Anne Betz, Daniel Hörnig, Klaus Lepsky

Lehr- und Lernsystem Information Retrieval

(LIR)

- Dokumentation -

Stand: 15.12.03

Inhalt

1. Das System LIR

1.1 Aufgaben und Ziele

Das Lehr- und Lernsystem Information Retrieval (LIR) vermittelt Inhalte der automatischen Indexierung und des Information Retrieval experimentell und anwendungsnah. Die Grundlage dafür sind eine hohe Transparenz aller Teilschritte und die Option, das System möglichst variabel zu konfigurieren.

1.2 Übersicht über den LIR-Ablauf

LIR gliedert den Prozess des Aufbaus einer Umgebung für das Information Retrieval in mehrere Teile:

1. Export der zugrundeliegenden MIDOS-Datei
2. Automatische Indexierung der Daten mit IDXWin
3. Möglicherweise Verbesserung der Indexierungsergebnisse mit SelRVL und IDXWin
4. Statistische Gewichtung der gewonnenen Daten mit RVL2DB
5. Umsetzung der Ergebnisse in ein Speicherformat mit OPAC_STO
6. Import der Indexierungsergebnisse in MIDOS
7. Recherche in den Daten

Im folgenden werden die einzelnen Schritte des LIR-Prozesses zur Übersicht kurz beschrieben. Hierbei soll ersichtlich werden, welche Dateien und Programme in jedem Schritt verwendet werden und welche Ergebnisse in jedem Schritt entstehen. Eine genauere Übersicht der einzelnen Arbeitsschritte folgt dann in Kapitel 3.

1.2.1 Export der zugrundeliegenden MIDOS-Datei

Am Beginn des LIR-Prozesses steht eine Datei, die indexiert werden soll. Es handelt sich dabei um eine kleine Testdatenbank in MIDOS2000, die zuerst in das IDX-Eingabeformat überführt werden muss. Dazu muss diese Datenbank aus MIDOS über ein Ausgabeformat in eine Text-Datei (*.txt) umgewandelt werden, die den Anfordererungen von IDX entspricht.

1.2.2 Automatische Indexierung der Daten mit IDXWin

Im zweiten Teil des LIR-Prozesses erfolgt die eigentliche Indexierung der im ersten Schritt exportierten Datei. Für diese Indexierung wird neben der TXT-Datei eine Konfigurationsdatei (*.cfg) verwendet. In dieser CFG-Datei wird die Art der Indexierung genauer spezifiziert. Außerdem wird darin festgelegt, welche Wörterbücher IDXWin verwenden soll. Als Abschluss des Indexierungslaufs gibt IDXWin eine Ergebnisdatei (sog. RVL-Datei) aus.

1.2.3 Verbesserung der Indexierungsergebnisse mit SelRVL, IDXWin und WoEx_M

Der nun folgende (optionale) Schritt umfasst die intellektuelle Nachbearbeitung der zuvor automatisch generierten Ergebnisse. Alle Wörter, die den Indexierungswörterbüchern unbekannt waren, wurden in der RVL-Datei als solche markiert. Mit SelRVL können diese unbekannten Wörter aus der RVL-Datei extrahiert und in ein leeres Wörterbuch geschrieben werden. Dieses Wörterbuch dient dann als Grundlage für einen zweiten Indexierungslauf mit IDXWin. In einem ähnlichen Verfahren

können potenzielle Mehrwortgruppen extrahiert und eingetragen werden. Hierfür wird WordExtract (WoEx_M) verwendet. Mit den so entstandenen Mehrwortbegriffswörterbüchern kann dann eine neue Indexierung durchgeführt werden.

1.2.4 Statistische Gewichtung der gewonnenen Daten mit RVL2DB

Im nächsten Teil des LIR-Prozesses sollen die indexierten Daten statistisch gewichtet werden. Dazu wird das Programm RVL2DB verwendet. Dieses gewichtet die RVL-Datei unter Zuhilfenahme eines bestehenden Relationenwörterbuches und eines leeren Relationenwörterbuchs. Als Ergebnis entsteht eine STO-Datei, die zusätzlich Gewichtungswerte enthält.

1.2.5 Umsetzung der Ergebnisse in ein Speicherformat mit OPAC_STO

Die durch IDXWin generierte RVL-Datei ist für die Nutzung in einem Datenbanksystem wenig geeignet. Deshalb wird die RVL-Datei in diesem Schritt in ein leicht zu importierendes Datenformat überführt. Dies geschieht mit Hilfe des Programmes OPAC_STO. Die RVL-Datei wird also in eine STO-Datei umgewandelt.

1.2.6 Import der Indexierungsergebnisse in MIDOS

Die zuvor generierte STO_Datei mit den Indexierungsergebnissen wird nun in MIDOS importiert. Dabei ist es sinnvoll, die bibliografischen Daten (die ja Ausgangspunkt des LIR-Prozesses waren) mit den Indexierungsergebnissen in einer neuen Datenbank zu verbinden. Diese Datenbank ist das Ergebnis des LIR-Prozesses.

2. Die einzelnen Programme

2.1 MIDOS

Die für den LIR-Prozess zugrunde liegenden Daten bilden eine Teilmenge der Datenbank „Literatur zur Inhaltserschließung“. Für diese Teilmenge ist eine neue Midos2000-Datenbank anzulegen. Midos2000 verfügt über leistungsfähige Im- und Export-Möglichkeiten für Daten, ist v.a. aus diesem Grund als Datenbank für LIR ausgewählt worden.

2.2 IDXWin

IDX existiert seit 1986 als DOS-basierte Software zur maschinellen Indexierung von Textdaten. Seither bildet IDX die Grundlage für verschiedene Einzelanwendungen (im bibliothekarischen Bereich z.B. das Programmpaket IDX/MILOS). Mit **IDXWin** liegt eine Windows-Version für IDX vor.

IDX führt die folgenden Arbeitsschritte durch:

- Lemmatisierung, d.h. Rückführung von Wortformen auf die entsprechende Grundform (Beispiel: Prinzipien - Prinzip)
- Identifikation von Mehrwortbegriffen (Beispiel: wissenschaftliche Bibliothek)
- Zuordnung von Grundformen und Mehrwortbegriffen zu ihren Teilwörtern (Beispiele: Schlagwortkatalog – Schlagwort, Katalog / wissenschaftliche Bibliothek – wissenschaftlich, Bibliothek)
- Identifikation und Beseitigung von Teilworttilgungen (Beispiel: Haus- und Gartenwirtschaft)

- Bereitstellung von Wortrelationierungen, also von Synonymen oder Wortderivaten (Beispiele: Bücherei – Bibliothek / Anwendung – anwenden)
- Bereitstellung von Übersetzungsäquivalenten (bei Zuschaltung entsprechender Übersetzungswörterbücher)

IDX ermöglicht dadurch folgende Retrievalmöglichkeiten:

- Retrieval mit Grundformen und sinntragenden Teilwörtern (das Trunkierungsproblem wird vermindert)
- Retrieval mit sinntragenden Mehrwortbegriffen (das Adjacencyproblem wird verringert)
- Retrieval unter Einbeziehung von Übersetzungsäquivalenten

2.3 RVLShow

RVLShow ist ein JAVA-Tool für die Anzeige der von IDXWin erzeugten Ergebnisdateien (sog. RVL-Dateien). Zu beachten ist, dass RVLShow ein reiner Viewer ist, der keine Bearbeitung der RVL-Datei ermöglicht.

2.4 SELRVL

Das Programm SELRVL erlaubt die Selektion von Einträgen aus einer RVL-Datei. So wird es möglich, durch IDXWin vorgenommene Einträge eines bestimmten Typs aus der RVL-Datei in ein (leeres) Wörterbuch zu übertragen.

2.5 WoEX_M (WordExtract)

Mit WoEx_M (WordExtract) können Mehrwortgruppen in Dokumenten erkannt und übernommen werden. Dies geschieht auf der Grundlage syntaktischer Regeln (beispielsweise sind alle Wortpaare, bestehend aus einem Adjektiv und folgendem Substantiv wie „öffentliche Bibliothek“ gute Kandidaten für eine Mehrwortgruppe). Diese Mehrwortgruppen werden in ein leeres Relationenwörterbuch eingetragen.

2.6 WBTool

WBTool dient zur Pflege von Wörterbüchern. Hierbei können bereits vorhandene Einträge in den Wörterbüchern modifiziert oder gelöscht werden. Außerdem ist es möglich, neue Einträge oder Korrekturvorschläge vorzunehmen.

2.7 OPAC_STO

Das Programm OPAC_STO wandelt RVL-Dateien in leicht zu importierende sog. STO-Dateien um.

2.8 RVL2DB

RVL2DB nimmt eine statistische Gewichtung der Deskriptoren in einer RVL-Datei vor. RVL2DB kennt 5 Gewichtungsalgorithmen.

3. Der Ablauf von LIR

3.1 Export der zugrundeliegenden MIDOS-Datei

Für den ersten Schritt des LIR-Prozesses wird eine MIDOS-Datei mit bibliografischen Daten benötigt.

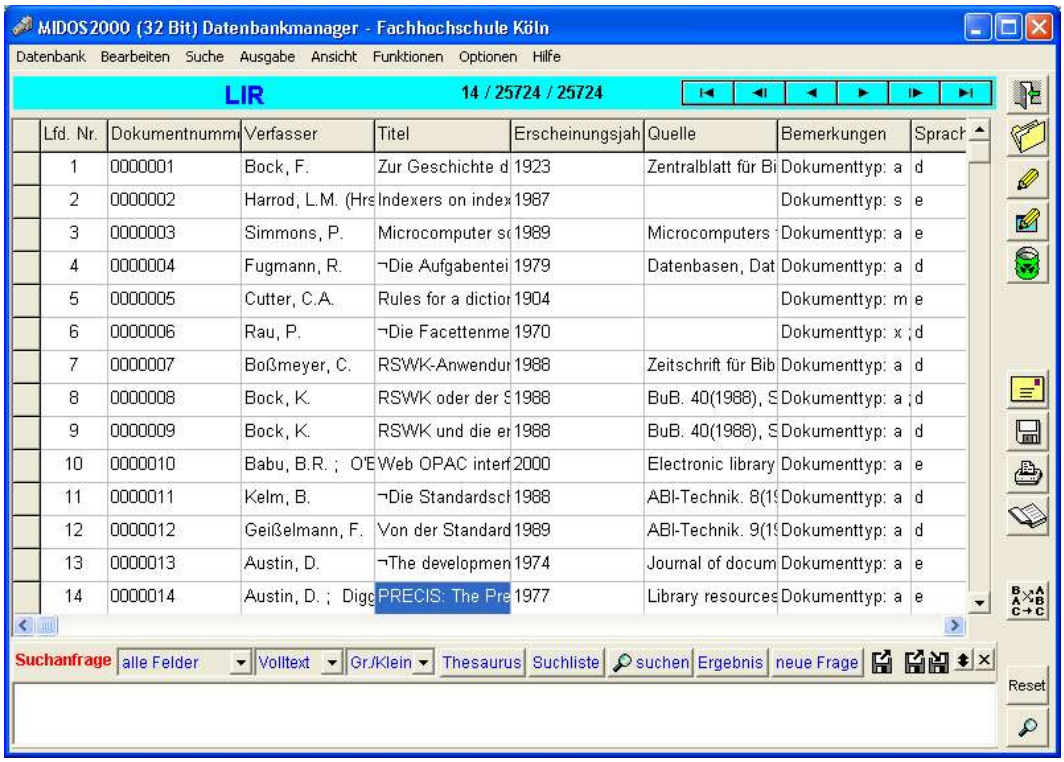


Abb.1: MIDOS-Datenbank LIR

Für die Indexierung werden alle Kategorien, die einen inhaltlichen Bezug zum Dokument haben exportiert. Der Inhalt der Datenbank muss also analysiert werden. IDXWin erfordert eine TXT-Datei nach folgendem Muster:

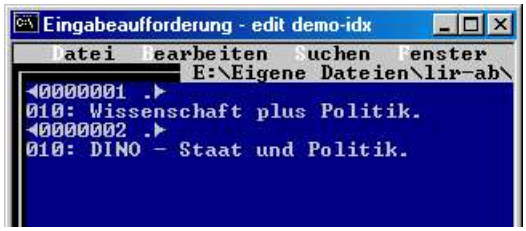


Abb. 2: IDX-Eingangsformat

Zu beachten ist, dass vor den Identnummern „ASCII 17“ stehen muss. Hinter den Identnummern folgt dann „Spatium/Leerzeichen, Punkt, ASCII 16“:

Abb. 3: für IDX geeignete Identnummer

Kategorien werden eingeleitet durch die Kategoriennummer, gefolgt von einem Doppelpunkt. Am Ende einer Kategorie steht „ .“ (Spatium/Leerzeichen, Punkt). Die Kategorienummern können auch entfallen.

Die MIDOS-Datenbank muss nun so exportiert werden, dass eine geeignete TXT-Datei erzeugt wird. Dazu muss in MIDOS unter „Bearbeiten/Ausgabeformate“ ein entsprechendes Ausgabeformat erstellt werden:

Abb. 4: MIDOS – Ausgabeformate erstellen

Nun muss in MIDOS unter Ausgabe/Datei der Export erfolgen. Dabei muss ein Dateiname vergeben werden. Als Speicherformat sollte „Text (Ausgabeform)“ verwendet werden. MIDOS gibt die gewählten Kategorien nun in einer TXT-Datei aus.

Abb. 5: MIDOS - Export

Die erzeugte TXT-Datei sollte etwa so aussehen (<< und >> sind durch ASCII 17 u. 16 zu ersetzen):

```
<<00018 .>>
020: Entwicklung und Grundprinzipien von PRECIS, einem computergestütztem Inde-
xierungssystem .
<<00019 .>>
020: DIN 31623: Indexierung zur inhaltlichen Erschließung von Dokumenten .
025: T.1: Begriffe, Grundlagen; T.2: Gleichordnende Indexierung mit Deskriptoren;
T.3: Syntaktische Indexierung mit Deskriptoren .
<<00020 .>>
020: Entwicklung und Fortschritt bei Klassifikation und Indexierung .
<<00021 .>>
020: PASSAT: Programm zur automatischen Selektion von Stichwörtern aus Texten .
<<00022 .>>
020: Methodische Rahmenregelung zur Erarbeitung und Anwendung sachbezogener Inde-
xiermuster .
<<00023 .>>
020: DIN 31623 oder die Problematik des genormten Indexierens .
```

3.2 Automatische Indexierung der Daten mit IDXWin

Der nächste Schritt des LIR-Prozesses ist die automatische Indexierung der im vorhergehenden Schritt erzeugten TXT-Datei.

IDXWin ist als wörterbuchbasiertes Indexierungssystem von den verfügbaren und eingebundenen elektronischen Wörterbüchern direkt abhängig. Es benötigt die folgenden 3 Arten von Wörterbüchern: Systemwörterbücher, Benutzerwörterbücher und Musterwörterbücher.

Systemwörterbücher sind alle Wörterbücher, die zum Systemumfang des Indexierungsprogramms gehören und die durch Benutzer nicht verändert werden können. Dazu gehören: WBDSTX (Rechtschreibwörterbuch Deutsch, enthält die Grundformen der deutschen Sprache, erledigt Grundformerzeugung der Quellwörter und algorithmische Kompositumzerlegung), SWBRELD1 (ein Relationenwörterbuch, das v.a. Synonymrelationen aber auch lexikalisierte Kompositumzerlegung enthält), SWBRELD3 (Relationenwörterbuch, das v.a. Kompositumzerlegungen enthält) und Relneu (Relationenwörterbuch, das ausschließlich Mehrwortgruppen enthält).

Benutzerwörterbücher werden benötigt, um Erweiterungen des Wortschatzes zu erfassen. Für jeden Wörterbuchtyp – Rechtschreib- und Relationenwörterbücher – gibt es entsprechende Benutzerwörterbücher.

Musterwörterbücher sind leere Wörterbücher, die kopiert und umbenannt werden können. Dadurch werden sie zu Benutzerwörterbüchern. Die beiden Dateien für das Muster-Rechtschreibwörterbuch sind ID.dat und ID.ind, die beiden Dateien für das Muster-Relationenwörterbuch sind Rel.dat und Rel.ind.

Nach dem Start von IDXWin erwartet das Programm eine Eingabe-Datei (Dateiendung TXT) und eine Konfigurations-Datei (Datei-Endung CFG). Die Eingabe-Datei ist die aus MIDOS exportierte TXT-Datei . Als Konfigurations-Datei sollte die Datei „idx_lir.cfg“ (in C:\LIR\IDX).

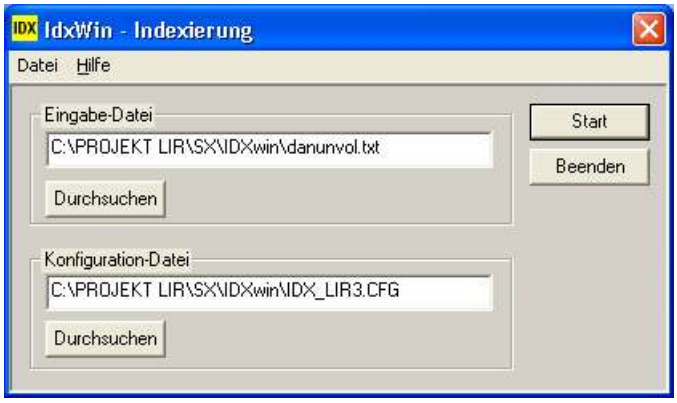


Abb. 7: IDXWin

Die Konfigurations-Datei kann mit Hilfe des Programms IDXCFG (idxcfg.jar, zu finden unter: C:\Lir\Tools\IDXCFG) bearbeitet werden. IDXCFG verfügt über eine eigene Hilfedatei, in der die IDX-Parameter genauer erklärt werden. Diese Hilfe wird über das grüne Fragezeichen aufgerufen.

Abb. 8: IDXCFG - Button „Hilfe“

Wenn beide Dateien angegeben worden sind, kann IDXWin gestartet werden. Die Dauer des nun startenden Indexierungslaufs ist abhängig von der Größe der verwendeten Eingabe-Datei.

IDXWin identifiziert mit Hilfe von Rechtschreibwörterbüchern. Je nach Konfiguration durchläuft das Programm nun bis zu 9 mögliche Indexierungsphasen:

Phase 0: Grundformermittlung auf Einzelwortebene

Erfasst werden hierbei eine automatisch vergebene laufende Nummer, die Stammwortklasse und die Wortformenklassen. Dabei berücksichtigt IDXWin nur das einzelne Wort, nicht aber den Kontext. In dieser Phase protokolliert IDXWin die Wörter in der Reihenfolge ihres Auftretens.

Phase M: Mehrwortbegriffe

Die Phase M dient zur Identifizierung von Mehrwortbegriffen; dazu wird ein spezielles Mehrwortwörterbuch verwendet. Hierbei werden die möglichen Mehrwortbegriffe vor einer Bestätigung lexikalisch überprüft. Durch Bindestrich verbundene Wörter werden als solche erfasst.

Phase B: Ermittlung getilgter Teilwörter

Als nächstes erfolgt die Ermittlung getilgter Teilwörter aus mehrteiligen Wendungen („aus- und einsteigen“):

Phase 1: Strukturanalyse, Stoppwortermittlung

In Phase 1 werden die Ergebnisse der vorangehenden Phasen 0 und M zusammengeführt. Außerdem werden alle gefundenen Stoppwörter gelöscht oder entsprechend markiert.

Phase 2: Derivation und Dekomposition, Mehrwortkontrolle

Die durch Derivation (Wortableitung) und Dekomposition (Wortzerlegung) ermittelten Wörter werden als neue Einträge für das Rechtschreib- und das Relationenwörterbuch aufbereitet. Außerdem werden die Mehrwortbegriffe mit Hilfe eines Wörterbuches kontrolliert.

Phase 3: Aufbau der Indexierungsergebnisdatei (IX3)

In Phase 3 wird dann die Indexierungsergebnisdatei (Datei-Endung IX3) produziert. Diese Datei ist als ein vorläufiges Ergebnis zur Indexierung zu sehen.

Phase G: Aufbau der Relationendatei (RVL)

Phase G führt die bisherigen Ergebnisse der Indexierung nun mit den entsprechenden Wortrelationen aus den Relationenwörterbüchern zusammen. Dabei wird eine zweite Ergebnisdatei (Datei-Endung RVL) ausgegeben. Da die RVL-Datei auch die Wortrelation genauer spezifiziert, ist sie das eigentliche Ergebnis des Indexierungslaufs. Sie wird für die weiteren Schritte im LIR-Prozess benötigt.

Phase T: Übersetzung

In dieser Phase würden die ermittelten einzelnen Wörter und Mehrwortgruppen übersetzt. Voraussetzung ist ein geeignetes Übersetzungswörterbuch.

Am Ende des Indexierungslaufs erzeugt IDXWin eine RVL-Datei mit dem gleichen Namen wie die indexierte TXT-Datei. Diese Datei enthält sämtliche Deskriptoren und Relationierungen. Zur Betrach-

tung der Datei wird das Programm RVLShow benutzt (Näheres siehe 3.3). Die erzeugte RVL-Datei kann im nächsten Schritt noch verbessert werden.

3.3 Optionale Verbesserung der Indexierungsergebnisse mit SelRVL, IDXWin und WoEx_M

Zunächst einmal ist es ratsam, die Indexierungsergebnisse, die von IDXWin erzeugt wurden, zu überprüfen. Die Anzeige der RVL-Datei erfolgt mit dem JAVA-Tool RVLShow.

Nachdem RVLShow geöffnet wurde, muss die zu öffnende RVL-Datei ausgewählt werden.

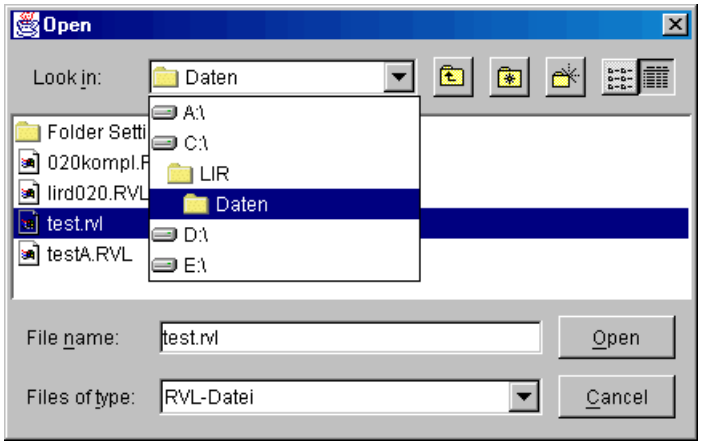


Abb. 29: Öffnen einer RVL mit RVLShow

Als nächstes sollte der zu benutzende Zeichensatz festgelegt werden. Es existieren 3 Optionen:

1. ASCII (CP437)
2. ISO 8859-1 (auch ISO Latin-1, Standard-Zeichensatz für westeuropäische Zeichen)
3. ISO 8859-15 (auch ISO Latin-9, erweiterter Zeichensatz für westeuropäische Zeichen, enthält u.a. das Euro-Zeichen)

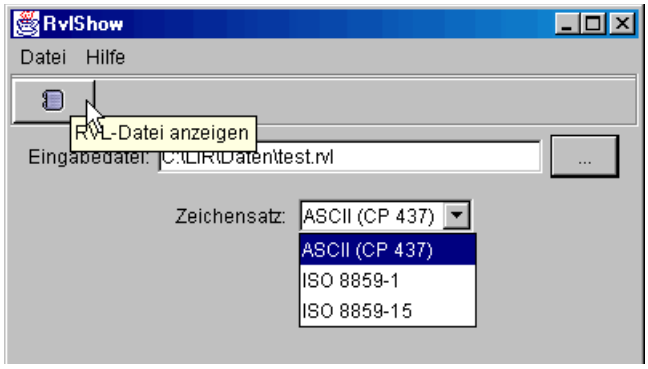
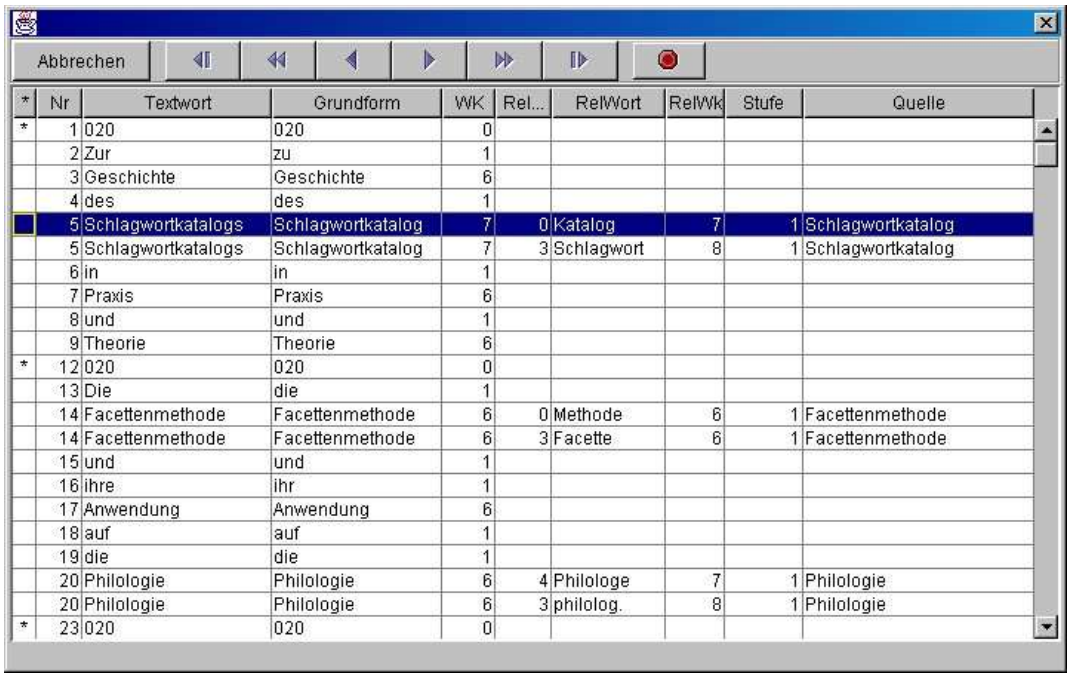


Abb. 30: Auswahl des Zeichensatzes in RVLShow

Hier sollte der Zeichensatz ISO 8859-1 ausgewählt werden. Die Anzeige erfolgt in einer Tabellenansicht, in der die Textwörter, ihre Grundformen und die laufenden Nummern erscheinen.



Abbrechen

*	Nr	Textwort	Grundform	WK	Rel...	RelWort	RelWk	Stufe	Quelle
*	1	020	020	0					
	2	Zur	zu	1					
	3	Geschichte	Geschichte	6					
	4	des	des	1					
	5	Schlagwortkatalogs	Schlagwortkatalog	7	0	Katalog	7	1	Schlagwortkatalog
	5	Schlagwortkatalogs	Schlagwortkatalog	7	3	Schlagwort	8	1	Schlagwortkatalog
	6	in	in	1					
	7	Praxis	Praxis	6					
	8	und	und	1					
	9	Theorie	Theorie	6					
*	12	020	020	0					
	13	Die	die	1					
	14	Facettenmethode	Facettenmethode	6	0	Methode	6	1	Facettenmethode
	14	Facettenmethode	Facettenmethode	6	3	Facette	6	1	Facettenmethode
	15	und	und	1					
	16	ihre	ihr	1					
	17	Anwendung	Anwendung	6					
	18	auf	auf	1					
	19	die	die	1					
	20	Philologie	Philologie	6	4	Philologe	7	1	Philologie
	20	Philologie	Philologie	6	3	philolog.	8	1	Philologie
*	23	020	020	0					

Abb. 31: Ansicht der RVL-Datei mit RVLShow

Unter anderem werden auch die Wortklassenkodes und die Relationenkodes angezeigt (s. Anhang). Innerhalb von RVLShow kann die RVL-Datei nicht weiter bearbeitet oder (auch unabsichtlich) verändert werden. Die Dateiansicht kann mit „Abbrechen“ verlassen werden.

IDXWin kann aufgrund seiner wörterbuchgestützten Indexierung nur diejenigen Wörter erkennen, die auch in den verwendeten Wörterbüchern enthalten sind. Wörter, die IDXWin während der Indexierung nicht erkennt, legt es in einer Datei mit der Datei-Endung LST ab. Um herauszufinden, ob sich unter den nicht erkannten Wörtern eventuell brauchbare Deskriptoren befinden, sollte die LST-Datei mit einem Editor geöffnet und betrachtet werden. Alle unbekannten Wörter könnten, wenn man sie in die Indexierung einbezieht (sie also in die Indexierungswörterbücher einfügt), das Ergebnis verbessern.

Diese unbekannten Wörter können mit Hilfe des Programms SelRVL aus der RVL-Datei extrahiert und in die Indexierungswörterbücher eingetragen werden. Dies ist deswegen möglich, weil IDXWin alle unbekannten Wörter in der RVL-Datei mit einem <U> markiert hat.

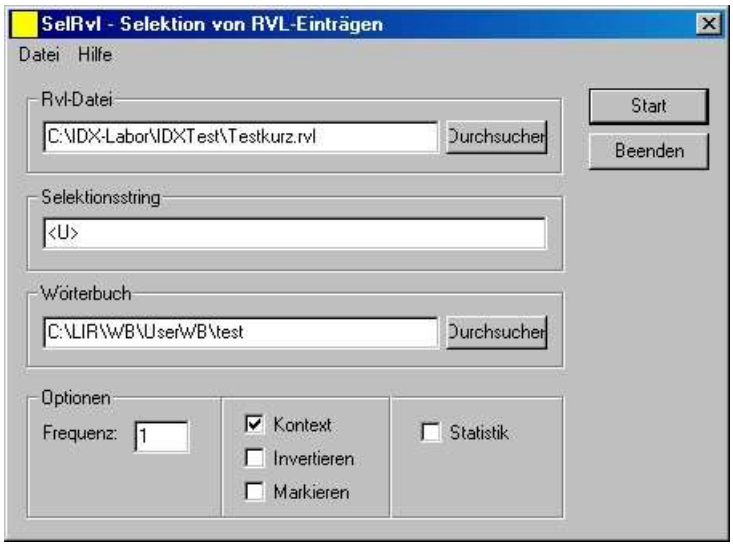


Abb. 32: SelRVL

Nachdem SelRVL gestartet wurde, benötigt es eine RVL-Datei (die im vorigen Schritt erstellte), einen Selektionsstring (die zu suchende Zeichenkette, in diesem Fall <U> für unbekannte Wörter) sowie ein Wörterbuch (in diesem Fall ein leeres, in welches die Ergebnisse geschrieben werden). Wenn dabei nach der Wortklassenkennung <U> gesucht wird, sollte diese groß geschrieben werden.

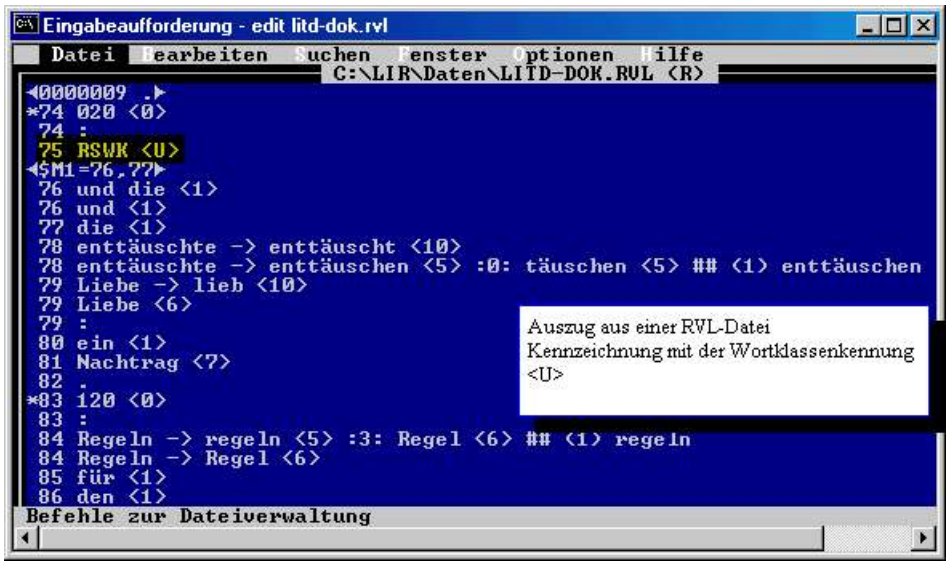


Abb. 33: RVL-Datei, Wortklassenkennung <U>

Unter den Optionen ist es möglich, auch den Kontext des Wortes mit in das leere Wörterbuch aufzunehmen, dies kann auch invertiert geschehen (statt der Wort-Kontext-Relation wird eine Kontext-Wort-Relation erzeugt). Außerdem kann das Wort im Kontext auch markiert werden. Darüber hinaus besteht auch die Möglichkeit, mit SelRVL eine Statistik zu erstellen, in der die Häufigkeit eines Eintrages gezählt wird.

Nachdem SelRVL die unbekannten Wörter in ein leeres Rechtschreibwörterbuch geschrieben hat, sollte eine neue Indexierung durchgeführt werden, bei der das neue Wörterbuch verwendet wird (Ablauf wie in 3.2). Die Ergebnisse sollten mit RVLShow betrachtet werden. Danach sollte auch das neue, von SelRVL beschriebene Wörterbuch analysiert werden. Eine Anzeige des Wörterbuches erfolgt mit dem Programm WBTool, das mit der Datei „wbtool.jar“ im Verzeichnis C:\LIR\TOOLS gestartet wird. Nach dem Start von WBTool kann ein Wörterbuch über den Button „Wörterbuch öffnen“ oder über Datei/Öffnen ausgewählt werden.

Es öffnet sich eine Tabellenansicht des Wörterbuches, in der Wortklassenkennung (WK), Endungskode (EN), Fugenkennung (FU), Frequenz (FQ, zu vernachlässigen) und Wortlaut (WL1 und WL2) angezeigt werden.

Abb. 34: WBTool - Wörterbuchansicht

In dieser Ansicht können die Einträge des von SelRVL generierten Wörterbuches betrachtet und editiert werden.

Analog verläuft die Verbesserung der Ergebnisse aus der Mehrwortgruppenerkennung. Hierzu wird das Tool WoEx_M gestartet.

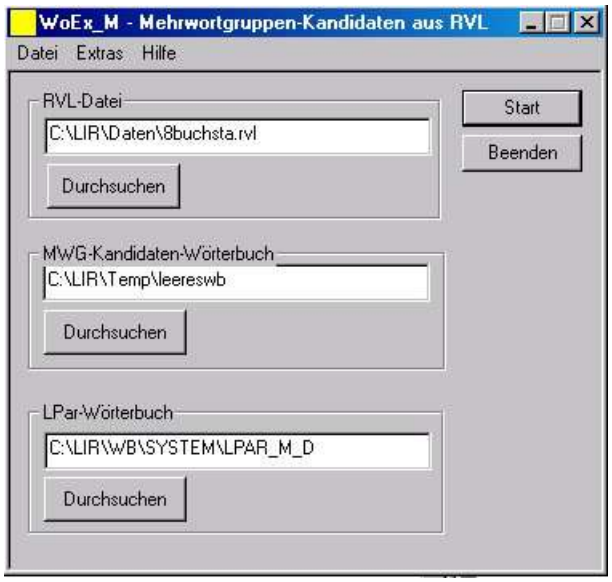


Abb. 35: WoEx_M

Nach dem Programmstart sind drei Angaben zu machen: RVL-Datei (die zuvor erzeugte), MWG-Kandidaten-Wörterbuch (ein leeres Relationenwörterbuch) und Lpar-Wörterbuch (ein Regelwörterbuch, welches die syntaktischen Muster beinhaltet, zu verwenden ist LPAR_M_D). Unter Extras kann noch ein Relationenwörterbuch (für die Analyse, hier: Relneu) angegeben werden.

Nach dem Programmdurchlauf sollte ein neuer Indexierungslauf mit IDXWin gestartet werden. Das neu erzeugte Mehrwortbegriffswörterbuch sollte dabei verwendet werden.

Im nächsten Schritt werden die indexierten Daten in der RVL-Datei einer statistischen Gewichtung unterzogen.

3.4 Statistische Gewichtung der gewonnenen Daten mit RVL2DB

Mit RVL2DB werden die durch IDXWin erzeugten Indexterme statistisch gewichtet.

Nach dem Programmstart erwartet RVL2DB drei Angaben: eine Eingabe-Datei, d.h. eine RVL-Datei; das zu verwendende Relationenwörterbuch (Datei-Endung: IND) und ein leeres Wörterbuch, in das geschrieben wird. Die Art und Weise der statistischen Gewichtung kann vom Anwender aus den fünf verfügbaren Gewichtungsalgorithmen ausgewählt werden. Dies geschieht durch editieren der Datei RVL2DB.ini, im Verzeichnis C:\LIR\Tools\RVL2DB.

Abb. 36: INI-Datei für RVL2DB

In der INI-Datei werden die wichtigsten Einstellungen für RVL2DB festgehalten. Die Auswahl des zu verwendenden Gewichtungsalgorithmus erfolgt in dem mit [HfkGew] beginnenden Abschnitt. Hier

sind die entsprechenden Algorithmen mit den dazugehörigen Nummern aufgelistet. Beispiel: Für die Inverse Document Frequency (IDF) wird die Zahl 4 in der Zeile „Gewicht=“ eingesetzt. Nach der Editierung der INI-Datei kann diese geschlossen werden. Nun kann RVL2DB gestartet werden.

Abb. 37: RVL2DB Hauptbildschirm

Die ausgegebene Ergebnisdatei enthält die Gewichtungswerte der Deskriptoren in geschweiften Klammern.

```
0025681* Informationsgesellschaft {000.0164}# 0025680* Interest {000.2500}#
0025679* Bibliothek {000.0023}#
0025677* Notation {000.1250}# Chance {000.0323}#
0025673* Application {000.2500}#
0025672* File {000.3333}# Master {000.3333}#
0025671* Apostroph {000.5000}# Entwicklung {000.0057}#
0025660* Software {000.0500}#
0025659* Informationszentrum {000.0179}# Friedrich-Ebert-Stiftung {000.0159}
0025652* Table {000.2500}# Time {000.1667}#
0025598* Fakt {000.0444}# Vorstellung {000.0286}# Kopf {000.0267}# Welt {000
0025597* Prosa {000.4000}# Poesie {000.4000}# Seitenblick {000.4000}# Elixie
0025594* Fakt {000.0444}# Vorstellung {000.0286}# Bilder {000.0267}# Macht {
0025593* Orden {000.2500}# Verbeugung {000.2500}# Mathematiker {000.1000}# G
0025591* Ferne {000.3333}# Tod {000.1111}# Bot {000.1111}# Hermeneutik {000.
0025590* Feuereifer {000.4000}# Philosoph {000.0800}# Diskurs {000.0444}# Ge
0025551* Normung {000.1176}#
0025544* Portal {000.1667}# Recht {000.0256}# Wissen {000.0033}#
0025543* Surfer {000.3333}# Des {000.0513}#
0025540* Web-Suchwerkzeug {000.6667}# Retrievaltest {000.0667}# Evaluation {
0025539* Kognitionswissenschaft {000.1333}#
```

Abb. 38: Ergebnisdatei von RVL2DB

3.5 Umsetzung der Ergebnisse in ein Speicherformat mit OPAC_STO

Die von IDXWin produzierte RVL-Datei ist für die Nutzung in einem bibliografischen Datenbanksystem nicht geeignet. Bevor der Inhalt der RVL also zurück nach MIDOS importiert werden kann, muss die RVL zunächst in ein gut zu importierendes Datenformat konvertiert werden. Dafür wird das Tool OPAC_STO verwendet.

Nach dem Programmstart von OPAC_STO müssen zwei Dateien benannt werden, die Quell-Datei (also die RVL-Datei, die umgewandelt werden soll) und die Ergebnis-Datei (die zu erzeugende STO-Datei). Außerdem ist anzugeben, welche Wortarten in die Zieldatei übernommen werden sollen (Teil-

wörter, Substantive, Verben, Adjektive, Wortformen). Die Konfiguration von OPAC-STO erfolgt mit der OPAC_STO.cfg, zu finden in C:\LIR\Tools.

```
##################################
# Konfiguration fuer Opac_Sto    #
##################################

##################################
# Trennzeichen nach ID
##################################
# IdDelim = *

##################################
# Trennzeichen zwischen Woertern
##################################
# WordDelim = #

##################################
# Startzeichen fuer ID-Eintrag
##################################
# ID-Start = <

##################################
# Endezeichen fuer ID-Eintrag
##################################
# ID-Ende = >

##################################
# Relationsnummer ausgeben
##################################
# Relation = TRUE

##################################
# Woerter ohne ID ausgeben
##################################
# NoIdWords=TRUE
```

Abb 39.: Konfiguration von OPAC_STO

In dieser CFG-Datei kann festgelegt werden, welches Trennzeichen nach der Identnummer steht (IdDelim, Voreinstellung ist |), welches Trennzeichen zwischen den Wörtern steht (WordDelim, Voreinstellung ist ;), was Anfangs- und Endzeichen für einen Kommentar für Quelldaten sind (ID-Start, ID-Ende) sowie ob die Relationsnummer mit ausgegeben werden soll oder ob Wörter ohne ID ausgegeben werden sollen.

Wenn OPAC_STO gestartet wurde, konvertiert es die RVL-Datei in eine STO-Datei . Diese STO-Datei enthält nun alle Ergebnisse der Indexierung und der statistischen Gewichtung in einem gut zu importierenden Format und wird im nächsten Schritt zurück nach MIDOS importiert. Unter den Einstellungen können Spezifikationen für bestimmte Formate angegeben werden.

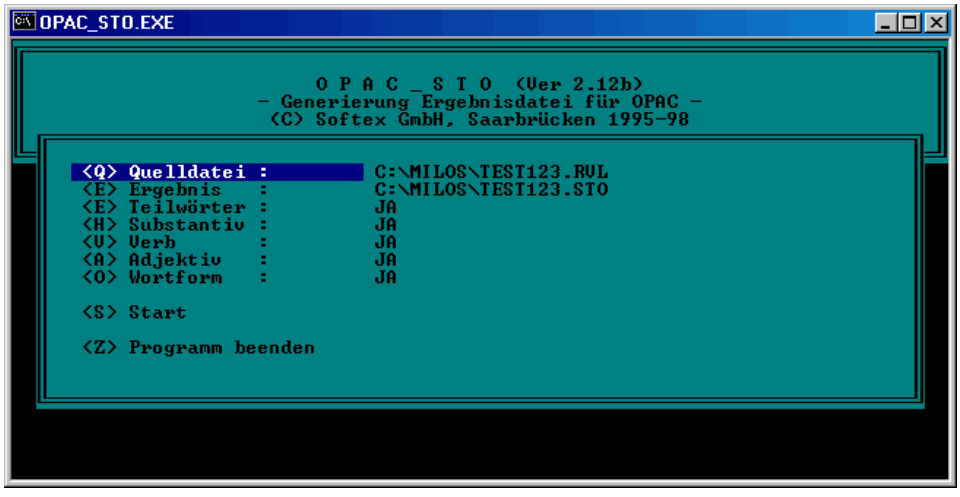


Abb. 40: Parameter von OPAC_STO

3.6 Import der Indexierungsergebnisse in MIDOS

Im nächsten Schritt werden die Indexierungsergebnisse des LIR-Prozesses mit den ursprünglichen bibliografischen Quelldaten verbunden. Dazu ist es notwendig, den Inhalt der zuvor erzeugten STO-Datei nach MIDOS zu importieren.

MIDOS erwartet dazu zwei Datenbanken:

Datenbank 1 : die Datenbank mit den bibliografischen Daten

Datenbank 2: die Datenbank mit den Indexierungsergebnissen

Datenbank 1 existiert bereits. Datenbank 2 muss neu angelegt werden. Dazu wird MIDOS gestartet und unter Datenbank der Punkt „Datei importieren“ angewählt. Es öffnet sich eines der MIDOS-Hilfsprogramme zum Datei-Import. Hier sind nun verschiedene Angaben zu machen: das Format der Quelldatei (Delimited), sowie der Pfad zur Quelldatei (die importiert werden soll) und zur Zieldatei (die angelegt werden soll). Außerdem kann ein Titel für die neue Datenbank vergeben werden. Beim Import ist zu beachten, dass entweder ein Kopfsatz vorhanden sein muss oder eine Feldzuordnung (000; 200 usw.) getroffen werden muss. Als Feldtrenner dient das Pipe-Zeichen (|), als Satzende ist &&& anzugeben.

Ein Satz könnte also wie folgt aussehen:

000 | 100 ; 200 &&&

Wenn beide Datenbanken vorhanden sind, werden sie in einer neuen Datenbank zusammengeführt. Diesen Prozess nennt MIDOS „Mischen“. Dazu wird unter „Funktionen“ der Punkt „Job einrichten“ angewählt. Unter „Einzelprogramm“ und „Daten mischen“ kann das Programm „MISCHDAT“ gestartet werden. In MISCHDAT wird als erster Parameter die 1. Quelldatei angegeben, als zweiter Parameter die 2. Quelldatei, und als dritter Parameter die zu erzeugende Zieldatei. Als Keyfeld dient hier „000“, die Option „copyall“ sollte auf jeden Fall verwendet werden.

Als Endergebnis der LIR-Prozesses existiert dann eine neue Datenbank. Diese enthält die bibliografischen Quelldaten aus der MIDOS-Datei sowie die Ergebnisse der IDX-Indexierung und der RV-L2DB-Gewichtung. Sie kombiniert also bibliografische Daten mit automatisch generierten und gewichteten Deskriptoren.

Anhang

Übersicht über die wichtigsten Wortklassenkodes

00 Briefwort (*Du; Sie*)
01 Wortform nur klein (*aufwärts; ab; allg.*)
02 Wortform nur groß (*GmbH; Bsp.*)
03 Wortform klein oder groß (*Dreierlei/dreierlei; Decresc./decresc.)*
04 Verb stark (*geben*)
05 Verb schwach (*kaufen*)
06 Substantiv feminin Singular (*Karte*)
07 Substantiv maskulin Singular (*Stuhl*)
08 Substantiv neutrum Singular (*Buch*)
09 Adverb (nicht kodieren – stattdessen WK 1)
10 Adjektiv nur klein (*schön; gekauft*)
11 Adjektiv groß oder klein (*englisch/Englisch)*
12 Substantiv maskulin und feminin Singular (*Angestellte*)
13 Verbstamm schwach (*kauf-*)
14 Zahl (*zwei*)
15 Verbstamm stark (*gib-*)
16 Eigenname feminin Singular (*Michaela; Adria*)
17 Eigenname maskulin Singular (*Wolfgang; Apennin)*
18 Eigenname neutrum Singular (*Rom; Word*)
19 Adjektiv-Suffix
20 Infinitiv-Suffix
21 Kundenwort nur klein
22 Kundenwort nur groß
23 Kundenwort klein oder groß
24 Substantiv maskulin, feminin und neutrum Plural
25 blockiertes Morphem.
26 Substantiv feminin Plural. (*Kosten; Mütter*)
27 Substantiv maskulin Plural (*Azzurri; Flüge*)
28 Substantiv neutrum Plural (*Daten; Bücher)*
29 Verbinfinitiv mit 'zu' (*anzugeben*)
30 Präfix
31 Substantiv-Suffix feminin
32 Substantiv-Suffix maskulin
33 Substantiv-Suffix neutrum
35 Substantiv maskulin und feminin Plural (*Grüne*)
36 Eigenname feminin Plural (*Antillen*)
37 Eigenname maskulin Plural (*Donkosaken*)
38 Eigenname neutrum Plural (*Arabische Emirate*)
39 Substantiv-Suffix feminin Plural
40 Substantiv-Suffix maskulin Plural
41 Substantiv-Suffix neutrum Plural
42 Verb-Suffix stark
43 Verb-Suffix schwach
44 Substantiv maskulin und neutrum Singular (*Band*)
45 Substantiv maskulin und neutrum Plural (*Graffiti*)
46 Substantiv feminin und neutrum Singular (*Raclette*)
47 Substantiv feminin und neutrum Plural (*Studien*)
48 Substantiv maskulin, feminin und neutrum Singular (*Süddeutsche*)

Übersicht über die wichtigsten Relationenkodes

```
000 Kompositum -> Headword
001 Synonym
002 Teilwort -> Kompositum
003 Kompositum -> Teilwort (nicht Headword)
004 Teilwortderivation
005 Akronym -> Langform
007 Antonym (alle Richtungen)
008 Sprachvarianten (Deutsch/Schweiz/Österr.)
009 siehe-auch
010 orthograph. Variante
011 Stichwort -> Nichtstichwort
013 (Syntakt.) Homograph
014 Quasi-Synonym
015 Unterbegriff -> Oberbegriff
017 Grundform -> Ablaut/Umlaut
018 Ablaut/Umlaut -> Grundform
019 Wort (Begriff) -> Klassifizierungscode
021 Worterläuterung
022 Einzelwort -> Mehrwort
023 Mehrwort -> Dekompisition (nicht Headword)
024 Mehrwort -> Derivation
026 Homonym (alle Richtungen; mit Semantik-Differenzierung)
029 weiblich -> männlich (Personen, Berufe)
031 Nominativ Plural -> Nominativ Singular
033 Pseudokompositum -> Teilwort
034 Kompositum mit Präfix -> Teilwort
035 Korrektur bei Regionalvariante
036 Rechtschreibfehler -> Korrekte Schreibweise
037 Zusammenbildung -> Element
039 chemische Formel -> Langform
041 Mehrwort kanonisch -> Mehrwort Text
042 Kategorie -> Kategorie
043 Klasse -> Unterklasse
044 Wortform -> Prop-Kategorie
045 Übersetzung
046 Übersetzung (falsch in Quellwort)
047 Regelbeispielgruppe -> Regelnummer
048 Kategorie -> Unterkategorie
049 Kategorie -> antonyme Kategorie
051 Deskriptor (ohne Relationierung)
052 Nichtdeskriptor (ohne Relationierung)
053 Assoziation (alle Richtungen)
054 Teilwort-Übersetzung (Teil im Zielwort)
```

Literatur

Knorz, Gerhard: Automatische Indexierung. In: Wissensrepräsentation und Information-Retrieval. Universität Potsdam 1994. S. 138-198.

Lepsky, Klaus: Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen. Köln 1994. (Kölner Arbeiten zum Bibliotheks- und Dokumentationswesen; Heft 18)

Lepsky, Klaus: Automatische Indexierung zur Erschließung deutschsprachiger Dokumente. In: nfd Information – Wissenschaft und Praxis. 50(1999)

Nohr, Holger: Automatische Indexierung: Einführung in betriebliche Verfahren, Systeme und Anwendungen. Potsdam 2000. (Materialien zur Information und Dokumentation ; Band 13)