gesis
Leibniz-Institut für Sozialwissenschaften

Datenbank SOLIS: Automatische Erschließung von Dokumenten mit dem MindServer - Categorizer

Trends aus ersten Analysen

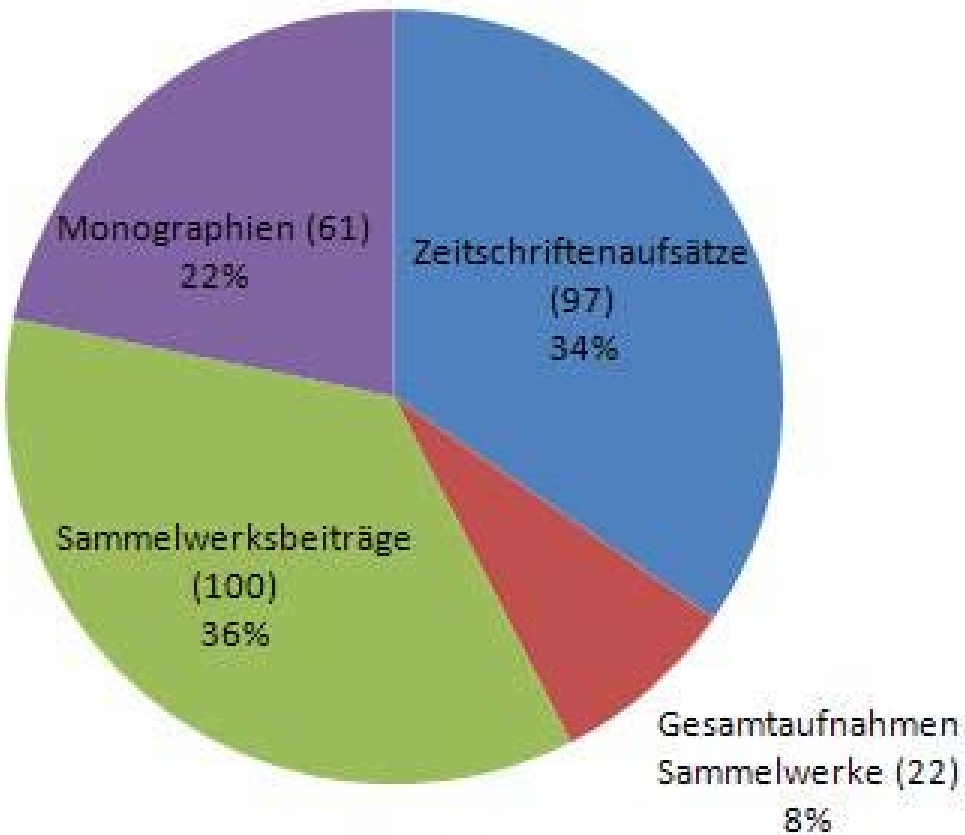
Monika Zimmer

25. November 2010

MindServer - Categorizer

- ***Arbeitsweise / Prinzip:***
  „Probabilistische latente semantische Analyse“ (PLSA) in Verbindung mit „Support Vector Machine“
- ***Trainingsbasis:***
  368.000 Dokumente aus SOLIS

Stichprobe: 280 Dokumente



Gesamtmenge n=280, absolute Werte in Klammern

Untersuchte Variablen

- Anzahl Schlagwörter und Klassifikationen - automatische vs. intellektuelle Bearbeitung
- Überschneidungen beider Bearbeitungen
- Einbezug Titelbegriffe
- „Falsche“ Schlagwörter /Klassifikationen

für

- Gesamtstichprobe
- Dokumentart
- Textlänge des Abstracts
- Art bzw. Herkunft des Abstracts

Automatische vs. intellektuelle Indexierung:

*Vergleich **Schlagwörter** - Gesamtstichprobe* (n = 280 Dok.)

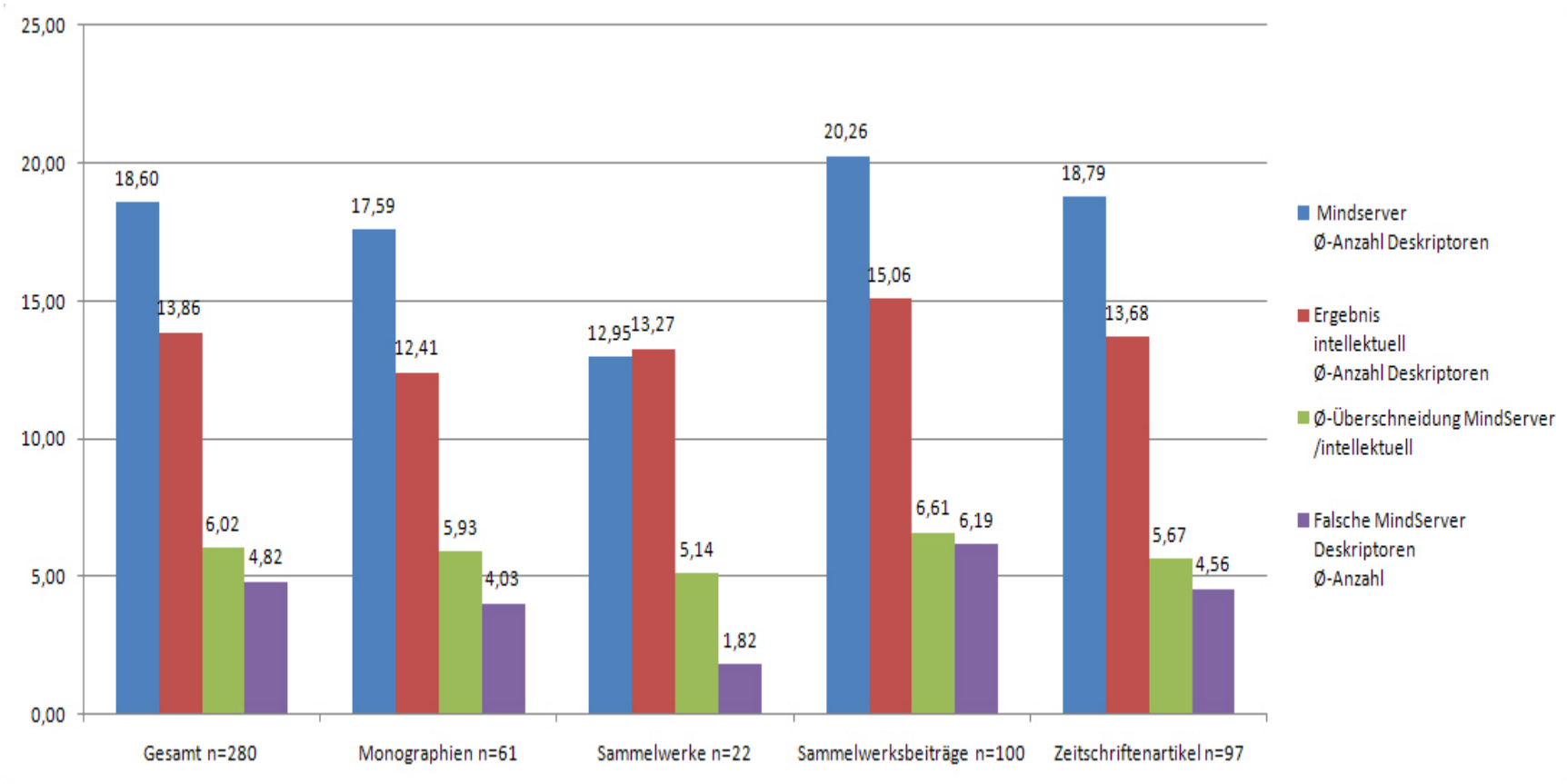
- Anzahl SW intellektuelle Indexierung i.D. => 13,9
- Anzahl SW automatische Indexierung i.D. => 18,6
- Anzahl identischer Schlagwörter i.D. => 6,0
- zusätzliche „passende“ SW (MS-Cat.) => 0,6
- nicht berücksichtigte Titelbegriffe (MS-Cat.) => 0,6
- „falsche“ Schlagwörter (MindServer-Cat.) => 4,8

Automatische vs. intellektuelle Indexierung:

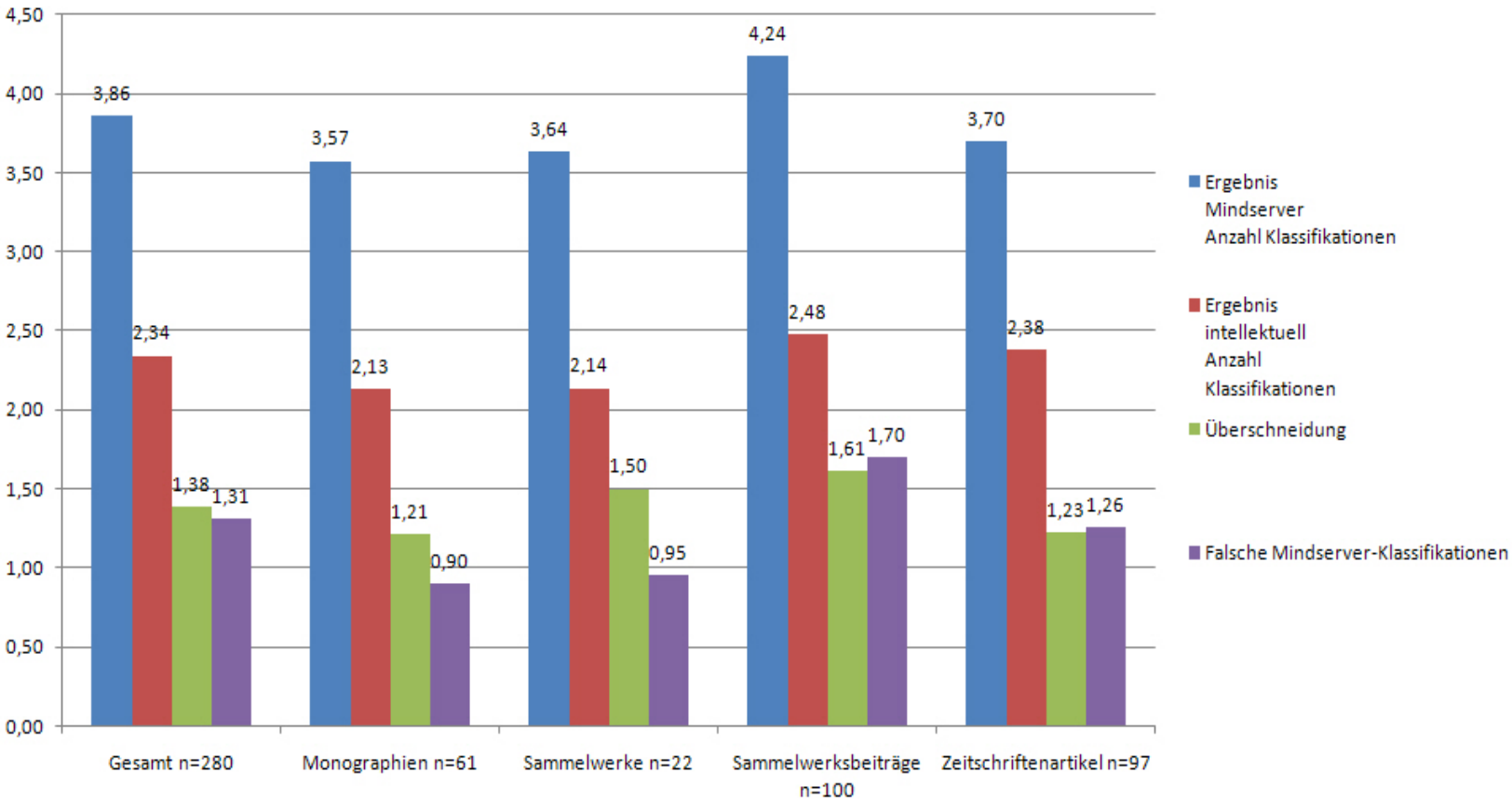
Vergleich ***Klassifikationen*** *- Gesamtstichprobe* (n = 280 Dok.)

- Anzahl Klass. intellektuelle Indexierung i.D. => 2,3
- Anzahl Klass. automatische Indexierung i.D. => 3,9
- Anzahl identischer Klassifikationen i.D. => 1,4
- zusätzliche wichtige Klass. (MS-Cat.) => 0,2
- „falsche“ Klassifikationen (MindServer-Cat.) => 1,3

Automatische vs. intellektuelle Indexierung nach **Dokumentarten** - *Vergleich* ***Schlagwörter***



Automatische vs. intellektuelle Indexierung nach **Dokumentarten** - *Vergleich **Klassifikationen***



Automatische vs. intellektuelle Erschließung nach **Textlänge des Abstracts**

Ergebnisse: mit zunehmender Textlänge

- wird die Anzahl der von MindServer vergebenen Schlagwörter geringer,
- steigt der Anteil der identischen Schlagwörter, die automatisch und intellektuell vergeben wurden,
- werden Hauptklassifikationen häufiger gefunden und
- sinkt die Zahl der von MindServer falsch vergebenen Klassifikationen und insbesondere Schlagwörter (um mehr als die Hälfte)

Die **Art des Abstracts** ergab keine Rückschlüsse auf Unterschiede zwischen automatischer und intellektueller Erschließung

Vielen Dank für Ihre Aufmerksamkeit!