung unvollständig

Uncontrolled Terms (engl)

Uncontrolled Terms (germ)

Start Mozilla Microsoft Excel Microsoft Pow... Referat AUTI... AutindexShot... STAR Client ~... DE 15:26

Datenbasis: 15.521 PSYINDEX-Dokumente verschiedener Literaturgattungen aus den Jahren 2006-2010

Die Dokumente wurden von insgesamt neun Humanindexierern mit Unterstützung durch AUTINDEX inhaltlich erschlossen

Zentrale Fragen:

- Unterscheiden sich intellektuelle und automatische Indexierung in der Indexierungsbreite?
- Wie viele Deskriptorenvorschläge von AUTINDEX werden übernommen?

Indexierungsbreite und Übereinstimmungen für ein "durchschnittliches" PSYINDEX-Dokument (Ausgangsbasis: 15.521 Dokumente)

Anzahl der vom Auswerter vergebenen Deskriptoren	7,5 (43 % übernommen)
Anzahl der von AUTINDEX vorgeschlagenen Deskriptoren	9,2 (65 % "unnötig")
Anzahl der Deskriptoren, die sowohl vorgeschlagen als auch vergeben wurden	3,2
Indexierungskonsistenz (nach L. Rolling)	38,3 %
Für 4,5 % der Dokumente werden keine Vorschläge generiert	

Indexierungsbreite und Übereinstimmungen für unterschiedliche Literaturgattungen in PSYNDEX

	Zeitschriften- aufsätze (11.504 D.)	Sammelwerks- beiträge (851 D.)	Bücher/ Sammelwerke (1.987 D.)
intellektuell	7,5 (44 % übern.)	6,3 (46 % übern.)	6,9 (39 % übern.)
automatisch	8,5 (61 % unnötig)	7 (59 % unnötig)	13,9 (81 % unnötig)
Übereinstimmung	3,3	2,9	2,7
Indexierungs- konsistenz (Rolling)	41,5 %	43,3 %	26 %

Weitere Tendenzen:

- Nach Einführung von AUTINDEX werden etwas mehr Deskriptoren vergeben (vorher: 6,7, nachher: 7,5)
- Der erfahrene Humanindexierer vergibt spezifischere Deskriptoren als AUTINDEX
- Neue Mitarbeiter neigen dazu, mehr Deskriptoren von AUTINDEX zu übernehmen (Problem: hohe Konsistenz bedeutet nicht notwendigerweise hohe Indexierungsqualität)

ZPID-interne Umfrage:

v.a. im Zeitschriftenbereich hohe Zufriedenheit der Mitarbeiter mit AUTINDEX

Vielen Dank für Ihre Aufmerksamkeit.

Michael Gerards

