Psychologie Information
ZPID – Leibniz Institut

PETRUS-Workshop "Automatische Erschließungsverfahren" 21./22.03.2011

Dipl.-Psych. Michael Gerards

Semiautomatische Erschließung von Psychologie-Information

Kontext

Die Literaturdatenbank PSYNDEX:

- 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 APA

Zielsetzung

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)

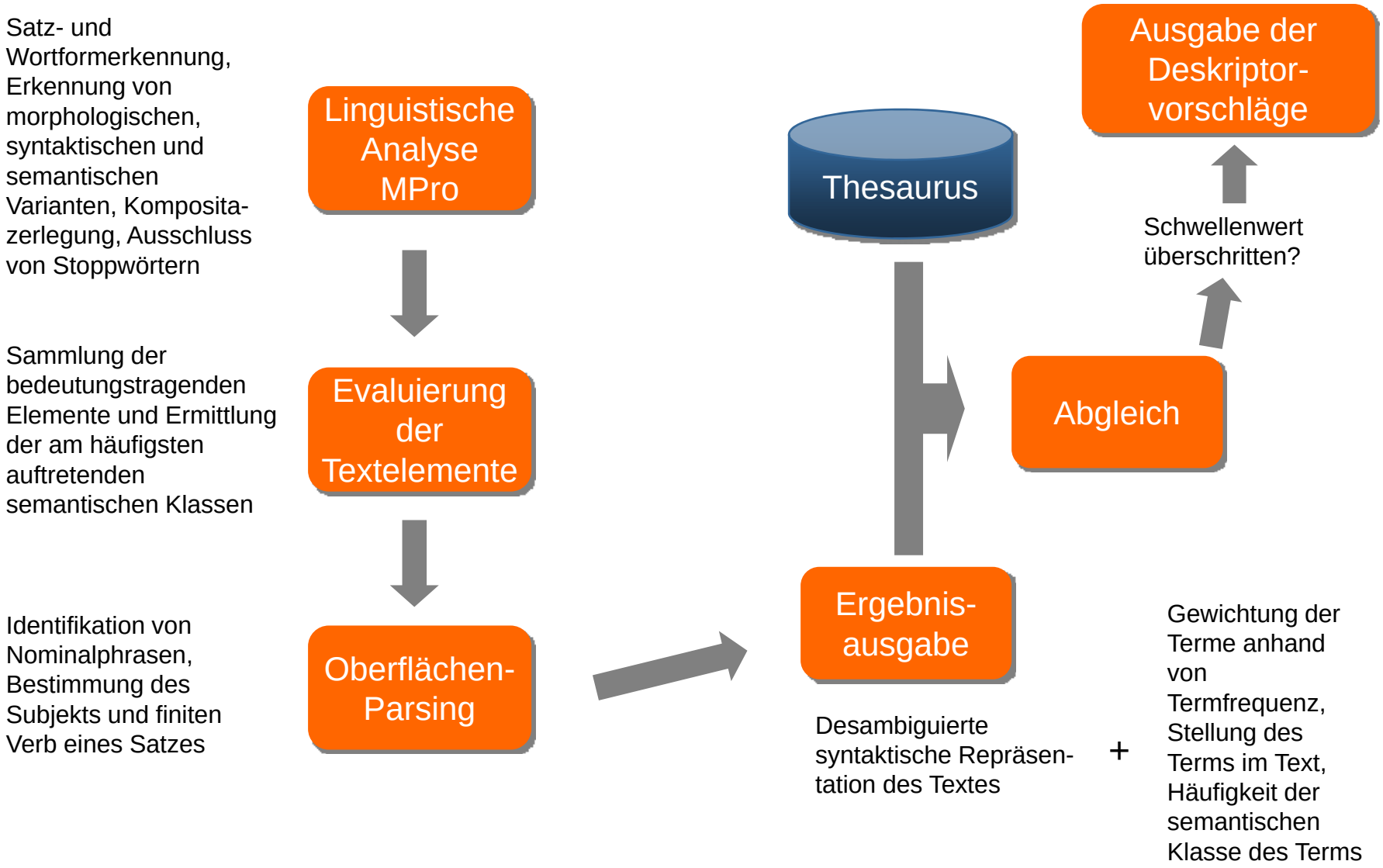
Arbeitsweise von AUTINDEX

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

AUTINDEX-Textanalyse



Voraussetzungen

Voraussetzung für die linguistische Analyse:

Fachsprache muss in das AUTINDEX-Wörterbuch integriert sein

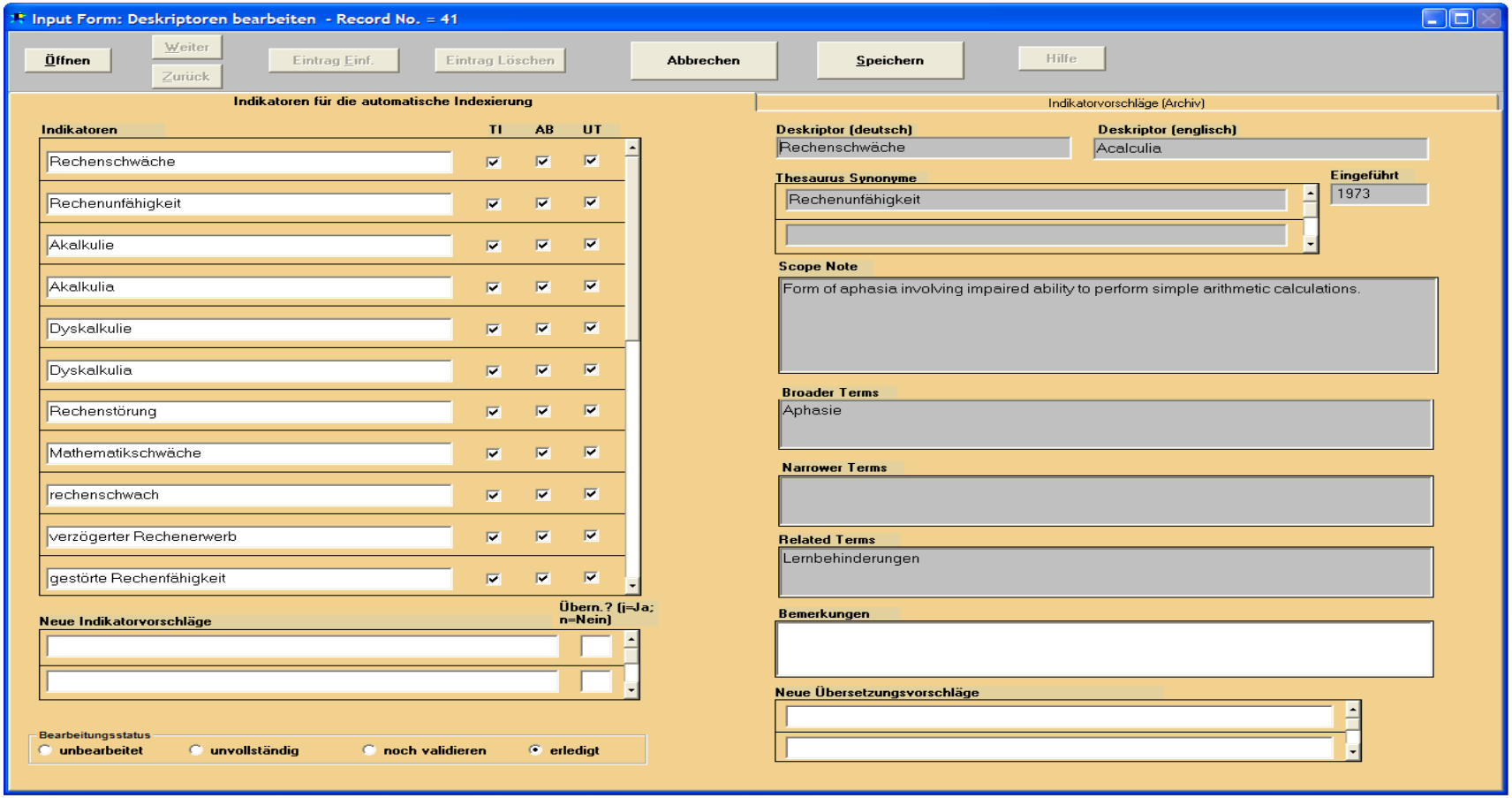
- Einbau der ca. 6.000 Deskriptorterme und der im Thesaurus enthaltenen ca. 1.340 Synonyme

  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)

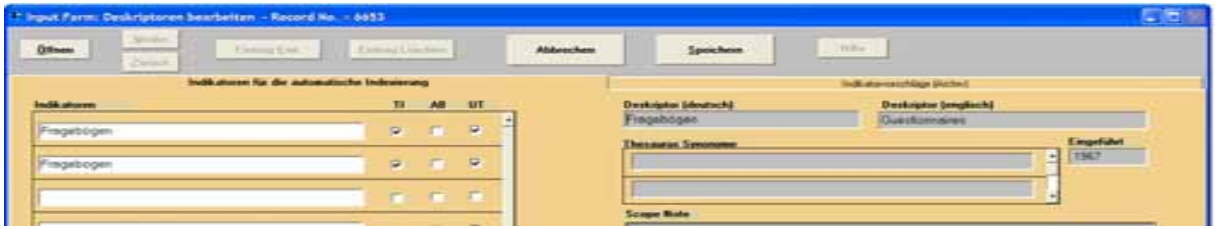
1. Erweiterung des Thesaurus durch Indikatoren

Bearbeitungsmaske für Indikatoren; Beispiel: Rechenschwäche (Acalculia)



2. Überarbeitung des Thesaurus

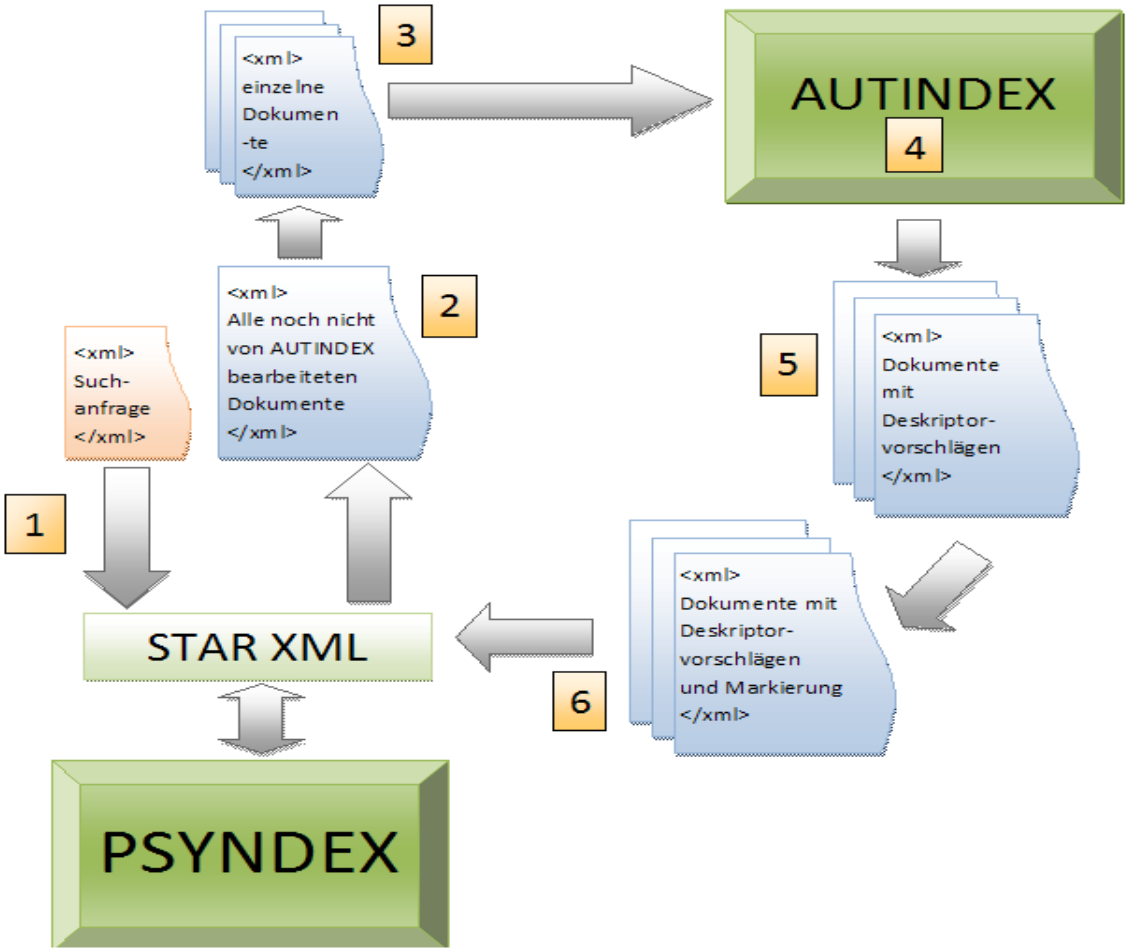
- Einschränkung auf Felder, in denen der Thesaurus-/Indikatorbegriff vorkommen muss



•Ausschluss ungeeigneter Thesaurusbegriffe (kommen auch in der Alltagssprache oder zu häufig in der Fachsprache vor)

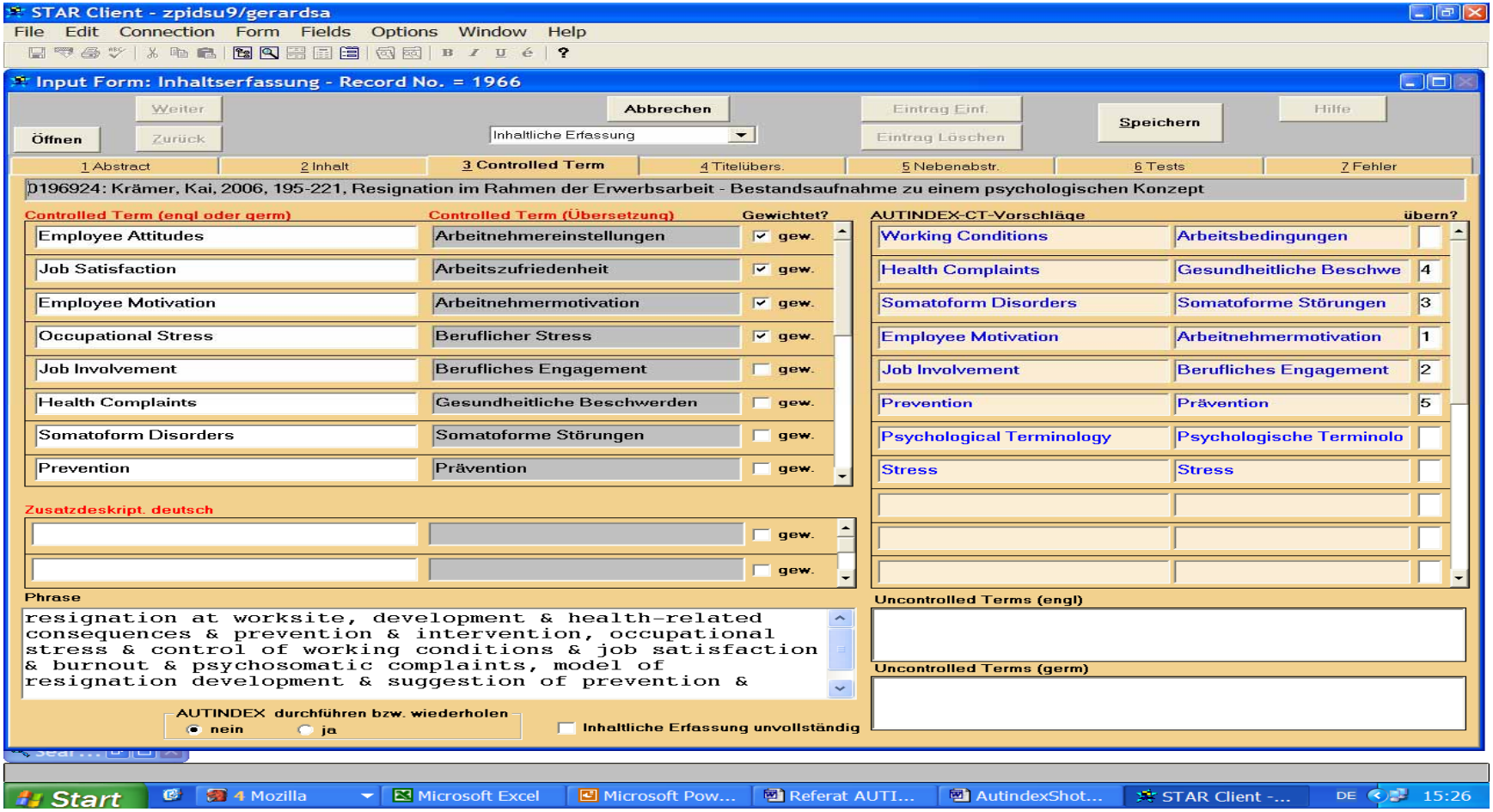
•Bei nicht trennscharfen Thesaurusbegriffen:
Normierung
oder
Zuweisung der Indikatoren zu beiden Begriffen

3. Schaffung der Indexierungsperipherie



4. Integration in die PSYNDEX-Erfassung

PSYNDEX Erfassungsmaske: Deskriptorvorschläge übernehmen



Evaluation 2011

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

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

Zentrale Fragen:

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

Evaluation 2011

Indexierungsbreite und Übereinstimmungen für ein "durchschnittliches" PSYNDEX-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	

Evaluation 2011

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 %

Evaluation 2011

Weitere Tendenzen:

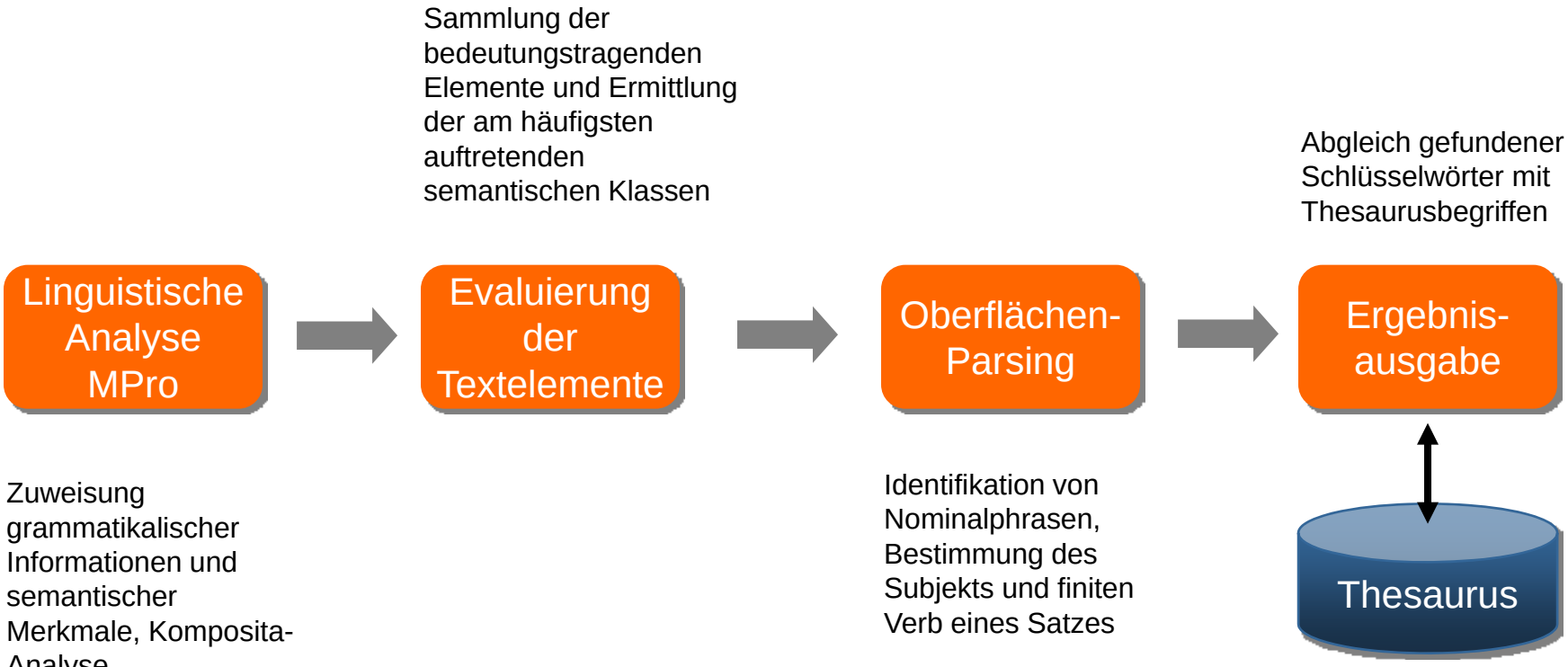
- Nach Einführung von AUTINDEX werden etwas mehr Deskriptoren vergeben (vorher: 6,7, nachher: 7,5)
- Der erfahrene Humanindexierer vergibt spezifischere Deskriptoren als AUTINDEX
- Neue Mitarbeiter neigen dazu, mehr Deskriptoren von AUTINDEX zu übernehmen (Problem: hohe Konsistenz bedeutet nicht notwendigerweise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

ZPID-interne Umfrage:
v.a. im Zeitschriftenbereich hohe Zufriedenheit der Mitarbeiter mit AUTINDEX

Vielen Dank für Ihre Aufmerksamkeit.

Michael Gerards

AUTINDEX-Textanalyse



Evaluationsstudien im Vergleich

Studie*	Konsistenz nach Rolling
Automatisch-Intellektuell 2004 (63 Dok., ohne Indikatoren)	25 %
Automatisch-Intellektuell 2006 (63 Dok., mit Indikatoren)	40 %
Automatisch-Intellektuell 2011 (11.504 Dok., mit Indikatoren)	41,5 %
Intellektuell-Intellektuell (37 Dok., Interindexer-Konsistenz)	57 %

* Es wurden nur Zeitschriftenaufsätze ausgewertet

Arbeitsweise von AUTINDEX 2

Linguistische Analyse der Titel, Abstracts und Autorenschlagworte mit u.a.

- Satz- und Wortformerkennung
- Erkennung von morphologischen (Haus/Häuser), syntaktischen (Reduktion von Kosten -> Kostenreduktion) und semantischen (Synonyme) Varianten von Worten
- Kompositazerlegung
- Ausschluss von Stopwörtern und ihren morphologischen Varianten
- Zuordnung des ermittelten Begriffs zu einer semantischen Klasse

Ergebnis: desambiguierte syntaktische Repräsentation des Textes

Arbeitsweise von AUTINDEX 3

Nach der Bearbeitung linguistischer Einheiten:
Gewichtung der ermittelten Terme unter Berücksichtigung der

- Häufigkeit des Vorkommens des Terms im Text (Termfrequenz)
- Stellung des Terms im Text
- im Text vorkommenden semantischen Klassen

Ergebnis: Wörter und Phrasen, die häufig bzw. an prominenter Stelle im Text vorkommen bzw. zu den häufigsten semantischen Klassen gehören, werden als Deskriptorkandidaten gesammelt und mit den Begriffen aus dem Thesaurus abgeglichen.
Wenn Übereinstimmung gefunden und ein definierter Schwellenwert überschritten wird, erfolgt Deskriptorvorschlag.