Jan-Helge Jacobs (DNB)

Wörterbucharbeit für die maschinelle Beschlagwortung

Grundlagen maschinelle Beschlagwortung

- Textpassagen **und** Terme aus der GND werden jeweils in eine Matching-Repräsentation umgewandelt und für diese Repräsentationen wird dann eine Übereinstimmung geprüft (Matching).
- Bei Sachschlagwörtern wird auf Segment-Ebene gematcht. Das ermöglicht die Identifikation vieler grammatikalischer Varianten.
- Bei allen anderen Satzarten werden fast ausschließlich zuvor festgelegte Genitivvarianten zu Grundformen reduziert.

Verschiedene Mappingmodi

die	Myokarditis	ist	eine	sammelbezeichnung
für	entzündliche	erkrankungen	des	herzmuskels
mit	unterschiedlichen	ursachen		

Original

die	myokarditis	ist	eine	sammelbezeichn
für	entzünd	erkrankung	des	herzmuskel
mit	unterschied	ursach .		

Stem

die	myo kard itis	ist	eine	sammel bezeich
für	entzuend	krank	des	herz muskel
mit	unterschied	ursache		

Segment

Positivbeispiele

Text:perkutane transluminale Koronarangioplastie
SW: Perkutane transluminale koronare Angioplastie (Syn.)

Text: Beugesehnennahtmaterialien
SW: Beugesehne + Nahtmaterial (Dekomposition)

Text: Cholinesterasen
SW : Enzym (über Segment „ase“)

Text: Schriftspracherwerb
SW: Schriftsprache + Spracherwerb (über Hinweissatz)

Matching von Sachschlagwörtern: Segmentierung

Sach-SW/Syn.	Mappingform	**Segmente**	Text
Held	Held	**held**	Helden
Roman	Roman	**roman** [held]	Romanhelden
Thema	Thema	**thema**	Themen
Verfremdung	Verfremdung	**verfremd**	Verfremdendes
Sprachstil	Sprachstil	**sprech stil**	Sprache und Stil
Zuteilung **$g**Menge	Zuteilung	**zuteil**	mitzuteilen
Sozialfeld	Sozialfeld	**soz feld**	sozialen Feld

Wörterbuchprofile im averbis-System

Wörterbuchprofile sind Anweisungssammlungen, die festlegen, welche Terme aus den ausgewählten GND-Terminologien in welcher Weise für das Matching bereitgestellt werden sollen.
Es gibt drei verschiedene Mappingmodi:

DEFAULT – Segmentierung für Sach-SW, Genitivmodus für die anderen Satzarten

EXACT – Term muss genau so im Text stehen (auch Groß-/Kleinschreibung)

IGNORE – Term steht nicht für die maschinelle Beschlagwortung zur Verfügung

Mappingmodus: DEFAULT

GND-Term (Sach-SW): **Werk**
matcht im DEFAULT-Modus mit folgenden Textstellen:

Werke

Werken

Werkschau (Kompositumsbestandteil)

GND-Term (kein Sach-SW): **Thüringen**
matcht im DEFAULT-Modus mit folgenden Textstellen:

Thüringen

THÜRINGEN

Thüringens

Mappingmodus: EXACT

GND-Term (Sach-SW):
SOKRATES$gBibliotheksinformationssystem
matcht im EXACT-Modus mit folgender Textstelle:
SOKRATES

Terminologie-Filter

Die Terme aus den GND-Terminologien können mit Hilfe von regelbasierten Filtern systematisch mit passenden Mappingmodi (DEFAULT/EXACT/IGNORE) versehen werden.

Es werden im Folgenden einige Filter und ihre Arbeitsweise vorgestellt.

DNB Ignore Short Synonyms Filter

Selektiert alle synonymen Terme eines Begriffs, deren Länge weniger als vier Zeichen beträgt und setzt deren Mappingmodus auf *ignore*.

Ausnahmen:

DDR

USA

EU

NRW

Ulm

neulich als Ausnahme hinzugefügt: **DNA** (Sach-SW)

DEUTSCHE NATIONAL BIBLIOTHEK

GND-Motiv-Filter

Selektiert alle synonymen Terme eines Begriffs, die den Homonymzusatz <Motiv> aufweisen und setzt deren Mappingmodus auf *ignore*.

SWD-spezifischer Filter

Selektiert sämtliche Terme eines Begriffs, deren Vorzugsbenennung einen der folgenden identifizierenden Zusätze aufweist und setzt deren Mappingmodi auf *ignore*:

<Wort>
<Morphem>
<Phonem>
<Personenname>
<Ortsname>
<Familienname>
<Druckschrift>

PND-spezifischer Filter

Selektiert synonyme Personennamen und setzt deren Mappingparameter auf ignore, wenn mindestens eine Mapping-Variante

- aus weniger als vier Zeichen besteht,
- nur aus Akronymen besteht,
- einen Asterisk (*) oder Auslassungspunkte (…) enthält,
- nur aus einem Stoppwort oder einem Vornamen besteht. (GND-Bsp.: **$PWilly**)

DNB PND Kurzwort

Selektiert synonyme, aus nur einem Wort bestehende Terme, bei deren Ansetzungsformen Kommata auftreten [also moderne Namen mit Nachname, Vorname] und setzt die Mappingparameter der synonymen Ein-Wort-Terme auf *ignore*.

Bsp.:

100 Knackfuss, Eduard
400 $P Lucas [-> ignore]

manuelle Modifikationen

Neben der Möglichkeit, Filter direkt anzuwenden, kann man auch die Filterungen als Datei ausgeben lassen und bei Bedarf nur einzelne Modifikationen aus den Filterergebnissen übernehmen.

Der Modus kann bei einzelnen Terme oder Schlagwörtern auch manuell modifiziert werden.

Bsp. für manuelle Modifikation : **Battle** [Tg1] -> ignore
Grund: Weil für den Begriff “Battle” nur das oben genannte Geografikum als Kandidat zur Verfügung steht.