Terminologieextraktion von Mehrwortgruppen in kunsthistorischen Fachtexten

Bachelorarbeit

Studiengang Bibliothekswesen
Fakultät für Informations- und Kommunikationswissenschaften
Fachhochschule Köln

vorgelegt von:

Juliane Bredack

am 29.08.2013 bei Prof. Dr. Klaus Lepsky

Abstract

Mit Hilfe eines algorithmisch arbeitenden Verfahrens können Mehrwortgruppen aus elektronisch vorliegenden Texten identifiziert und extrahiert werden.
Als Datengrundlage für diese Arbeit dienen kunsthistorische Lexikonartikel des *Reallexikons zur Deutschen Kunstgeschichte*. Die linguistisch, wörterbuchbasierte Open-Source-Software Lingo wurde in dieser Studie genutzt. Mit Lingo ist es möglich, auf Basis erstellter Wortmuster, bestimmte Wortfolgen aus elektronisch vorliegenden Daten algorithmisch zu identifizieren und zu extrahieren. Die erstellten Wortmuster basieren auf Wortklassen, mit denen die lexikalisierten Einträge in den Wörterbüchern getaggt sind und dadurch näher definiert werden. So wurden individuelle Wortklassen für Fachterminologie, Eigennamen, oder Adjektive vergeben. In der vorliegenden Arbeit werden zusätzlich Funktionswörter in die Musterbildung mit einbezogen. Dafür wurden neue Wortklassen definiert. Funktionswörter bestimmen Artikel, Konjunktionen und Präpositionen. Ziel war es fachterminologische Mehrwortgruppen mit kunsthistorischen Inhalten zu extrahieren unter der gezielten Einbindung von Funktionswörtern. Anhand selbst gebildeter Kriterien, wurden die extrahierten Mehrwortgruppen qualitativ analysiert. Es konnte festgestellt werden, dass die Verwendung von Funktionswörtern fachterminologische Mehrwortgruppen erzeugt, die als potentielle Indexterme weitere Verwendung im Information Retrieval finden können.

Schlagwörter:
Terminologieextraktion, Informationsextraktion, automatisches Indexieren

Multiword groups can be detected and extracted from electronic resources not only manually but also with the help of an algorithmic approach.

For this work, articles of the web version of the *Reallexikon zur Deutschen Kunstgeschichte,* an encyclopaedia for art history, were analysed. For this purpose, the open-source software, Lingo, which is based on linguistic methods and works with pre-defined dictionaries, was used. It is possible to identify and extract multiword groups algorithmically. This is based on the formation of word patterns. These word patterns are constituted of the specifically defined word classes. The different words in the dictionaries are assigned with special word classes, which add information to every entry. Different word classes were found for specialized terminology, proper names or adjectives. In the present study, additionally, function words were included in the formation of word patterns. Therefor, new word classes for these function words were defined. Functions words are articles, conjunctions and prepositions. The aim of this study was to identify and extract specific multiword groups from art historic data with the specific integration of function words in the formation of the word patterns. With self-defined criteria, the found multiword groups were qualitatively assessed.

It could be shown, that the use of function words leads to the generation of specific multiword groups. These groups can be used as potential index terms in an information retrieval. Therefore it is beneficial to include function words in the formation of word patterns for the automatic extraction of multiword groups from art historic articles with Lingo.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

MWG	Mehrwortgruppe
MWGs	Mehrwortgruppen
RDK	Reallexikon zur Deutschen Kunstgeschichte
WB	Wörterbuch
WBs	Wörterbücher

1 Einleitung

Mehrwortgruppen sind als lexikalische Einheit zu betrachten und bestehen aus mindestens zwei miteinander in Verbindung stehenden Begriffen. Durch die Verbindung mehrerer Fachwörter transportieren sie in Fachtexten aussagekräftige Informationen. Sie vermitteln eindeutige Informationen, da aus den resultierenden Beziehungen zwischen den in Verbindung stehenden Fachbegriffen die inhaltliche Bedeutung eines Fachtextes ersichtlich wird. Demzufolge ist es sinnvoll, Mehrwortgruppen aus Fachtexten zu extrahieren, da diese die Inhalte eindeutig repräsentieren. So können Mehrwortgruppen für eine inhaltliche Erschließung genutzt und beispielsweise als Indexterme im Information Retrieval bereitgestellt werden.

Mehrwortgruppen enthalten Informationen eines Textes, die in natürlicher Sprache vorliegen. Zur Extraktion von Informationen eines elektronisch vorliegenden Textes kommen maschinelle Verfahren zum Einsatz, da Sprache Strukturen aufweist, die maschinell verarbeitet werden können.[1] Eine mögliche Methode Mehrwortgruppen innerhalb von elektronisch vorliegenden Fachtexten zu identifizieren und extrahieren ist ein algorithmisches Verfahren. Diese Methode erkennt Wortfolgen durch das Bilden von Wortmustern, aus denen sich eine Mehrwortgruppe in einem Text zusammensetzt. Die Wortmuster repräsentieren somit die einzelnen Bestandteile einer Mehrwortgruppe. Bereits an mathematischen Fachtexten wurde dieses Verfahren untersucht und analysiert. Relevante Mehrwortgruppen, die ein mathematisches Konzept oder mathematischen Inhalt repräsentierten, konnten erfolgreich extrahiert werden. Zum Einsatz kam das Indexierungssystem Lingo, mit dessen Programmodul *sequencer* eine algorithmische Identifizierung und Extraktion von Mehrwortgruppen möglich ist.[2] In der vorliegenden Arbeit wird dieses algorithmische Verfahren unter Einsatz der Software Lingo genutzt, um Mehrwortgruppen aus kunsthistorischen Fachtexten zu extrahieren. Als Datenquelle dienen kunsthistorische Lexikonartikel aus dem *Reallexikon zur Deutschen Kunstgeschichte*, welches in deutscher Sprache vorliegt. Es wird untersucht, ob positive Ergebnisse im Sinne von fachterminologischen Mehrwortgruppen mit kunsthistorischen Inhalten erzeugt werden können. Dabei soll zusätzlich die Einbindung von Funktionswörtern innerhalb einer Mehrwortgruppe erfolgen. Funktionswörter definieren Artikel, Konjunktionen und Präpositionen, die für sich alleinstehend keine inhaltstragende Bedeutung besitzen, allerdings innerhalb einer

[1] Vgl.: Dengel (2012), S. 205, weiterführende Informationen und Verfahren zur Informationsextraktion s. bes. Kapitel 8
[2] Vgl.: Gödert (2012a)

Mehrwortgruppe syntaktische Funktionen erfüllen. Anhand der daraus resultierenden Ergebnisse wird analysiert, ob das Hinzufügen von Funktionswörtern innerhalb einer Mehrwortgruppe zu positiven Ergebnissen führt. Ziel soll es demnach sein, fachterminologische Mehrwortgruppen mit kunsthistorischen Inhalten zu erzeugen, unter Einbindung von Funktionswörtern.

Bei der Extraktion fachterminologischer Mehrwortgruppen wird im Folgenden insbesondere auf die Erstellung von Wortmustern eingegangen, da diese die Basis liefern, mit welchen das Programmmodul *sequencer* Wortfolgen innerhalb der kunsthistorischen Lexikonartikel identifiziert. Eine Einordung der Indexierungsergebnisse erfolgt anhand selbst gebildeter Kriterien, die definieren, was unter einer fachterminologischen Mehrwortgruppe zu verstehen ist.

2 Methodische Grundlagen

Im Kapitel 2.1 wird die Bedeutung von Mehrwortgruppen erläutert, welche Merkmale diese aufweisen und in welcher Form sie aus den Datenquellen extrahiert werden. Für diese Arbeit wird ein algorithmisches Verfahren genutzt. Zum Einsatz kommt das linguistisch basierende System Lingo (s. Kapitel 2.2), mit dem Ziel, Mehrwortgruppen zu extrahieren, die kunsthistorischen Inhalt repräsentieren. Die Extraktion fachterminologischer Mehrwortgruppen erfolgt am Beispiel kunsthistorischer Daten aus dem *Reallexikon zur Deutschen Kunstgeschichte*. (s. Kapitel 2.3).

2.1 Bedeutung von Mehrwortgruppen

Inhaltsrelevante Informationen eines Fachtextes werden durch Fachwörter repräsentiert. Ein Fachterm sollte allerdings nicht für sich alleinstehend betrachtet werden, „[…] sie sind immer im Zusammenhang mit ihrer sprachlichen Umgebung zu sehen. Bei der Textproduktion bereitet die Einbettung des Fachwortes in den Fachtext, d.h. die Wahl der korrekten Verben, Präpositionen usw., vielfach Schwierigkeiten. Damit Fachwörter ihre kommunikative Leistung erbringen können, sind solche zusätzlichen sprachlichen Elemente jedoch unerlässlich.“[3]

Demnach sind Mehrwortgruppen[4] als zusammenhängende lexikalische Einheiten[5] zu sehen, welche einzelne Konzepte wiedergeben bzw. Inhalte repräsentieren. Um eine hohe Aussagekraft und Deutlichkeit zu erreichen, müssen mehrere, miteinander in Verbindung stehende Begriffe in syntaktischer Hinsicht in sich als abgeschlossen angesehen werden, z.B. „Baum der Erkenntnis“ oder „Symbolik und Mythologie der Natur“.

Wesentliche Informationen eines Textes lassen sich durch MWGs vermitteln, da diese mehr Informationen beinhalten und ausdrücken als ein einzelnes Wort.[6] Allein der Lexikonartikel „Christus“[7] des Reallexikons enthält verschiedenste Schlagworte wie „Leben“, „Szene“ oder „Kunst“. Durch den Einsatz von MWGs lassen sich eindeutigere Aussagen zum Inhalt des Artikels treffen, z.B. „dargestellte Szene aus dem Leben Christi“, „Einzug in Jerusalem“, „monumentale Kunst der altchristlichen Zeit“ oder „Szene aus dem Leben Jesu“. Hinzu kommt, dass ein einzelner Begriff verschiedene Inhalte ausdrücken kann. Z.B. kann der Begriff „Zeit“ einen Augenblick, die Dauer oder

[3] Arntz (2004), S. 34
[4] In der vorliegenden Arbeit nachfolgend im Singular abgekürzt mit MWG bzw. Plural MWGs.
[5] Vgl.: Huo (2012), S. 9
[6] Vgl.: ebd., S. 11
[7] Band 3, Spalte 611

einen Tag beschreiben, an dem Beispiel „Kunst der altchristlichen Zeit“ allerdings auf eine Epoche bzw. ein Zeitalter Bezug nehmen. Erst durch die Verbindung mit zusätzlichen Termen innerhalb einer MWG wird der inhaltsrelevante Sinn eines Begriffes sichtbar. MWGs lassen somit eine Bedeutungsdifferenzierung auf der Ebene eines einzelnen Wortes zu. Vor allem im Information Retrieval verbessern sie durch die Aufnahme in den Index die Suchmöglichkeiten. Ein Dokument kann durch den Einsatz einer MWG gezielt mit dieser gefunden werden, anstelle von mehreren Dokumenten, die die einzelnen Begriffe beinhalten und unnötig viele Informationen bereitstellen, die nicht im Sinne des Suchenden sind. Somit wird eine präzisere Suche möglich, da die Mehrdeutigkeit einzelner Indexterme aufgehoben wird.[8]

MWGs können aus unterschiedlich vielen Teilen bestehen, mindestens aber aus zwei sprachlichen Elementen.[9] In der vorliegenden Arbeit erfahren MWGs, welche sich aus drei bis sechs Bestandteilen zusammensetzen, besondere Aufmerksamkeit (s. Kapitel 4.3 und 4.4).

MWGs bestehen nicht nur aus feststehenden Verbindungen von Substantiven oder Adjektiv–Substantiv–Verbindungen, sondern, wie an den bereits genannten Beispielen gezeigt wurde, auch aus Funktionswörtern, wie Präpositionen, Artikeln oder Konjunktionen. Diese vermitteln selbst keine inhaltlichen Informationen, das heißt sie haben keine eigene Bedeutung, besitzen allerdings wichtige syntaktische Funktionen innerhalb einer MWG. Funktionswörter stellen die nötige Verbindung zwischen den inhaltsrelevanten Begriffen her, sodass sich diese zu einer Einheit verbinden und je nach fachlich ausgeprägtem Kontext einen Sachverhalt vermitteln.

2.1.1 Syntaktische Funktionen von Funktionswörtern

Artikel, Präpositionen und Konjunktionen stellen Funktionswörter[10] dar, die in MWGs eine wichtige Rolle spielen können. Die verschiedenen Funktionswörter weisen folgende syntaktische Merkmale auf, weshalb sie bei der Extraktion von MWGs in den Daten des Reallexikons besondere Beachtung finden.

Ein Artikel dient u. a. dazu, das Genus eines Wortes näher zu bestimmen. Infolgedessen wird oft erst durch den Einsatz eines Artikels sichtbar, durch welchen Kasus

[8] Vgl.: Gödert (2012b), S. 286
[9] Vgl.: Arntz (2004), S. 34
[10] Es werden in diesem Abschnitt nur überblicksartig Merkmale der genutzten Funktionswörter und dessen syntaktische Funktionen beschrieben. Für detailliertere Beschreibungen Vgl.: Pittner (2004): Deutsche Syntax und Hoberg (2004): Deutsche Grammatik.

und Numerus ein Begriff bestimmt wird. In der Regel sind es Artikel, die als Begleiter vor einem Substantiv stehen[11], z.B. „Anbetung *der* Könige".

Konjunktionen verbinden in der Regel Sätze, aber auch einzelne Wörter, wie beispielsweise Personen und / oder Sachverhalte miteinander, z.B. „Chor *und* Querschiff der Marienkirche".[12]

Präpositionen setzen Begriffe innerhalb einer MWG miteinander in Beziehung, sodass ein zusammenhängender Inhalt durch die Verbindung mehrerer Begriffe vermittelt wird. Dieser Sachverhalt wird an dem Beispiel „große Bilderhandschrift *von* Wolfram Willehalm" deutlich. Die Präposition „von" kennzeichnet die inhaltliche Verknüpfung zwischen den Begriffen „Bilderhandschrift" und dem Personennamen „Wolfram Willehalm", sodass daraus hervorgeht, dass eine Bilderhandschrift von Wolfram Willehalm stammt. Abhängig von den eingesetzten Begriffen, kann ein und dieselbe Präposition unterschiedliche Beziehungen ausdrücken. So kann z.B. „von" auch ein örtliches Verhältnis beschreiben, wie in der MWG „goldenes Evangelienbuch *von* Echternach". Ein bestimmter Sachverhalt lässt sich allerdings auch durch verschiedene Präpositionen ausdrücken, z.B. „Darstellung *aus* dem Leben Jesu" oder „Darstellungen *zum* Leben Jesu".[13] Präpositionen stehen immer in Verbindung mit einem anderen Wort und bestimmen dessen Kasus. Es sind Substantive, mit denen sie bevorzugt innerhalb einer sogenannten Präpositionalgruppe auftreten, z.B. „Christus am Kreuz".[14]

All diese beschriebenen Merkmale verdeutlichen, dass Funktionswörter eine wichtige Rolle bei der Extraktion von fachterminologischen MWGs einnehmen. Aus diesem Grund werden bei dem Versuch, fachterminologische MWGs zu erzeugen, Artikel, Präpositionen und Konjunktionen bei der Erstellung von Wortmustern zur Extraktion genutzt, um zusammenhängenden kunsthistorischen Inhalte zu extrahieren (s. Kapitel 4.1 und 4.3).

2.2 Automatisches Indexieren mit Lingo

Im Folgenden wird ein algorithmisch basiertes Verfahren vorgestellt, mit dem MWGs aus kunsthistorischen Lexikonartikeln erkannt und extrahiert werden. Die praktische Durchführung erfolgt mit der Open-Source-Software Lingo.[15]

[11] Vgl.: Hoberg (2004), S. 223
[12] Vgl.: ebd. S. 313 f
[13] Vgl.: Srikumar (2013), S. 231
[14] Vgl.: Hoberg (2004), S. 300 f
[15] Software Download unter: http://lex-lingo.blogspot.de/; eingesetzte Version Lingo 1.8.2

Lingo[16] ist ein linguistisch arbeitendes System, welches sich bei der Indexierung auf elektronische Wörterbücher stützt. Es ermöglicht eine Identifizierung von Grundformen aller Zeichenketten[17], damit alle vorkommenden grammatikalischen Varianten von Termen innerhalb eines zu indexierenden Textes erkannt werden. Komposita werden ebenfalls identifiziert und lassen sich in deren Bestandteile zerlegen. Die Erkennung von MWGs erfolgt entweder algorithmisch (*sequencer*) oder lexikalisch (*multiworder*) (s. Kapitel 2.2.3 und 2.2.4). Weiterhin ist eine lexikalische Relationierung von Begriffen möglich.[18]

2.2.1 Indexierungskonfigurationen und Programmmodule von Lingo

Die Indexierung eines elektronisch vorliegenden Textes bzw. einer Dokumentkollektion kann durch zwei Indexierungskonfigurationen erfolgen. Die Konfigurationsdatei lingo.cfg dient u.a. der Analyse kollektionsspezifischer Daten. Nach einer Indexierung können die daraus resultierenden Ergebnisse dafür genutzt werden, um geeignete (abhängig vom Anwendungsfall), kollektionsspezifische Wörterbücher, welche fachspezifisches Vokabular enthalten sollen, aufzubauen.[19] Die Konfigurationsdatei lir.cfg ermöglicht, z.B. durch den Einsatz zuvor erstellter fachspezifischer Wörterbücher, eine fachspezifische Dokumentkollektion zu indexieren. Die generierten Terme können als potentielle Indexterme direkt den Datensätzen einer Dokumentkollektion zugewiesen werden, um sie für ein Retrieval bereitzustellen.[20] Die Indexierungskonfigurationen lingo.cfg und lir.cfg sind in einen Verarbeitungsbereich und einen Ausgabebereich eingeteilt.[21] Abhängig vom Indexierungsvorhaben und den gewünschten Ergebnissen können diese flexibel durch eigene Einstellungen verändert werden.[22]

---

[16] Es soll lediglich ein Überblick auf die Funktionalitäten von Lingo geschaffen werden. Für die vorliegende Arbeit sind insbesondere die Programmmodule sequencer und multiworder relevant. Für umfassende Informationen zur Arbeitsweise von Lingo und dessen Verarbeitung von Dateien, Wörterbuchformaten, statistische Funktionalitäten und Einstellungsmöglichkeiten der Konfigdateien vgl.: Gödert (2012b), Kapitel 5

[17] Lingo verarbeitet Zeichenketten, die z.B. Wörter in einer Textdatei entsprechen können. Weitere Informationen siehe Gödert (2012b), S. 273 f. In der vorliegenden Arbeit werden neben der Benennung Zeichenkette synonym dazu auch die Begriffe Wörter, Terme oder Indexterme gebraucht.

[18] Vgl.: Gödert (2012b), S. 270

[19] Vgl.: Gödert (2012a), S. 5. Wörterbücher können als Textdateien erstellt werden und mit einem Texteditor bearbeitet werden vgl.: Gödert (2012b), S. 282, 304. Zu indexierende Dateien, Konfigurationsdateien und Ergebnisdateien können ebenfalls mit einem Texteditor geöffnet und bearbeitet werden. Zum Nutzen kollektionsspezifischer WBs siehe Kapitel 2.2.2 der vorliegenden Arbeit.

[20] Vgl.: (Gödert 2012b), S. 294

[21] Siehe Abbildungen 4 und 5 der Indexierungskonfigurationen lingo.cfg und lir.cfg als Standardfunktionsbestandteile von Lingo im Anhang.

[22] Vgl.: Gödert (2012b), S. 273

Einzelne Programmodule, auch als *attendees* bezeichnet, realisieren eine Indexierung nach den gewünschten Anforderungen. Im Einzelnen sind dies der *textreader, tokenizer, wordsearcher*, *decomposer*, *multiworder*, *sequencer synonymer*, *abbreviator*, *variator*, *dehyphenizer* und *stemmer*.[23] Die attendees bauen in ihrer Arbeitsweise und erzeugten Ergebnissen aufeinander auf.[24] Schwerpunkt in der vorliegenden Arbeit bilden die attendees *sequencer* und *multiworder* von Lingo. Deren Einsatz besitzt für die praktische Durchführung der Indexierung zur Identifizierung, Extraktion und Weiterverarbeitung von MWGs besondere Relevanz.

2.2.2 Wörterbuchkonzept und Nutzen der Wortklassen für eine Indexierung

Eine Indexierung mit der Software Lingo ist in Bezug auf die zuverlässige Identifizierung aller Zeichenketten innerhalb einer Datei abhängig von den eingebundenen Wörterbüchern.[25] Alle Wörterbücher, die für eine Indexierung genutzt werden, müssen in der dafür vorgesehenen Wörterbuchkonfiguration de.lang und in den eingesetzten attendees der Indexierungskonfigurationen lingo.cfg oder lir.cfg definiert sein.[26] Als Standardwörterbücher stehen Rechtschreibwörterbücher bereit, sowie ein Mehrwortgruppen- und Synonymwörterbuch.[27] Für die Grundformreduzierung werden die bereits genannten Rechtschreibwörterbücher eingesetzt. Sie enthalten Einträge in folgender Form:

```
darstellung=darstellung #s
abendländisch=abendländisch #a
```

Links vom Gleichheitszeichen befinden sich alle Wortformen und rechts davon die zugeteilte Grundform.[28] Die Abkürzungen #s und #a sind die Bezeichnungen für die Wortklassen. In diesem Fall steht das S für Substantive und das A für Adjektive. Somit befinden sich in den Rechtschreibwörterbüchern Einträge von Begriffen, die immer mit einer individuell vereinbarten Wortklasse gekennzeichnet sind. Wortklassen können nicht nur die in der deutschen Sprache gebräuchlichen, wie Substantive, Adjektive oder Verben sein. Zum Zwecke einer Indexierung können unterschiedlichste Wortklassen definiert werden, z.B. können Eigennamen oder Fachwörter eine gesonderte Kenn-

[23] Detaillierte Erläuterungen zu den einzelnen attendees vgl.: Gödert (2012b), Kapitel 5
[24] Vgl.: ebd. S. 272
[25] Vgl.: ebd. S. 271
[26] Vgl.: ebd. S. 283
[27] Siehe Abbildung 6 im Anhang. Standardwörterbuchkonfigurationen de.lang.
[28] Vgl.: Gödert (2012b), S. 275

zeichnung erhalten, wie es auch in der vorliegenden Arbeit praktiziert wird (s. Kapitel 4.2.3). Die angegebenen Wortklassen spielen für eine Indexierung mit dem attendee *sequencer* eine weitere wichtige Rolle (s. Kapitel 2.2.3).

Die Einträge der Wörterbücher werden während eines Indexierungslaufes mit denen aus der zu indexierenden Datei abgeglichen. Für eine zuverlässige Identifizierung eines flektierten Begriffs kommt zusätzlich eine Suffixliste zum Einsatz. In dieser können für jede genutzte Wortklasse in den Wörterbüchern Suffixe (Wortendungen) definiert werden, damit grammatikalische Varianten einer Zeichenkette innerhalb eines Textes erkannt werden können. Die Einstellungen an der Suffixliste werden in der Wörterbuchkonfiguration de.lang vorgenommen.[29]

Für eine erfolgreiche Identifizierung und Generierung von Indextermen einer fachspezifischen Dokumentkollektion, wie in der vorliegenden Arbeit, reichen die systemintegrierten Standardwörterbücher von Lingo nicht aus. Diese sind hinsichtlich des fachspezifischen Vokabulars nicht vollständig und würden bei der Indexierung eines Fachtextes zu vielen nicht erkannten Wörtern führen. Die systemintegrierten Wörterbücher werden bei einer Neuinstallation von Lingo überschrieben und sollten deswegen nicht mit eigenen Einträgen erweitert werden.[30] Die Indexierung einer fachspezifischen Kollektion erfordert den Aufbau und die Nutzung von Wörterbüchern mit fachspezifischen Vokabular und dem Einsatz selbst definierter Wortklassen, wie es auch in der vorliegenden Arbeit geschehen ist (s. Kapitel 4.2.2 und 4.2.3).

2.2.3 Indexierung mit dem attendee Sequencer

Das Kennzeichnen von lexikalisierten Wörtern mit ihrer Wortklasse wird als tagging bezeichnet. Es spielt eine besondere Rolle, da dadurch Wissen über die Wortarten und deren Bedeutung bereitgestellt wird.[31] Somit sind die Wortklassen das entscheidende Instrument für den *sequencer* zur Erkennung von fachterminologischen MWGs. Wortklassen definieren einen lexikalisierten Begriff in seiner Funktion in den erstellten Lingo Wörterbüchern. Durch individuell festgelegte Wortklassen kann beispielsweise eine Abgrenzung zwischen Fachterminologie und Gemeinsprache erfolgen oder eine Unterscheidung zwischen Substantiven, Adjektiven oder Verben vorgenommen werden, wie sie in der deutschen Sprache vorkommen. MWGs weisen syntaktische

[29] Vgl.: ebd., S. 275 f
[30] Vgl.: ebd., S. 283, 303. Nicht erkannte Wörter werden in der non-Datei ausgegeben s. Gödert (2012b), S. 279, 302, für weitere Ausführungen zum Aufbau neuer Wörterbücher vgl.: Gödert (2012b), Kapitel 5.3.7
[31] Vgl.: Dengel (2012), S. 211 und weiterführende Informationen zum *Part of Speech Tagging* im Rahmen der Informationsextraktion, Kapitel 8.2.1

Strukturen auf, welche sich im deutschen Sprachgebrauch wiederholen (s. Kapitel 2.1.1 und 4.3). Deshalb können durch die Erstellung von sequences[32] bestimmte Wortfolgen aus einem Text algorithmisch erkannt und extrahiert werden. Die Wortmuster basieren im Falle des attendee *sequencer* von Lingo auf den Abkürzungen der Wortklassen, wie sie auch in den Rechtschreibwörterbüchern zur Grundformidentifizierung zum Einsatz kommen. Somit ist der Aufbau und Einsatz der Wörterbücher und der für jeden Eintrag definierten Wortklassen auch zur Nutzung des *sequencer* notwendig. Nur so kann der *sequencer* auf das vorhandene Wissen über die Wörter zurückgreifen. Die Ausgabe der MWGs nach erfolgreicher Identifizierung durch den *sequencer* erfolgt in der Grundform, z.B. „allegorisch auslegung auf sündenfall“ und werden mit der Wortklasse q ausgewiesen.
Die Wortmuster werden in der Wörterbuchkonfiguration de.lang unter den Parametern des *sequencer* angegeben.

```
sequences: [ [AAS, "1 2 3"], [SS, "1 2"] ]
```

Dieses Beispiel verdeutlicht, dass eine Wortkombination, welche in der Form Adjektiv-Adjektiv-Substantiv und Substantiv-Substantiv-Verbindungen in einem Text auftreten, erkannt und extrahiert werden sollen. Des Weiteren lässt sich festlegen, wie die erkannten MWGs ausgegeben werden, also beispielsweise in der Reihenfolge (Adjektiv Adjektiv Substantiv oder Substantiv Substantiv), wie sie auch in dem zu indexierenden Text vorkommen. Das wird durch die Zahlenkombinationen "1 2 3" bzw. "1 2" zwischen den Anführungszeichen realisiert.[33]
Nach einer algorithmischen Identifizierung von MWGs durch den *sequencer* ist nicht davon auszugehen, dass diese immer die Merkmale einer MWG aufweisen.[34] Der Nutzen von MWGs im Retrieval liegt u.a. darin, diese zusätzlich als Suchmöglichkeit in einer Dokumentkollektion bereitzustellen. Nach der Indexierung einer fachspezifischen Kollektion unter Einsatz des attendee *sequencer* ist es demnach erforderlich, die erzeugten MWGs vorab intellektuell, hinsichtlich ihrer fachspezifischen Qualität, zu sichten. Da die MWGs durch den *sequencer* in ihrer Grundform erzeugt werden, ist es notwendig eine manuelle Überarbeitung an den MWGs vorzunehmen, sodass sie in einer grammatikalisch korrekten Variante vorliegen, z.B. „allegorische auslegung auf

[32] Als synonyme Benennungen von Sequences werden in der vorliegenden Arbeit die zusätzlichen Begrifflichkeiten Wortmuster oder Muster genutzt.
[33] Vgl.: Gödert (2012b), S. 287
[34] Eigenschaften von MWGs siehe Kapitel 2.1; Kriterien, welche für die vorliegende Arbeit eine fachterminologische MWG definieren siehe Kapitel 4.1

sündenfall“.[35] Die aus den erforderlichen Arbeitsschritten heraus selektierten MWGs können in ein Mehrwortgruppenwörterbuch übertragen und durch Einsatz des attendee *multiworder* weiterverarbeitet werden (s. Kapitel 2.2.4).[36]

2.2.4 Indexierung mit dem attendee Multiworder

Der attendee *multiworder* erlaubt es, Wörterbücher mit lexikalisierten Mehrwortgruppen in folgender Form aufzubauen:

> wort und bild in der graphik
> zusammengekommenes wissen über engel und dämon
> ärztlicher schutzpatron in der bildenden kunst

Mehrwortgruppen können zusammenhängend in ihrer grammatikalisch korrekten Form eingetragen werden, ohne das für jedes einzelne Wort die Wortklasse angegeben werden muss. Dies ist möglich, indem jeder einzelne Begriff, aus denen sich eine MWG zusammensetzt, durch das Hinzuschalten eines zusätzlichen Wörterbuchs bei einer Indexierung identifiziert wird. In diesem Wörterbuch ist jeder Term mit seiner entsprechenden Wortklasse lexikalisiert.
In der Wörterbuchkonfiguration de.lang ist als Standard das systemintegrierte Mehrwortgruppenwörterbuch sys-mul[37] festgelegt, welches zur Identifizierung einzelner Terme auf das Rechtschreibwörterbuch sys-dic zurückgreift.[38]

sys-mul: { name: de/lingo-mul.txt, txt-format: SingleWord, use-lex: 'sys-dic', def-wc: m }

Die beschriebenen Arbeitsschritte zur Erstellung eines Mehrwortgruppenwörterbuchs sind notwendig (s. Kapitel 2.2.3) und lassen sich nicht vermeiden. Nur so ist es möglich, dass keine Abstriche hinsichtlich der Qualität eines aufgebauten Mehrwort-gruppenwörterbuchs und dessen erzeugten Ergebnisse hingenommen werden müssen. Die Indexierung einer Dokumentkollektion mit einem aufgebauten Mehrwort-gruppenwörterbuch durch die Einbindung in den attendee *multiworder* mit der Indexierungskonfiguration lir.cfg, realisiert eine Bereitstellung der erkannten MWGs als

[35] Lingo unterscheidet nicht zwischen Groß-und Kleinschreibung, weswegen die Einträge in den WBs in Kleinschreibung vorgenommen werden, vgl.: Gödert (2012b), S. 276
[36] Vgl.: Gödert (2012b) S. 288
[37] Bezeichnung der Wörterbücher, wie sie in der Wörterbuchkonfiguration de.lang definiert sind.
[38] Vgl.: Gödert (2012b), S. 286

Indexterme (s. Kapitel 4.5). Die MWGs werden als Indexterme den jeweiligen Datensätzen einer Dokumentkollektion zugeordnet.[39] Somit wird das Suchen und Finden von Dokumenten mit dem Einsatz von Mehrwortgruppen möglich, wenn diese in einem weiteren Arbeitsschritt in eine entsprechende Datenbank integriert werden.

2.3 Reallexikon zur Deutschen Kunstgeschichte – RDK

Das *Reallexikon zur Deutschen Kunstgeschichte*, im Weiteren abgekürzt mit RDK, ist ein Nachschlagewerk zur Realienkunde der Kunstgeschichte, welches seit 1937 erscheint.[40] Inhaltliche Schwerpunkte der Lexikonartikel liegen auf Architektur, Bildende Künste, Kunsthandwerk, Materialien, Technik und Ikonographie des deutschen Sprachgebiets. Herausgeber ist das Zentralinstitut für Kunstgeschichte in München. Die Bände des RDK erscheinen als fortsetzendes Lieferwerk.[41]

Die Artikel des RDK weisen eine hohe Zahl an kunsthistorischen Sachbegriffen und Eigennamen (Personennamen, Ortschaftsbezeichnungen bzw. Geografika) auf. Sie bieten spezifische Fachinformationen zu einem Stichwort. Abkürzungen und Sonderzeichen, wie Punkte, Striche oder Klammern sind häufig eingesetzte Mittel (s. Abbildung 1). Die Länge der Artikel variieren[42], was typische Merkmale von Lexikonartikeln als Fachtextsorte darstellen.[43]

[39] Vgl.: Gödert (2012b), S. 296
[40] Für detailliertere Informationen zur Geschichte des RDK vgl.: Augustyn (2004)
[41] Vgl.: Homepage: Zentralinstitut für Kunstgeschichte: Forschungsstelle Realienkunde / Reallexikon zur Deutschen Kunstgeschichte
[42] Vgl.: Lepsky (2006), S. 171
[43] Vgl.: Fandrych (2011), S. 90 f

Zu Sp. 696 und 710: Attribut des Christus Iudex

5. Meister Francke, um 1400, Hamburg.

Demuth Ihm das Passwort underschreibt" (*Biblisches Engel- und K.werck,* Augsb. 1694 [Ndr. Portland, Or. 1972], Bl. 28).

Ein Kupferstich von Jos. Seb. und Joh. Bapt. Klauber nach Gg. Balth. Probst zeigt den Engel neben dem Zugang zum Himmel, der durch die Fürsprache Mariens – die selbst als „Ianua coeli" angerufen wird – offensteht (Lauretanische Litaney. In 57 Kupffer-Stichen ... fürgestellt und ... erkläret von *Franc. Xaver Dornn,* Augsb. 1749 [benutzte Ausg.: Augsb. 1771], Nr. 40; [50] Bl. 92; auf einer weiteren Illustration zur Lauretanischen *Litanei bewacht der Engel den „Hortus conclusus": *F. X. Dornn* a.a.O. Nr. 16; [50] Bl. 86).

In einem Kupferstich zur „Pietas Eucharistica" des Hauses Österreich von Karel Skréta (Entw.) und Barth. Kilian (Stich), dat. 1668, steht vor den Stufen zu einem Palast, in dessen Mittelfenster im Piano Nobile eine Monstranz ausgestellt ist, ein Engel mit F. Er bewacht den Zugang zum Allerheiligsten [61, S. 63 Abb. 33].

B. In Bildern zu alchimistischen Texten konnte das F. den Zugang zum Garten der Philosophen („Rosarium philosophorum") bewachen.

In dem von Raphaël Baltens, gen. Custos, gestochenen Lehrbild zur Tinkturgewinnung, Taf. 4 der bei Steffan Michelpacher verlegten „Cabala", sind zwei gekreuzte F. vor dem Einlaß aufgerichtet (*Cabala, Spiegel der K. unnd Natur: in Alchymia,* Augsb. 1615; lat. Ausg. Augsb. 1616 u. ö.). Im philosophischen Garten knien Sol und Luna zu seiten eines Lebensbrunnens, aus dem ihnen Christus den aus seiner Seitenwunde gespeisten Trank als Zeichen der „Multiplication" reicht (gemeint ist die „Vermehrung der Tinktur" als „letzte Stufe des Prozesses": *Emil Ernst Ploss u. a.,* Alchimia, Ideologie und Technologie, Mchn. 1970, S. 163).

V. Von Deutungen abhängige Darstellungen.

A. Eigenschaften.

Anlaß zu bildlicher Wiedergabe von F. boten manche der ihm zuerkannten Eigenschaften und die mit diesen zu erreichenden Wirkungen (zu solchen, die man in der exegetischen Literatur gewöhnlich dem F. beilegte, wie z. B. die Eigenschaft des Feuers, beweglich – „versatilis" – zu sein, vgl. Ambrosius [3], Gregor [16], Paterius [19], Beda Venerabilis [20], Alkuin [23], Glossa ordinaria [26, Sp. 97], Ps.-Hugo von St-Victor [31a], Andreas von St-Victor [32], Aelred von Rievaulx [32a, Sp. 292], *Lauretus* [42] S. 496). Da dem F. die Kräfte des Schwerts und des Feuers zu eigen sind, konnten sowohl dem einen wie dem anderen zuerkannte Deutungen wirksam werden. Ging man von der Wirkung des Feuers aus, war diese gesteigert durch die Kraft des Schwertes, betonte man die letztere, war sie vermehrt um jene des Feuers. Dies ermöglichte die Darstellung eines F. auch dort, wo die von Schwert oder Feuer allein unzulänglich gewesen wäre ([59a]; vgl. Sp. 706).

1. Wegen seiner Schärfe konnte das F. schneiden und brennen: Nach *Picinelli* befreit das F. die arme Seele mittels der Schneide von irdischen Verstrickungen, während das Feuer des F. die Seele peinigt ([43] lib. XXII cap. 9 Nr. 58, Bd. 2 S. 213).

In einer Zeichnung des Maarten van Heemskerck (Düsseldorf, K.mus.) und einem Heemskerck folgenden, 1554 datierten Stich des Cornelis Bos führt Christus selbst das F., das die Stricke durchtrennt, mit denen das Herz des Sünders an die Welt gefesselt ist (*Abb. 8;* vgl. *Heinz Peters,* Psychomachia. Zu einer unbekannten Zchg. von H. und ihren ikonol. Voraussetzungen, in: *Fs. Eduard Trautscholdt ...,* Hbg. 1965, S. 90–100).

Im fünften Blatt des „Triumphus Patientiae"-Zyklus von M. van Heemskerck (Entw.) und Dirk Volkertsz. Coornhert (Ausf. 1559) führt der auf einer Schildkröte reitende Iob in seinem Banner ein F.; dieses scheidet einen Globus von einer Waage, die an einem geflügelten Herzen hängt ([51] Bd. 4 S. 230 Nr. 155; *Ilja M. Veldman,* M. v. H. and Dutch humanism in the sixteenth c., Maarsen 1977, S. 66 Abb. 43). Das F. ist wohl Bild für die Entsagung Iobs, der sich von aller Verfangenheit in Weltliches gelöst hat.

Veritas durchtrennt mit einem F. den Strick, mit dem der „Discipulus Christi" an „Mors", „Diabolus", „Caro" und „Peccatum" gebunden ist (Kupferstich des Dirck Volkertsz. Coornhert [†1590] nach Adriaan de Weert: *ders.,* De Wereld tussen Goed en Kwaad, s'Gravenhage 1990, S. 100f. mit Abb.; *Henricus Oraeus,* Aeroplastes Theo-Sophicus sive Eicones mysticae ..., Ffm. 1620, Emblem 74 „Veritas liberabit nos"; zum F. als Attribut von „Veritas" s. Sp. 726).

Abbildung 1: Seite aus dem RDK, Spalten 701, 702 aus Band 9, Artikel „Flammenschwert"

2.3.1 RDK-Web

Bis zum Jahr 2007 lag das RDK nur als Printversion vor. Durch den unvollständigen Charakter eines Fortsetzungswerkes gab es bis dahin keine ausreichenden Recherchemöglichkeiten, wie Register, um gezielt auf die Inhalte der Artikel zurückzugreifen. Lediglich die Stichwörter und Querverweise als Vernetzung zwischen den Lexikonartikeln boten Navigationsmöglichkeiten.[44] Nach der Beendigung eines Erschließungsprojektes, welches durch das Zentralinstitut und der Fachhochschule Köln realisiert wurde, entstand eine Online Version des RDK.[45] RDK-Web bietet die Möglichkeit, umfassend auf kunsthistorische Informationen zurückzugreifen.

Die Realisierung einer RDK-Web-Variante erfolgte durch den Einsatz automatischer Indexierung. Ziel war es, Zugriffspunkte auf die Lexikonartikel zu schaffen, um eine gezielte Suche auf deren Inhalte zu ermöglichen. Sucheinstiege sind u. a. über die Stichworte möglich und durch die von der Indexierung geschaffenen Sach-, Personen- und Ortsregister.

Für die Erzeugung dieser Register mittels automatischer Indexierung mit Lingo, wurden umfangreiche fachspezifische Wörterbücher zur Identifizierung relevanter Einträge aufgebaut. Zu diesem Zwecke wurden diverse kunstgeschichtliche Quellen herangezogen.[46] Für kunsthistorische Fachterminologie, Personennamen, Ortschaftsbezeichnungen und Mehrwortgruppen wurden separate Wörterbücher erstellt. Zusätzlich zu diesen wurden noch Synonymwörterbücher, für eine Synonymrelationierung angefertigt, um die Zugriffspunkte auf die Lexikonartikel auszuweiten.[47]

Das aus der damaligen Indexierung hervorgegangene Datenmaterial (s. Kapitel 3) konnte als Ausgangsbasis zur Weiterverarbeitung genutzt werden und fand Eingang in die Extraktion von Mehrwortgruppen in der vorliegenden Arbeit.

[44] Vgl.: Lepsky (2006) , S. 171
[45] Vgl.: Homepage: Über das RDK – RDK Web
[46] Vgl.: Lepsky (2006), S. 173 f
[47] Vgl.: ebd. S. 174, zum Nutzen einer Synonymrelationierung vgl.: ebd. S 174 ff und Gödert (2012b), S. 288 ff

3 Ausgangslage und Nutzung der vorhandenen RDK–Daten

Das vorliegende Ausgangsmaterial der damaligen RDK-Web-Indexierung besitzt für die angestrebte algorithmische Extraktion von Mehrwortgruppen durch den *sequencer* eine besondere Relevanz. Die Wörterbücher mit umfangreich erstellten Einträgen kunsthistorischer Sachbegriffe, Personennamen und Geografika konnten ohne inhaltliche Erweiterung der bestehenden Einträge nachgenutzt werden. Der Zeitaufwand, welcher in die Erstellung kollektionsspezifischer Wörterbücher investiert werden muss, blieb bei der Umsetzung des eigenen Vorhabens erspart.

3.1 RDK-Web-Datenmaterial

In den folgenden Kapiteln wird erläutert, in welcher Form die RDK-Daten und die mit der damalig in Verbindung stehenden Einstellungen der Software Lingo (Wörterbücher, Wortklassen, Konfigurationsdateien) für die aktuelle Indexierung vorlagen. Diese Daten können für die vorliegende Arbeit genutzt, aber auch Anpassungen müssen vorgenommen werden. Es wird beschrieben welche Probleme mit einer direkten Nachnutzung für eine neue Indexierung zur Mehrwortgruppenextraktion aufgetreten sind.

3.1.1 RDK-Lexikonartikel als Datenbank

Die RDK–Lexikonartikel sind mit ihren Textinhalten und den zugehörigen Band- und Spaltenangaben in der Datenbank seite.dbm hinterlegt, welche aus der RDK-Web-Indexierung hervorgegangen ist. Mit dem Datenbankmanagementsystem MIDOS 6 wird auf die Daten zugegriffen und für die aktuelle Indexierung zur Weiterverarbeitung genutzt (s. Kapitel 4.2.1).[48] Die Band- und Spaltenangaben dienen gleichzeitig als Dokumentnummern. Der Datenbankausschnitt seite.dbm ist für eine direkte Nachnutzung besonders relevant. Die darin hinterlegten Textinhalte der Lexikonartikel konnten für die Indexierung mit dem *sequencer* direkt herangezogen werden. Durch die angegebene Dokumentnummer wird eine Generierung und eine direkte Zuordnung der fachterminologischen MWGs nach einer Indexierung mit der Indexierungs-

[48] Softwaredownload unter: http://www.progris.de/index.html?/midost.htm (Download einer kostenlosen Demoversion)

konfiguration lir.cfg und dem darin aktivierten attendee *multiworder* zu den jeweiligen Artikeln möglich.

3.1.2 RDK-Web-Wörterbücher

Zu den damals festgelegten Wörterbüchern gehörten Rechtschreibwörterbücher zur Identifizierung kunsthistorischer Sachbegriffe, Personen, Orten sowie Abkürzungen[49]. Diese wurden bei der damaligen Indexierung für die Grundformreduktion herangezogen. Mehrwortgruppenwörterbücher für die Identifizierung von MWGs und Synonymwörterbücher für die Relationierung von Synonymen kamen ebenfalls zum Einsatz. Die damals aufgebauten Mehrwortgruppen- und Synonymwörterbücher spielen für das aktuelle Indexierungsvorhaben allerdings keine Rolle. Die damals genutzten Mehrwortgruppenwörterbücher wurden beispielsweise intellektuell und aus externen Ressourcen heraus zusammengestellt, z.B. zur Identifizierung von Werktiteln aus der Kunst oder von Personennamen.[50] Eine erneute Indexierung durch den Einsatz des *sequencer* hat u.a. das Ziel, neue Einträge durch eine algorithmische Identifizierung von MWGs zu schaffen, neben den bereits bestehenden.
Als direkte Nachnutzung für das geplante Indexierungsvorhaben standen die Rechtschreibwörterbücher im Vordergrund, da diese lexikalisierte Einträge mit getaggten Wortklassen enthalten. Diese werden für das geplante Indexierungsvorhaben zur Identifizierung von Wortfolgen durch den *sequencer* weitere Verwendung finden (s. Kapitel 4.2.2). Im Einzelnen kommen folgende Wörterbücher für die weitere Nutzung in Frage:

- rdk-usr: enthält kunsthistorische Sachbegriffe
- rdk-us3: enthält kunsthistorische Sachbegriffe und Einträge, die eine ungewollte Kompositazerlegung verhindern sollten
- per-big: enthält alle Personennamen (Vor- und Nachname)
- rdk-ort: enthält alle Ortsnamen bzw. Geografika
- dic-sys: Allg. Rechtschreibwörterbuch, welches umfangreich überarbeitete Einträge des damals vorhandenen Lingo-Systemwörterbuchs enthält.[51] Des Weiteren sind Adjektive, weitere Fachbegriffe, Substantive und Verben lexikalisiert**.**

[49] Abkürzungen können bei einer Indexierung mit dem attendee abbreviator erkannt werden. In der vorliegenden Arbeit findet dieser allerdings keinen Einsatz. Auf die Identifizierung von Abkürzungen innerhalb einer MWG durch das Einbinden mit einer extra Wortklasse wird verzichtet.
[50] Vgl.: Lepsky (2006) S. 173 f
[51] WB dic-sys ist nicht das vorinstallierte Lingo-Systemwörterbuch.

3.1.3 Wortklassen der RDK-Web-Indexierung

Die lexikalisierten Begriffe in den unterschiedlich genutzten Wörterbüchern wurden mit einer Vielzahl verwendeter Wortklassen gekennzeichnet. Im Folgenden sind das die Wortklassen:

- A: Adjektive (sowohl Gemeinsprache als auch kunsthistorische Terminologie)
- B: Sachbegriffe (kunsthistorische Terminologie, nur Sachbegriffe)
- E: Eigennamen (kunsthistorische Terminologie, sowohl Sachbegriffe, Personennamen und Ortsnamen)
- N: Vorname einer Person
- O: Ortsnamen bzw. Geografika
- P: Nachname einer Person
- R: Nachname einer Person ohne Suffixe[52]
- S: Substantive (Gemeinsprache)
- T: take it as it, ohne Suffixe, Einträge, unabhängig davon ob es sich um Sachbegriffe, Personen, Orte, Gemeinsprache etc. handelt
- V: Verben
- W: Wortform (z.B. Funktionswörter, Sachbgriffe, etc.), ohne Suffixe
- Z: Ortsnamen bzw. Geografika ohne Suffixe

Das Tagging von Begriffen in den Wörterbüchern mit einer Vielzahl individuell definierter Wortklassen hatte für die vollständige Indexierung der RDK-Lexikonartikel diverse Gründe. Im Kapitel 3.2 wird darauf eingegangen, welche Probleme für die aktuelle Indexierung damit verbunden waren. Dies hatte Einfluss auf die weitere praktische Umsetzung des eigenen Verfahrens zur Identifizierung von MWGs unter Zuhilfenahme von Funktionswörtern.

3.1.4 Lingo-Konfigurationsdateien der RDK-Web-Indexierung

Die in der damalig verwendeten Wörterbuchkonfiguration de.lang definierten Wörterbücher sowie auch die damals erstellte bzw. mit Angaben für die individuellen

[52] Ohne Suffixe bedeutet, dass diese Einträge, in ihrer vorliegenden Form nicht verändert werden, z.B. Namenseintrag „Peters“, welcher nicht auf den Namen „Peter“ relationiert wird, sondern beibehalten wird. Gleicher Sachverhalt findet sich auch in den restlichen WBs wieder, allerdings durch die Nutzung anderer Wortklassen.

Wortklassen erweiterte Suffixliste dienten als Ausgangsbasis zur Weiterverarbeitung. Die festgelegten Wortendungen der einzelnen Wortklassen für die Grundform-identifizierung dienten auch der aktuellen Indexierung (s. Kapitel 4.2.4.1). Für die angestrebte algorithmische Identifizierung von MWGs wurden eigene Einstellungen an den Standardkonfigurationen von Lingo (lingo.cfg, lir.cfg und de.lang) vorgenommen und diese als Ausgangsbasis weiterverwendet (s. Kapitel 4.2.4).

3.2 Probleme einer direkten Nachnutzung des RDK–Datenmaterials

Die Wörterbucheinträge, welche in Vorbereitung auf eine vollständige RDK-Web-Indexierung erstellt wurden, waren nicht durchgängig und konsequent in den dafür vorgesehenen Wörterbüchern lexikalisiert und zudem nicht mit einheitlichen genutzten Wortklassen getaggt. Das lässt sich bereits an obiger Aufzählung der vielfältigen Wortklassen erkennen. Orts- und Personennamen, welche die Wortklassen O/Z (Orte) und P/R bzw. N (Namen) in den dafür festgelegten Wörterbüchern per-big und rdk-ort erhielten, lexikalisierte man z.T. auch im allgemeinen Rechtschreibwörterbuch dic-sys und kennzeichnete sie durch die Wortklasse E. Weiterhin wurden in dem Wörterbuch dic-sys ebenfalls Orts- und Personennamen lexikalisiert, die sich nicht in den dafür vorgesehenen Wörterbüchern (per-big bzw. rdk-ort) wiedergefunden haben. Dieses Problem zeigt sich auch an den kunsthistorischen Sachbegriffen. Diese wurden im Top-Rechtschreibwörterbuch rdk-usr nicht nur mit der Wortklasse B lexikalisiert sondern auch mit einem E für Eigenname gekennzeichnet. Zusätzlich fanden sich fachterminologische Sachbegriffe in dic-sys mit der Wortklasse E. Die aus der RDK-Web-Indexierung hervorgegangenen Wortklassen T und W erwiesen sich insbesondere als problematisch, da mit diesen sowohl Fachterminologie, Gemeinsprache, als auch Personen, Orte und Funktionswörter getaggt waren. Bei den Wortklassen T und W fand somit keine klare Abgrenzung, einheitliche Nutzung und damit Kennzeichnung von Begriffen statt. Eine Umarbeitung und Streichung der Wortklassen T und W und die Verteilung der damit gekennzeichneten Begriffe auf die restlichen Wortklassen wurde nicht durchgeführt.[53] Erstens ist es fachlich schwer einschätzbar bzw. nicht immer ersichtlich, ob ein Begriff fach- oder gemeinsprachliche Verwendung finden sollte oder ob ein Begriff als Orts- oder Personenname definiert werden muss. Zweitens ist dieser Arbeitsschritt sehr zeitintensiv und für den Umfang der vorliegenden Arbeit nicht möglich.

[53] Weitere Verwendung von Begriffen mit den Wortklassen T und W, siehe Kapitel 4.2.3

Für die damalige RDK-Indexierung zur Erstellung einer RDK-Web-Variante mit nützlichen Zugriffspunkten auf die Inhalte der Lexikonartikel waren, z.B. die individuellen Kennzeichnungen fachspezifischer Terme durch Verwendung differenzierter Wortklassen, damit diese Eingang in ein Register fanden, notwendige Arbeitsschritte.[54] Für die Indexierung mit dem attendee *sequencer*, dessen Erkennung von Wortfolgen auf Wortmustern basiert, sind diese Schritte hinderlich. Eine Vielzahl von Wortklassen erschwert das Erstellen von Wortmustern und die Kombinationsmöglichkeiten derer wären nicht überschaubar.

Aufgrund dessen wurden Maßnahmen getroffen, mit denen die kollektionsspezifischen Wörterbücher und die darin vergebenen Wortklassen für den aktuellen Einsatzzweck angepasst wurden.

[54] Durch die Verwendung unterschiedlicher Wortklassen, kann eine konkrete Zuordnung der Indexterme für bestimmte Anwendungszwecke erfolgen.

4 Automatische Extraktion von Mehrwortgruppen

Für die Extraktion von MWGs in der vorliegenden Arbeit werden eigene Kriterien gebildet, die definieren, was unter einer fachterminologischen MWG mit kunsthistorischem Aspekt zu verstehen ist. Dies ist relevant, um eine Beurteilung der erzeugten Indexierungsergebnisse des attendee *sequencer* vorzunehmen. Anhand dessen kann geprüft werden, ob die algorithmische Identifizierung von MWGs in kunsthistorischen Fachtexten unter Einbindung von Funktionswörtern in den Wortmustern zu positiven Ergebnissen führt.

4.1 Kriterien fachterminologischer Mehrwortgruppen

Im Kapitel 2.1 wurde bereits beschrieben, welche allgemeinen Merkmale eine MWG kennzeichnet. Zum Zwecke der Extraktion von MWGs, die kunsthistorische Inhalte repräsentieren, werden Kriterien erstellt, durch die sich eine fachterminologische MWG auszeichnet. Nach einer erfolgreichen Identifizierung des *sequencer* werden die extrahierten MWGs anhand der aufgestellten Kriterien geprüft und weiterverarbeitet, um sie für den Aufbau eines Mehrwortgruppenwörterbuchs zu verwenden (s. Kapitel 4.5).

Damit eine MWGs als fachterminologisch bezeichnet werden kann, muss sie zunächst das Kriterium der Abgeschlossenheit erfüllen. Das bedeutet MWGs müssen als zusammenhängende Einheit extrahiert werden und nicht als unvollständige Satzfragmente vorliegen. Zum Beispiel stellen „abgesehen von Kloster und Wallfahrtskirche können“ oder „abendländisch von Anfang“ keine repräsentablen MWGs dar. Durch ihre Unvollständigkeit wird kein kunsthistorischer Inhalt repräsentiert, der z.B. für ein Retrieval in Frage kommen würde.

Alle weiteren Merkmale, welche Mehrwortgruppen als positiv definieren, müssen auch das Kriterium der Abgeschlossenheit bzw. Vollständigkeit erfüllen.

Ein weiteres Kriterium für fachterminologische MWGs bilden Kombinationen ab, welche sich aus Personennamen oder Ortschaftsbezeichnungen, spezifiziert durch zusätzliche Begriffe zusammensetzen, zum Beispiel „Barocke Freskomalerei in Schlesien“ oder „Holzschnitt von Georg Lemberger“. Diese weisen das Merkmal auf, kunsthistorische Inhalte gut wiedergeben und beschreiben zu können. Die Erzeugung eines einzelnen Personennamens bzw. Ortes wäre dazu noch nicht in der Lage. Außerdem existieren bereits Personen- und Ortsregister für einen gezielten Zugriff auf die Lexikonartikel des RDK und deren Inhalte.

MWGs, welche sich aus zwei Fachtermen zusammensetzen, sind ebenfalls als fachterminologisch anzusehen, wenn sich durch deren Verknüpfung Inhalte veranschaulichen lassen. Diese sind dafür sehr gut geeignet, da sie bereits auf der Wortebene kunsthistorische Terminologie repräsentieren, wie durch die MWGs „Wallfahrt und Pfarrkirche Maria“ oder „Sündenfall und Erlösung Christus“ deutlich wird. In diesem Zusammenhang werden auch MWGs als positiv erachtet, wenn sich diese aus Begriffen zusammensetzen, die allgemeinerer Natur sind, jedoch durch Fachtermini präzisiert werden, wie „Verständnis der biblischen Symbolsprache“ zeigt.

4.2 Nutzung der RDK-Daten für die Extraktion fachterminologischer Mehrwortgruppen

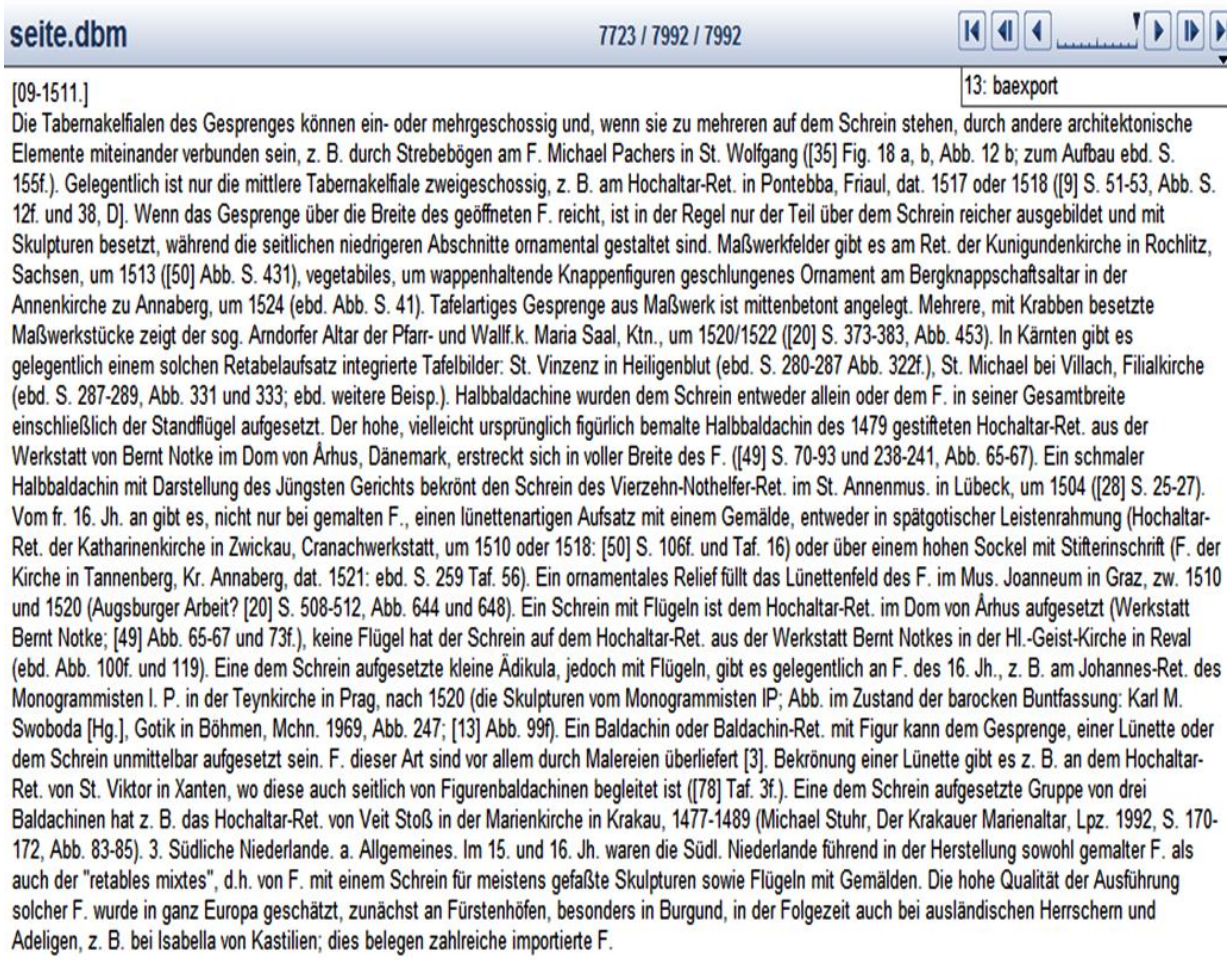
In den folgenden Kapiteln wird die praktische Umsetzung der durchgeführten Indexierungen mit den attendees *sequencer* und *multiworder* der Software Lingo beschrieben. Im Kapitel 4.2.1 wird erläutert, welche Arbeitsschritte notwendig waren, um die Textinhalte der RDK-Lexikonartikel für eine Indexierung zu nutzen. Um eine Indexierung nach den aktuellen Erfordernissen durchführen zu können, mussten die Einstellungen an den Konfigurationsdateien (Wörterbuch- und Indexierungskonfigurationen) von Lingo verändert werden (s. Kapitel 4.2.4.1 und 4.2.4.2). Weiterhin wird dargestellt, welche Wörterbücher für die Indexierungen zum Einsatz kamen und welche individuell vereinbarten Wortklassen erstellt und zur Musterbildung genutzt wurden (s. Kapitel 4.2.2 und 4.2.3). Im Kapitel 4.3 wird beschrieben, welche Kriterien der Musterbildung zu Grunde lagen. Im darauffolgenden Abschnitt 4.4 wird erläutert, mit welchen Wortmustern die Sequencerindexierung zur Identifizierung von MWGs durchgeführt wurde und in welchen Dateien die dadurch extrahierten MWGs vorliegen. Zuletzt werden die abschließenden Arbeitsschritte einer Indexierung mit dem attendee *multiworder* beschrieben (s. Kapitel 4.5).

4.2.1 Weiterverarbeitung der RDK-Daten mit Midos 6

Die RDK-Lexikonartikel, welche als Grundlage zur Identifizierung und Extraktion fachterminologischer MWGs mit dem attendee *sequencer* dienten, wurden aus der Datenbank seite.dbm exportiert (s. Kapitel 3.1.1). Mit dem Datenbankmanagementsystem Midos 6 wurde auf die Datenbank zugegriffen. Den Artikeln sind die jeweiligen Band- und Spaltenangaben des RDK zugeordnet. Alle Spaltenangaben eines Lexikon-

artikels sind mit den jeweiligen Inhalten in einem eigenen Datensatz hinterlegt, d.h. die Inhalte eines kompletten Artikels liegen nicht vollständig in einem Datensatz vor. Um eine Indexierung zu realisieren, wurde eine Textdatei mit den Inhalten der Lexikonartikel erstellt. Für diesen Zweck wurde mit Midos 6 ein neues Ausgabeformat angelegt, in dem lediglich die Textinhalte und die zugehörige Dokumentnummer (Band- und Spaltenangabe) der Lexikonartikel angegeben sind (s. Abbildung 2), ohne Angabe der dafür verwendeten Feldnamen aus der Datenbank. Die Inhalte der zusätzlich vorhandenen Feldkategorien[55], die in dem Datenbankausschnitt seite.dbm angegeben waren, spielen für die Indexierung mit dem *sequencer* keine Rolle. Allein die Textinhalte sind relevant für die Indexierung und die Extraktion von MWGs.[56] Die Dokumentnummer ist für die spätere Indexierung mit dem attendee *mulitworder* relevant, weil durch Verwendung der Konfigurationsdatei lir.cfg, erzeugte MWGs den Dokumenten zugeordnet werden können. Das genutzte Ausgabeformat hat die Bezeichnung baexport. Die neu erstellte Datei basetall.txt, welche nach dem Datenexport im Textformat vorliegt (s. Kapitel 2.2.1), enthält alle Artikel des RDK von Band 1 – Artikel „A-O“ bis zum Band 9 - Artikel „Flügelretabel“, die für das geplante Indexierungsvorhaben genutzt wurden.

[55] Kategorien: Deskriptoren, Sachbegriffe, Personen und Orte, welche den Lexikonartikeln nach der damaligen RDK-Web-Indexierung zugeordnet wurden.
[56] Bezeichnung der Feldkategorien werden von Lingo indexiert, falls sie nicht entfernt werden, vgl.: Gödert (2012b), S. 296


seite.dbm 7723 / 7992 / 7992

[09-1511.] 13: baexport

Die Tabernakelfialen des Gesprenges können ein- oder mehrgeschossig und, wenn sie zu mehreren auf dem Schrein stehen, durch andere architektonische Elemente miteinander verbunden sein, z. B. durch Strebebögen am F. Michael Pachers in St. Wolfgang ([35] Fig. 18 a, b, Abb. 12 b; zum Aufbau ebd. S. 155f.). Gelegentlich ist nur die mittlere Tabernakelfiale zweigeschossig, z. B. am Hochaltar-Ret. in Pontebba, Friaul, dat. 1517 oder 1518 ([9] S. 51-53, Abb. S. 12f. und 38, D]. Wenn das Gesprenge über die Breite des geöffneten F. reicht, ist in der Regel nur der Teil über dem Schrein reicher ausgebildet und mit Skulpturen besetzt, während die seitlichen niedrigeren Abschnitte ornamental gestaltet sind. Maßwerkfelder gibt es am Ret. der Kunigundenkirche in Rochlitz, Sachsen, um 1513 ([50] Abb. S. 431), vegetabiles, um wappenhaltende Knappenfiguren geschlungenes Ornament am Bergknappschaftsaltar in der Annenkirche zu Annaberg, um 1524 (ebd. Abb. S. 41). Tafelartiges Gesprenge aus Maßwerk ist mittenbetont angelegt. Mehrere, mit Krabben besetzte Maßwerkstücke zeigt der sog. Arndorfer Altar der Pfarr- und Wallf.k. Maria Saal, Ktn., um 1520/1522 ([20] S. 373-383, Abb. 453). In Kärnten gibt es gelegentlich einem solchen Retabelaufsatz integrierte Tafelbilder: St. Vinzenz in Heiligenblut (ebd. S. 280-287 Abb. 322f.), St. Michael bei Villach, Filialkirche (ebd. S. 287-289, Abb. 331 und 333; ebd. weitere Beisp.). Halbbaldachine wurden dem Schrein entweder allein oder dem F. in seiner Gesamtbreite einschließlich der Standflügel aufgesetzt. Der hohe, vielleicht ursprünglich figürlich bemalte Halbbaldachin des 1479 gestifteten Hochaltar-Ret. aus der Werkstatt von Bernt Notke im Dom von Århus, Dänemark, erstreckt sich in voller Breite des F. ([49] S. 70-93 und 238-241, Abb. 65-67). Ein schmaler Halbbaldachin mit Darstellung des Jüngsten Gerichts bekrönt den Schrein des Vierzehn-Nothelfer-Ret. im St. Annenmus. in Lübeck, um 1504 ([28] S. 25-27). Vom fr. 16. Jh. an gibt es, nicht nur bei gemalten F., einen lünettenartigen Aufsatz mit einem Gemälde, entweder in spätgotischer Leistenrahmung (Hochaltar-Ret. der Katharinenkirche in Zwickau, Cranachwerkstatt, um 1510 oder 1518: [50] S. 106f. und Taf. 16) oder über einem hohen Sockel mit Stifterinschrift (F. der Kirche in Tannenberg, Kr. Annaberg, dat. 1521: ebd. S. 259 Taf. 56). Ein ornamentales Relief füllt das Lünettenfeld des F. im Mus. Joanneum in Graz, zw. 1510 und 1520 (Augsburger Arbeit? [20] S. 508-512, Abb. 644 und 648). Ein Schrein mit Flügeln ist dem Hochaltar-Ret. im Dom von Århus aufgesetzt (Werkstatt Bernt Notke; [49] Abb. 65-67 und 73f.), keine Flügel hat der Schrein auf dem Hochaltar-Ret. aus der Werkstatt Bernt Notkes in der Hl.-Geist-Kirche in Reval (ebd. Abb. 100f. und 119). Eine dem Schrein aufgesetzte kleine Ädikula, jedoch mit Flügeln, gibt es gelegentlich an F. des 16. Jh., z. B. am Johannes-Ret. des Monogrammisten I. P. in der Teynkirche in Prag, nach 1520 (die Skulpturen vom Monogrammisten IP; Abb. im Zustand der barocken Buntfassung: Karl M. Swoboda [Hg.], Gotik in Böhmen, Mchn. 1969, Abb. 247; [13] Abb. 99f). Ein Baldachin oder Baldachin-Ret. mit Figur kann dem Gesprenge, einer Lünette oder dem Schrein unmittelbar aufgesetzt sein. F. dieser Art sind vor allem durch Malereien überliefert [3]. Bekrönung einer Lünette gibt es z. B. an dem Hochaltar-Ret. von St. Viktor in Xanten, wo diese auch seitlich von Figurenbaldachinen begleitet ist ([78] Taf. 3f.). Eine dem Schrein aufgesetzte Gruppe von drei Baldachinen hat z. B. das Hochaltar-Ret. von Veit Stoß in der Marienkirche in Krakau, 1477-1489 (Michael Stuhr, Der Krakauer Marienaltar, Lpz. 1992, S. 170-172, Abb. 83-85). 3. Südliche Niederlande. a. Allgemeines. Im 15. und 16. Jh. waren die Südl. Niederlande führend in der Herstellung sowohl gemalter F. als auch der "retables mixtes", d.h. von F. mit einem Schrein für meistens gefaßte Skulpturen sowie Flügeln mit Gemälden. Die hohe Qualität der Ausführung solcher F. wurde in ganz Europa geschätzt, zunächst an Fürstenhöfen, besonders in Burgund, in der Folgezeit auch bei ausländischen Herrschern und Adeligen, z. B. bei Isabella von Kastilien; dies belegen zahlreiche importierte F.

Abbildung 2: RDK–Datensatz Band 9, Spalte 1511, Artikel „Flügelretabel“, Ausgabeformat baexport als Ausgangsbasis für die Sequencerindexierung

4.2.2 Weiterverarbeitung und Nutzung der Wörterbücher

Die Wörterbücher, welche der angestrebten Indexierung als Ausgangsbasis zur Verfügung standen, konnten aus den bereits beschriebenen Gründen nicht direkt nachgenutzt werden (s. Kapitel 3.2). Grundsätzlich kamen zur Nutzung nur die Rechtschreibwörterbücher der damaligen RDK-Web-Indexierung in Frage, da diese alle lexikalisierten Einträge mit Wortklassen enthielten. Für eine Indexierung mit Lingo können die Wörterbücher als einfache Textdateien abgespeichert werden (s. Kapitel 2.2.1) und die Wörterbücher z.B. auch unkompliziert mit eigenen Einträgen erweitert werden. Zur Erstellung neuer Wörterbücher wurde der Texteditor notepad++ genutzt.[57]

[57] Software Download unter http://notepad-plus-plus.org/download/v6.4.3.html

Die vorhandenen Wörterbücher wurden für die aktuelle Indexierung mit dem attendee *sequencer* angepasst und folgendermaßen genutzt.

Fachrechtschreibwörterbuch brd-dic

Die Wörterbücher, welche kunsthistorische Einträge enthalten, sind für die Nachnutzung besonders relevant, da fachterminologische MWGs mit kunsthistorischen Inhalten durch den *sequencer* erkannt und extrahiert werden sollen. Die aus der RDK-Web-Indexierung entstandenen Wörterbücher und deren fachterminologischen Einträgen aus rdk-usr, rdkus3, per-big und rdk-ort (kunsthistorische Sachbegriffe, Personen- und Ortsnamen), wurden zusammengeführt und in einem neu erstellten Wörterbuch hinterlegt - brd-dic.

Funktionswortwörterbuch fkt-dic

Für das Einbinden von Funktionswörtern in den sequences wurde ein neues Wörterbuch angelegt in dem Artikel, Konjunktionen und Präpositionen lexikalisiert sind. Somit wurden die aktuell genutzten Funktionswörter einheitlich, mit jeweils individuellen Wortklassen in einem neuen Wörterbuch zusammengefasst – fkt-dic.

Allgemeines Rechtschreibwörterbuch dic-sys

Das aus der RDK-Web-Indexierung hervorgegangene Rechtschreibwörterbuch und dessen Einträge wurden ohne Veränderungen in die aktuelle Indexierung übernommen.

4.2.3 Erstellung und Verwendung der Wortklassen

Zur Identifizierung von Wortfolgen mittels Wortmuster gebildet durch Wortklassen, mussten die individuell erstellten Wortklassen der damaligen RDK-Web-Indexierung abgeändert werden. Ausgehend von den neu erstellten Wörterbüchern brd-dic und fkt-dic wurden die Einträge mit neuen Wortklassen versehen, um die Erstellung von Wortmustern zu erleichtern und deren Kombinationsmöglichkeiten zu begrenzen.
Die im fachterminologischen Wörterbuch brd-dic lexikalisierten Einträge wurden einheitlich mit der Wortklasse E getaggt. Die vormals verwendeten Wortklassen P, N und R, mit denen Personennamen gekennzeichnet waren, wurden komplett gestrichen.

Genauso verhält es sich mit den Wortklassen O und Z der Ortsnamen. Die Trennung der kunsthistorischen Sachbegriffe, welche vorab auch mit der Wortklasse B gekennzeichnet waren, wurde aufgehoben. Sie bekamen ebenfalls das E als neue Wortklasse zugeteilt.

Artikel, Präpositionen und Konjunktionen erhielten jeweils eine eigene Wortklasse und wurden im Wörterbuch fkt-dic lexikalisiert. Artikel sind mit dem R, Präpositionen mit einem C und Konjunktionen mit der Wortklasse U gekennzeichnet. Die differenzierte Kennzeichnung der Wortklassen soll den Nutzen jener speziellen Wortarten innerhalb einer MWG verdeutlichen. Weiterhin ließ dieses Vorgehen einen gezielteren Einsatz innerhalb der Musterbildung zu, sodass deren syntaktischer Zweck erfüllt wurde (s. Kapitel 4.3.3).

Die im allgemeinen Rechtschreibwörterbuch dic-sys lexikalisierten Adjektive blieben mit der Wortklasse A erhalten und wurden für die Musterbildung genutzt (s. Kapitel 4.3.3). Die in dic-sys mit der Wortklasse E enthaltenen fachterminologischen Sachbegriffe, Orts- und Personennamen (s. Kapitel 3.2) verblieben in diesem. Sie bestehen zusätzlich neben den Einträgen aus brd-dic.[58] Die mit den Wortklassen S, T, V und W gekennzeichneten Begriffe (s. Kapitel 3.1.3) aus dem Wörterbuch dic-sys wurden zur Bildung von Komposita genutzt. Somit wurden Komposita nicht nur aus fachterminologischen Termen (gekennzeichnet durch Wortklasse E) oder Adjektiven (Wortklasse A) erkannt und gebildet. Die aus der Indexierung heraus entstehenden Komposita sind mit der Wortklasse K gekennzeichnet (s. Kapitel 4.3.2). Die Wortklassen S, T, V und W fanden für die Sequencerindexierung und damit verbundene Erstellung von Wortmustern keine weitere Verwendung.

Für die Erstellung von Wortmustern kamen die Wortklassen A-Adjektive, C-Präposition, E-Fachbegriffe, K-Komposita, R-Artikel und U-Konjunktionen zum Einsatz (s. Kapitel 4.3).

[58] Eine Übertragung der mit Wortklasse E gekennzeichneten Einträge von dic-sys in brd-dic wurde nicht durchgeführt, da diese bei einer Indexierung identifiziert werden, unabhängig davon, in welchem WB sie lexikalisiert sind, solange die entsprechende WBs in die Konfigurationsdateien eingebunden sind.

4.2.4 Lingo Indexierungskonfigurationen

In den folgenden Kapiteln wird erläutert, welche Konfigurationsdateien von Lingo genutzt und welche Einstellungen vorgenommen wurden, um eine Indexierung mit den attendees *sequencer* und *multiworder* durchzuführen.[59]

4.2.4.1 Wörterbuchkonfiguration de.lang

Die neu erstellten Rechtschreibwörterbücher brd-dic und fkt-dic wurden in die Wörterbuchkonfiguration de.lang von Lingo hinzugefügt, damit diese mit ihren Einträgen für die Indexierungen genutzt werden konnten. Das allgemeine Rechtschreibwörterbuch dic-sys wurde aus der damals verwendeten RDK-Web-Konfigurationsdatei ohne Änderung in die für die aktuelle Indexierung genutzte de.lang übernommen (s. Abbildung 3).

```
#    Funktionswörterbuch
fkt-dic: { name: de/funkw-dic.txt, txt-format: WordClass, separator: '=' }

#    Mehrwortgruppenwörterbuch
ba-mul: { name: de/bamul-dic.txt, txt-format: SingleWord, use-lex: 'ba-dic', def-wc: m }

RDK-Wörterbücher
brd-dic: { name: de/bardk-dic.txt, txt-format: WordClass, separator: '=' }
ba-dic: { name: de/bardkall-dic.txt, txt-format: WordClass, separator: '=' }
dic-sys: { name: de/rdk_dic.txt, txt-format: WordClass }
```

Abbildung 3: Ausschnitt Wörterbuchkonfiguration de.lang

Suffixliste

Die Suffixliste der RDK-Web-Indexierung wurde als Ausgangsbasis genutzt und für den aktuellen Einsatzzweck angepasst (s. Kapitel 3.1.4).
Nachdem die Wortklassen B, P, O und N durch E ersetzt wurden, mussten die aus der RDK-Web-Indexierung erstellten Wortendungen der Klassen B, P, N und O zu E hinzugefügt werden, wenn sie noch nicht vereinbart waren (s. Abbildung 4). Somit wurde eine zuverlässige Grundformidentifizierung der mit Wortklasse E gekennzeichneten

[59] Über das Zusammenwirken der Konfigurationsdateien de.lang und lingo.cfg bzw. lir.cfg von Lingo Vgl.: Gödert (2012b), S. 283, 305 f

Terme möglich. Weitere definierte Wortendungen zur Grundformidentifizierung bestanden für die Wortklassen A (Adjektive), S (Substantive) und V (Verben).[60]

```
suffix:
  #    Suffixliste, Stand: 30-06-2005
  #    Suffixklasse:   s = Substantiv, a = Adjektiv, v = Verb, e = Fachbegriff, f = Fugung
  #    Suffixe je Klasse:   "<suffix>['/'<ersetzung>][ <suffix>['/'<ersetzung>]]"
  #    neue Klassen: c=präposition, r=artikel, u=konjunktion
  #    weitere Suffix-/Wortklassen: w=wortform t=stopwort
  - [s, "e en er ern es n s se sen ses"]
  - [a, "este ste ster sten stes ester estes esten e em en er ere eren erer eres es erem"]
  - [v, "e/en en/en est/en et/en st/en t/en te/en ten/en eten/en ete/en etest/en s"]
  - [e, "s - e en er ern nen es n s se sen ses ' 's s- en- n- isch ischen ische ischer isches
  - [f, "s n e en es er ch/che /en"]
```

Abbildung 4: Ausschnitt Suffixliste in de.lang

Sequencer-Parameter

Die Einstellungen an den Parametern des *sequencer* werden ebenfalls in der Wörterbuchkonfiguration de.lang vorgenommen. Die erstellten Wortmuster für die Indexierung wurden in den Parametern des *sequencer* festgelegt. Um zu realisieren, dass die Muster zur jeweiligen MWG mit in der dafür vorgesehenen Ergebnisdatei ausgegeben werden (s. Kapitel 4.4), mussten diese innerhalb der Anführungszeichen mit angeben werden (s. Abbildung 5).[61]

```
sequencer:
  sequences: [ [ERAK, "ERAK 1 2 3 4"], [KCAE, "KCAE 1 2 3 4"], [AKUE, "AKUE 1 2 3 4"]
```

Abbildung 5: Ausschnitt sequencer-Parameter in de.lang; Indexierung mit vierteiligen Wortmustern

4.2.4.2 Indexierungskonfiguration ba-rdk-ling.cfg

Die Indexierungen zur Identifizierung und Extraktion von MWGs durch den *sequencer* wurden mit der neu erstellten Indexierungskonfiguration ba-rdk-ling.cfg durchgeführt. Als Ausgangsdatei wurde die Standardindexierungskonfiguration von Lingo - lingo.cfg

[60] Wortklasse F=Fugung, Nutzen siehe hierzu Gödert (2012b), S. 281
[61] Weitere Ausführungen zur Ausgabe der MWGs, siehe Kapitel 4.4

genutzt. In ba-rdk-ling.cfg wurden alle nötigen Veränderungen vorgenommen, die eine Durchführung der Indexierung ermöglichten. Die Sequencerindexierung mit der Konfigurationsdatei ba-rdk-ling.cfg erzeugt zunächst eine Datei mit extrahierten MWGs, wie sie durch die angegebenen Wortmuster in den Parametern des sequencer der Wörterbuchkonfiguration de.lang festgelegt werden.[62] Mit Hilfe der erstellten Ergebnisdateien aller erkannten MWGs erfolgte deren Analyse, um MWGs ggf. in ein Mehrwortgruppenwörterbuch zu übertragen. Eine Indexierung mit der Konfigurationsdatei ba-rdk-ling.cfg und der daraus resultierenden Dateien diente deswegen ausschließlich der Ergebnisanalyse von erzeugten MWGs. So konnte u.a. die Hypothese geprüft werden, ob das Einbinden von Funktionswörtern in eine MWG zu positiven Ergebnissen führt. Eine Zuordnung potentieller MWGs zu den jeweiligen RDK-Datensätzen erfolgte durch eine zweite Indexierung mit dem *multiworder* durch die Konfigurationsdatei ba-rdk-lir.cfg (s. Kapitel 4.5)

Attendees wordsearcher, decomposer, multiworder

Die zur Indexierung mit dem *sequencer* genutzten Wörterbücher brd-dic, fkt-dic und dic-sys wurden für die Grundformidentifizierung aller Begriffe durch den attendee *wordsearcher* in ba-rdk-ling.cfg vereinbart. Die angegebene Reihenfolge der Wörterbücher bestimmt, wann ein Begriff zur Identifizierung herangezogen wird.[63] Zunächst wurden alle fachterminologischen Terme aus brd-dic bei der Grundformidentifizierung abgeglichen und die Funktionswörter aus fkt-dic vor dic-sys genutzt. Dies stellte beispielsweise sicher, dass die Funktionswörter mit den neu erstellten Wortklassen aus fkt-dic als erstes erkannt wurden und nicht die mit der Wortklasse W getaggten Terme aus dic-sys. Zur Identifizierung und Bildung von Komposita dienten die Wörterbücher brd-dic und dic-sys, welche für diesen Zweck im attendee *decomposer* festgelegt wurden (weitere Erläuterungen zum *decomposer* s. Kapitel 4.3.2). Attendees, die für die spezifische Indexierung mit dem *sequencer* keine Rolle spielten, wurden durch das Rautesymbol deaktiviert, wie z.B. der attendee *multiworder* oder *synonymer*.[64] Die Ausgabe der Indexierungsergebnisse erfolgte nach der Identifizierung und Extrahierung der MWGs im attendee *sequencer*, durch die Angabe out: syn (s. Abbildung 6).

[62] Welche Dateien erzeugt wurden s. Kapitel 4.4
[63] Nach erfolgreicher Identifizierung einer Zeichenkette bricht Lingo den Wörterbuchabgleich ab und bearbeitet die nächsten Einträge Vgl. hierzu Gödert (2012), S. 284, 308
[64] Synonymer findet keine Verwendung für die vorliegende Arbeit. Nutzen des attendees synonymer Vgl.: Gödert (2012b), S. 288 ff

```
    # Verbleibende Token im Wörterbuch suchen
    - word_searcher:    { source: brd-dic fkt-dic dic-sys, mode: first }

    # Schreibweisen variieren und erneut suchen
#   - variator:         { source: sys-dic }

    # Bindestrichergänzungen rekonstruieren
#   - dehyphenizer:     { source: sys-dic }

    # Wortstämme für nicht erkannte Wörter einfügen
#   - stemmer:          { }

    # Nicht erkannte Wörter auf Kompositum testen
    - decomposer:       { source: brd-dic dic-sys }

    # Mehrwortgruppen im Strom erkennen
    # multi_worder:     { source: ba-mul, out: syn }

    # Wortsequenzen anhand von Regeln identifizieren
    - sequencer:        { stopper: PUNC,OTHR, out: syn }

    # Relationierungen einfügen
    # synonymer:        { skip: '?,t', source: sys-syn, out: syn }
```

Abbildung 6: Ausschnitt Verarbeitungsbereich ba-rdk-ling.cfg mit vorgenommenen Einstellungen für Indexierung mit sequencer

4.3 Kriterien der Wortmusterbildung und deren Bedeutung für die Extraktion von Mehrwortgruppen

Für die Extraktion von Mehrwortgruppen ist die Bildung der sequences erforderlich. Diese bestimmen, welche Wortfolgen in den RDK-Lexikonartikeln erkannt werden sollen, um potentielle fachterminologische MWGs zu extrahieren. Weiterhin wurden Kriterien gebildet, die die Basis zur Musterbildung darstellen.

Bei der Bildung von Wortmustern zur Extrahierung von MWGs kommt die Wortklasse E als Zusammenfassung aller Fachbegriffe zum Einsatz. Die Wortklassen C, R und U dienen der Einbeziehung von Funktionswörtern (C=Präpositionen, R=Artikel, U=Konjunktionen). Das A wird als Wortklasse von Adjektiven und K als Abkürzung von Komposita genutzt.

Zum Zwecke einer Mehrwortgruppenextraktion durch den *sequencer* wurden Wortmuster gebildet, welche sich aus mindestens drei, maximal aus sechs Bestandteilen zusammensetzen. Dieses Vorgehen bot die Möglichkeit eine Vielzahl

von MWGs unterschiedlicher Länge, unter Berücksichtigung aller zu Verfügung stehender Wortklassen und deren als sinnvoll zu erachtender Kombinationen, zu bilden und zu testen.

4.3.1 Kunsthistorische Fachbegriffe – Wortklasse E

Der Wortklasse E kommt innerhalb der Musterbildung eine besondere Bedeutung zu. Mit dieser Klasse ist kunsthistorische Fachterminologie in den für die Indexierungen genutzten Wörterbüchern brd-dic und dic-sys gekennzeichnet. Kunsthistorische Fachterminologie umfasst Sachbegriffe, Personennamen und Ortschaftsbezeichnungen bzw. Geografika, welche in den Artikeln des RDK in hoher Anzahl auftreten (s. Kapitel 2.3). Somit ist die Einbindung der Wortklasse E entscheidend zur Erzeugung fachterminologischer MWGs. Jene Fachterminologie weist allerdings nicht immer klare Grenzen zwischen Gemein- und Fachsprache auf.[65] Ein einzelner Begriff kann, abhängig vom verwendeten Kontext verschiedene Bedeutungen besitzen. Begriffe, wie „Darstellung“, „Geburt“ oder „Leben“, würde man nach einer ersten Analyse auf der Wortebene der Gemeinsprache zuordnen. Der fachliche Hintergrund wird allerdings innerhalb einer MWG, wie z.B. „Geburt Christi“, „Leben Jesu“ oder „Darstellung der Geburt“ deutlich. Bei der Erzeugung der RDK–Web–Variante wurden Begriffe dieser Art mit den Wortklassen E oder B getaggt (s. Kapitel 3) und somit als fachterminologisch ausgewiesen. Die für die aktuelle Indexierung einheitlich festgelegte Wortklasse E kunsthistorischer Terminologie schließt für die aktuelle Indexierung mit dem attendee *sequencer* demnach nicht aus, dass auch MWGs erkannt und extrahiert werden, die offenbar nicht signifikant fachterminologisch zuzuordnen sind. Die MWG „Frau am Grab“ verdeutlicht, dass die Begriffe zu allgemein sind, um einen kunsthistorischen Inhalt zu verdeutlichen. Mit einem algorithmisch arbeitenden Verfahren, lassen sich Ergebnisse dieser Art allerdings nicht vermeiden.

4.3.2 Komposita – Wortklasse K

Als Besonderheit in der deutschen Sprache sind die vielfältigen Wortbildungsmöglichkeiten anzusehen. Komposita spielen im Kontext von Fachtexten eine besondere Rolle, da durch ihren Einsatz „[…] dem erhöhten Benennungsbedarf im Rahmen

[65] Zur Problematik einer klaren Abgrenzung zwischen Gemein- und Fachsprache vgl. Arntz (2004), bes. Kapitel 2.1 und 2.2

fachlicher Kommunikation leicht Genüge getan werden“[66] kann. Sie dienen der Ausdrucksökonomie und sollen Sachverhalte möglichst deutlich wiedergeben.[67] Als Merkmale von Fachtexten trifft dies auch auf die zu indexierenden Lexikonartikel des RDK zu (s. Kapitel 2.3). Da die Wortbildungsmöglichkeiten im deutschen Sprachgebrauch vielfältig sind, werden mit Lingo keine Wörterbücher lexikalisierter Komposita aufgebaut. Lingo besitzt durch den attendee *decomposer* die Möglichkeit Komposita algorithmisch durch einzelne Wortbestandteile zu identifizieren und somit auch zu extrahieren. Die extrahierten Terme sind durch ein K getaggt und können sowohl Fach- als auch Gemeinsprache kennzeichnen.[68] Lingo generiert Komposita nicht nur aus zusammengesetzten Substantiven, sondern auch aus Verben oder Verbindungen von Adjektiven und Substantiven, solange Einstellungen des decomposer-Parameters in der Wörterbuchkonfiguration de.lang dies nicht verhindern. Begriffe, wie „buntbemalen“, „kennzeichnen“ und „vielfältig“, werden als Komposita extrahiert, liegen allerdings in Verb- und Adjektivform vor. Diese prädestinieren eine MWG dazu unvollständig zu sein. Aus diesem Grund wurde bei der Erstellung der Wortmuster darauf geachtet, dass die Wortklasse K nur vereinzelt am Ende einer sequence platziert wird. Die MWGs „weltliches Möbelstück die buntbemalen“, „Opferstätte zu kennzeichnen“ und „Ornament in vielfältig“ verdeutlichen, das ein weiteres Wort zur Vervollständigung der MWGs fehlt, das durch die vorangestellten Begriffe spezifiziert wird.[69] Es wurde nicht komplett darauf verzichtet Komposita am Ende eines Wortmusters zu setzen, da sich diese auch aus Substantiven zu fachterminologischen Termen verbinden, wie an den MWGs „Museum für K*unstgewerbe*“ oder „Messe zur P*apstwahl*“ deutlich wird. Komposita wurden am Ende einer MWG beispielsweise in einer Adjektiv-Kompositum-Verbindung eingesetzt, sodass das Kompositum dadurch in substantivischer Form vorliegt, welches durch das vorangestellte Adjektiv näher spezifiziert wird, AERAK – „früheste Schöpfung der deutschen Bettelordensgotik“.

[66] Roelcke (2010), S. 79
[67] Vgl.: Roelcke (2010), S. 80
[68] Vgl.: Gödert (2012b), S. 280. Für eine detailliertere Funktionsbeschreibung des decomposer vgl.: bes. Kapitel 5.3.3
[69] Es werden keine Wortklassen in den Parametern des decomposer vereinbart, um eine Dekomposition zu optimieren. In der vorliegenden Arbeit sollen sich Komposita aus allen lexikalisierten Einträgen der WBs bilden, da sich Komposita auch aus Verbindungen von Verben und/oder Adjektiven usw. zu relevanten Termen verbinden, die zur Erzeugung fachterminologischer Inhalte dienen. Für diesen Zweck wurden die dafür vorgesehen WBs im attendee decomposer der Indexierungskonfigurationen ba-rdk-ling.cfg bzw. ba-rdk-lir.cfg definiert, s. Kapitel 4.2.3 und 4.2.4.2.

4.3.3 Adjektive – Wortklasse A

Adjektive sind innerhalb der Wortmuster durch die Wortklasse A gekennzeichnet. Diese sind in dem allgemeinen Rechtschreibwörterbuch dic-sys lexikalisiert und dienen sowohl einer fachsprachlichen-, als auch gemeinsprachlichen Identifizierung von Termen. Adjektive stehen als Beiwort vor einem Substantiv, da Substantive durch das Adjektiv näher definiert werden können.[70] Demzufolge wurden Adjektiv-Substantiv-Verbindungen innerhalb einer MWG als eine vielfach eingesetzte Kombination innerhalb von Wortmustern genutzt, wie z.B. ECAE - „Dreizack auf antiker Darstellung“.[71]

4.3.4 Funktionswörter – Wortklassen C, R, U

Die Bedeutung von Funktionswörtern innerhalb einer MWG wurde bereits im Kapitel 2.1.1 beschrieben. Aufbauend darauf wurden Artikel, Präpositionen und Konjunktionen innerhalb einer sequence gezielt platziert, damit sie ihren syntaktischen Zweck erfüllen.
Artikel wurden vorzugsweise vor Substantiven (Wortklasse E oder K) oder Adjektiv–Substantiv–Verbindungen (AE oder AK) platziert, wie es bei den sequences ERAE – „Symbol des eucharistischen Christus“, AERAE – „berühmte Komposition der italienischen Malerei“ oder ERE „Darstellung der Kreuzigung“ der Fall ist.
Konjunktionen wurden zwischen den Wortklassen E oder K gesetzt um substantivische Begriffe miteinander zu verbinden, z.B. erzeugt die sequence EUEE die MWG „Bistum und Hochstift Würzburg“.
Präpositionen, die als Verhältniswort die Fähigkeit besitzen, räumliche oder zeitliche Verhältnisse genauer anzugeben und somit inhaltliche Bezüge durch deren Einsatz näher definieren, bieten variable Einsatzmöglichkeiten. An den folgenden Beispielen wird dies deutlich. Bei dem Wortmuster ECE - „Christus am Kreuz“ oder ECEE – „Studie zum Hortus Deliciarum“ steht die Präposition vor einem Substantiv oder der Verbindung zweier Substantive. Präpositionen stehen innerhalb einer Wortgruppe immer vor einem Artikel, z.B. ECREE - „Szene aus der Kindheit Jesu“. In diesem Zusammenhang wurden auch Wortmuster gebildet, in denen Adjektiv-Substantiv-Verbindungen in Folge von Präpositionen und / oder Artikeln gezielt eingesetzt werden, wie es bei diesen MWGs deutlich wird: ECAE - „Ägypter *im* roten Meer“ oder ECRAE – „Apollo *in* der antik Kunst“.

[70] Vgl.: Hoberg (2004), S. 253
[71] Unter der allgemeinen Bezeichnung Substantiv, sind im Kontext die genutzten Wortklassen E/Fachbegriffe und K/Komposita zu verstehen.

4.3.5 Weitere Kriterien zur Wortmusterbildung

Anhand der einzelnen Wortklassen und deren Charakteristika, die sie innerhalb einer Mehrwortgruppe besitzen, wurden bereits einige Kriterien beschrieben, die bei der Musterbildung eine wichtige Rolle spielten. Diese wurden als Ausgangsbasis genutzt, um weitere Kriterien zu entwickeln.[72]

Es wurden maximal drei Fachbegriffe innerhalb eines Wortmusters hintereinander platziert, wie an dem genutzten Muster AECEEE deutlich wird. Dadurch können z. B. Personennamen, welche in der Form Vor- und Nachname auftauchen, innerhalb einer erweiterten Mehrwortgruppe, zu inhaltlich sinnvoll extrahierten Ergebnissen führen – „emblematisches Element im Werk Joris Hoefnagels“. Mehr als drei Fachbegriffe oder Komposita wurden nicht nebeneinander platziert, da diese passend gesetzten Funktionswörtern den Platz nehmen würden. Somit bliebe der Sinn dieser Wortart innerhalb einer MWG verborgen.

Ein Wortmuster endet nie mit einem Adjektiv. Diese treten in Verbindung mit einem Substantiv auf und wären andernfalls dafür prädestiniert unvollständige, nicht abgeschlossene MWGs zu erzeugen.

MWGs beginnen und enden nicht mit einem Funktionswort. Einerseits würden Wortmuster dieser Art unvollständige MWGs erzeugen, da ein Funktionswort immer in Verbindung mit einer zusätzlichen Wortklasse, wie dem Substantiv zu sehen ist. Andererseits wird in diesem Fall nur der Inhalt einer verkürzten MWG repräsentiert, da Funktionswörtern keine eigenständige, inhaltstragende Bedeutung besitzen. Demzufolge wurde diese Wortklasse nur innerhalb der eingesetzten Wortmuster positioniert, damit sie ihren syntaktischen Zweck erfüllen.

Es wurde vermieden, drei aufeinanderfolgende Funktionswörter einzusetzen. Der Einsatz von zu vielen Füllwörtern innerhalb eines Wortmusters würde inhaltstragenden Begriffen, welche in der Regel mit den Wortklassen A, E oder K gekennzeichnet sein können, den Platz nehmen. Deswegen wurde beispielsweise bei dem Aufbau vierteiliger Muster darauf geachtet, dass nur ein Funktionswort zum Einsatz kommt. Muster, welche sich aus fünf oder sechs Bestandteilen zusammensetzen, konnten zwei Funktionswörter enthalten, welche aus den bereits beschriebenen syntaktischen Gründen auch nebeneinander auftreten.

Eine Indexierung mit dem Einsatz von sequences, welche sich aus zwei Bestandteilen zusammensetzen entfällt. Das Einbinden eines Funktionswortes würde innerhalb

[72] Kriterien für die Musterbildung wurden z. T. nach einer ersten Testindexierung herausgearbeitet.

dieser Kombinationen nur einen inhaltstragenden Term extrahieren, was keine fachterminologische MWG im Sinne der vorliegenden Arbeit definiert.

4.4 Indexierungsdurchläufe mit dem attendee sequencer

Die Indexierungsdurchläufe erfolgen anhand der gebildeten Musterlängen. Ein Wortmuster kann aus mindestens drei, maximal aus sechs Bestandteilen bestehen. Somit wurden vier Indexierungen sortiert nach den Wortmusterlängen durchgeführt (s. Tabelle 1). Für die Indexierungen wurde immer dieselbe Wörterbuchkonfiguration (de.lang) genutzt und die sequences nach Bedarf ausgetauscht. Damit eine Analyse der Indexierungsergebnisse durchführbar ist, wurden die Ergebnisse in Dateien ausgegeben, in denen die MWGs nach Häufigkeit auftretend sortiert sind (ven.Datei, s. Tabelle 1). In einer weiteren Datei wurden die erzeugten MWGs nur mit dem jeweiligen Muster ausgegeben (seq.Datei, s. Tabelle 1). So konnte in der späteren Analyse untersucht werden, welche Muster fachterminologische MWGs erzeugen und bei welchen dies nicht der Fall war (s. Kapitel 5).
Im Ausgabebereich der Konfigurationsdatei ba-rdk-ling.cfg (s. Abbildung 7) wird festgelegt, welche Wortklassen bei einem Indexierungsdurchlauf zugelassen sind und die damit gekennzeichnete Terme aus den verwendeten WBs erkannt werden sollen. Weiterhin wurden Einstellungen vorgenommen, die es zuließen, zu jeder gewählten Musterlänge jeweils eine Indexierung durchzuführen und zwei Ausgabedateien mit extrahierten MWGs zu erstellen.

```
# Erstelle Datei mit Endung .vec für erkannte Indexterme
- vector_filter:    { in: syn, lexicals: '^[ksavebtwopnzxmcru]$' }
- text_writer:      { ext: vec, sep: "\n" }

# Erstelle Datei mit Endung .ven für erkannte Indexterme mit absoluter Häufigkeit
- vector_filter:    { in: syn, lexicals: '^[q]$', sort: term_abs }
- text_writer:      { ext: ven, sep: "\n" }

# Erstelle Datei mit Endung .ver für erkannte Indexterme mit relativer Häufigkeit
- vector_filter:    { in: syn, lexicals: '^[q]$', sort: term_rel }
- text_writer:      { ext: ver, sep: "\n" }

# Erstelle Datei mit Endung .mul für erkannte Mehrwortgruppen
- vector_filter:    { in: syn, lexicals: m }
- text_writer:      { ext: mul, sep: "\n" }

# Erstelle Datei mit Endung .seq für erkannte Wortsequenzen
- vector_filter:    { in: syn, lexicals: q }
- text_writer:      { ext: seq, sep: "\n" }
```

Abbildung 7: Ausgabebereich ba-rdk-ling.cfg

In den erstellten ven-Dateien wurden die extrahierten MWGs nach Häufigkeit sortiert ausgegeben. Durch die Wortklasse q, welche jeder erkannten Sequenz nach einer Indexierung zugeteilt wird, wurde festgelegt, das nur Sequenzen in der Ergebnisdatei hinterlegt sind. Durch sort: term_abs werden die MWGs nach absoluter Häufigkeit sortiert.

In den Ergebnisdateien mit der Endung seq wurden die MWGs nur mit den zugehörigem Muster angezeigt. Im Texteditor notepad++ werden diese automatisch in alphabetischer Reihenfolge sortiert.

Folgende Dateien, wurden nach den jeweiligen Indexierungen erzeugt und ausgegeben. Es wurden die angezeigten sequences für die jeweiligen Indexierungen genutzt (s. Tabelle 1).

Indexierung	**Erzeugte Datei**		**Genutzte Sequences**
	Dateiname	**Inhalt**	
Durchlauf mit dreiteiligen Mustern	basetall3.ven	Häufigkeitsangabe	EEE, KEE, EEK, EKK, ARE, ACE, AUE, ERE, KCE, ECK, EUE, EUK, KUE, KRE, ECE, EAE, KAE, AEE, KUK, EKE
	basetall3.seq	Nur Angabe sequence	
Durchlauf mit vierteiligen Mustern	basetall4.ven	Häufigkeitsangabe	ECEE, EUEE, EUEK, AECE, AKRK, AKUK, ECAE, ERAE, KRAE, ERAK, KCAE, AKUE, AKRE, KUEE, KCAK, EUKE, EAEE, ECKE, ERKE, EARE
	basetall4.seq	Nur Angabe sequence	
Durchlauf mit fünfteiligen Mustern	basetall5.ven	Häufigkeitsangabe	ACCEE, ACRAE, AERAK, AERAE EERAE, AECEE, ACEUE, ECRAE ECCAE, ECCEE, ECREE, EUERE KUERE, EUEAE, ECEUK, ECAEE KCREE, KCCAE, EEREE, EEUEE AECCEE,EECCAE,AKCRAE
	basetall5.seq	Nur Angabe sequence	
Durchlauf mit sechsteiligen Mustern	basetall6.ven	Häufigkeitsangabe	KRAEAE, EUECAK, EUKRAE, KACRAE, EACRAE, EUECKE, EUERKE, AECEUE, EUECRE, AKCKUE, ECAKRE, ACEUEE, AECEEE, AKCAEE, ECRAEE, EUKRAE, ERAEAE
	basetall6.seq	Nur Angabe sequence	

Tabelle 1: Ergebnisdateien nach Sequencerindexierung mit dafür verwendeten sequences

4.5 Indexierungsdurchlauf mit dem attendee multiworder

Aus allen Indexierungsdurchläufen wurde nach einer Stichprobenanalyse (s. Kapitel 5.5) MWGs in das neu erstellte Mehrwortgruppenwörterbuch ba-mul übernommen, die die Kriterien einer fachterminologischen MWG erfüllten. Die Indexierung erfolgte mit der Konfigurationsdatei ba-rdk-lir.cfg. Im attendee *multiworder* wurde das Wörterbuch

ba-mul zur Identifizierung von MWGs festgelegt (s. Abbildung 8). Für die Indexierungskonfiguration musste ein neues Datenformat definiert werden, weil die Standardeinstellungen von Lingo nicht mit dem vorliegenden RDK-Datensatzformat übereinstimmten. Dies ist notwendig, damit die Datensätze mit den zugehörigen Dokumentnummern erkannt werden und eine Zuordnung der Indexterme zu den jeweiligen Dokumenten möglich wird.

Durch die Einbindung des regulären Ausdrucks '^\[([\d-]+)\.\]' in die Indexierungskonfiguration ba-rdk-lir.cfg wurden die RDK-Datensätze in der Form [05-0933.][73] erkannt (s. Abbildung 8). Durch Verwendung der eckigen Klammern [05-0933.] erkennt Lingo bei einem Indexierungsdurchlauf, wann ein neues Dokument innerhalb einer gesamten Dokumentkollektion beginnt.[74]

```
attendees:

  ########################################
  # Text bereitstellen
  #

  # Angegebene Datei zeilenweise einlesen und verarbeiten
  - text_reader:     { files: $(files), records: '^\[([\d-]+)\.\]', progress: true }


  ########################################
  # Inhalte verarbeiten
  #

  # Zeile in einzelnen Sinnbestandteile (Token) zerlegen
  - tokenizer:       { }

  # Verbleibende Token im Wörterbuch suchen
  - word_searcher:   { source: ba-dic, mode: first }

  # Nicht erkannte Wörter auf Kompositum testen
  - decomposer:      { source: ba-dic }

  # Mehrwortgruppen im Strom erkennen
  - multi_worder:    { source: ba-mul, out: syn }

  # Wortsequenzen anhand von Regeln identifizieren
  # sequencer:       { stopper: PUNC,OTHR, out: syn }

  # Relationierungen einfügen
  # synonymer:       { skip: '?,t', source: rdk-syn, out: syn }
```

Abbildung 8: Ausschnitt Verarbeitungsbereich ba-rdk-lir.cfg mit vorgenommenen Einstellungen für Indexierung mit multiworder

[73] Innerhalb der eckigen Klammern befindet sich die Dokumentnummer welche gleichzeitig den Band- und Spaltenangabe des RDK entspricht. Band 5, Spalte 933

[74] Vgl.: Gödert (2012b), S. 295 f

Zur Indexierung mit dem Mehrwortgruppenwörterbuch wurde das Rechtschreibwörterbuch ba-dic erstellt und in dem attendee *wordsearcher* von ba-rdk-lir.cfg definiert, weiterhin festgelegt in der Wörterbuchkonfiguration de.lang (s. Abbildung 9). Das Wörterbuch ba-dic enthält alle Einträge aus dem fachspezifischen Wörterbuch brd-dic, die genutzten Funktionswörter aus fkt-dic und die Einträge aus dem allgemeinen Rechtschreibwörterbuch dic-sys, für eine zuverlässige Identifizierung aller Bestandteile einer MWG des Mehrwortgruppenwörterbuchs ba-mul.

```
#    Funktionswörterbuch
fkt-dic: { name: de/funkw-dic.txt, txt-format: WordClass, separator: '=' }

#    Mehrwortgruppenwörterbuch
ba-mul: { name: de/bamul-dic.txt, txt-format: SingleWord, use-lex: 'ba-dic', def-wc: m }

RDK-Wörterbücher
brd-dic: { name: de/bardk-dic.txt, txt-format: WordClass, separator: '=' }
ba-dic: { name: de/bardkall-dic.txt, txt-format: WordClass, separator: '=' }
dic-sys: { name: de/rdk_dic.txt, txt-format: WordClass }
```

Abbildung 9: Ausschnitt de.lang mit eingebundenen Wörterbüchern zur Indexierung mit dem multiworder

In der Ergebnisdatei des *multiworder*, basetall.mul, sind alle erzeugten MWGs den jeweiligen Dokumenten zugeordnet (s. Kapitel 5.5).

5 Analyse der Indexierungsergebnisse

Die erzeugten Ergebnisse des *sequencer* wurden nach jedem durchgeführten Indexierungslauf stichprobenartig analysiert. Zur Analyse dienten die erzeugten seq.Dateien und ven.Dateien. Für die folgende Hauptauswertung der Indexierungsergebnisse wurde eine Auswahl an Mustern getroffen.[75] Diese extrahierten fachterminologische MWGs mit kunsthistorischen Inhalten aus den Datensätzen des RDK. Anhand von Beispielen wird untersucht, welche sequences fachterminologische Inhalte erzeugt haben. Der Schwerpunkt bei der Analyse liegt auf MWGs, die aus vier- und fünf Bestandteilen bestehen. Bei der Aus- und Bewertung wurden die zuvor erstellten Kriterien fachterminologischer MWGs genutzt und eine Einteilung der Ergebnisse in diese vorgenommen. Die Einordnung erfolgt in folgende Kriterien:

- Zusammensetzung mehrerer Fachbegriffe bzw. die Verbindung eines gemeinsprachlichen Begriffs mit einem fachterminologischen Begriff
- Spezifizierung von Personennamen und Ortschaftsbezeichnungen durch Sachbegriffe

Diese unterliegen dem Merkmal der Abgeschlossenheit.

Die daraus geschlossenen Erkenntnisse werden vorgestellt. Im Anschluss wird anhand von Beispielen gezeigt, welche Probleme bzw. Fehlerquellen zu Ergebnissen führen, die nicht den Kriterien einer fachterminologischen MWG entsprechen. Weiterhin wird aufgeführt, welche Ergebnisse mit dem *multiworder* erzeugt werden können.

[75] Für die Analyse wurden Muster ausgewählt, welche sich aus vier- und fünf Bestandteilen zusammensetzen. Die Auswahl von dargestellten sequences zur Erzeugung fachterminologischer MWGs ist nicht vollständig. Es wurden weitere genutzt, welche die genannten Kriterien erfüllen.

5.1 Analyse Verbindungen kunsthistorische Sachbegriffe

Im Folgenden werden sequences untersucht, die zur Extraktion von MWGs geführt haben, welche vor allem Verbindungen von Sachbegriffen bzw. gemeinsprachlichen und Sachbegriffen herstellen.

Sequneces
ERAE / ECAE / KRAE / ERAK / KCAE / KCAK / AECE / AKRK / AKUE / AERAK / AERAE / ERKE

Abgeschlossene und fachterminologische MWGs werden extrahiert, wenn das Wortmuster mit einem substantivischen Begriff (Wortklasse E oder K) eingeleitet wird, gefolgt von einem Funktionswort (Artikel oder Präposition), welches den ersten Term mit einer Adjektiv–Substantiv–Verbindung in Beziehung bringt, z.B. durch das Wortmuster ERAE – „dogma der unbefleckt empfängnis“[76] oder ECAE – „adam auf altchristlich sarkophag“. Die Wortklasse E erzeugt als alleinstehender Begriff sowie in der Adjektiv Substantiv-Wortfolge AE Fachterminologie. Die hergestellte Verbindung durch die eingesetzten Funktionswörter erzeugen MWGs mit kunsthistorischem Inhalt.
Es werden keine unvollständigen MWGs extrahiert, wenn das Kompositum am Ende der getesteten Muster (ERAK, KCAK) eingesetzt wird. Durch die Kombination des Kompositums mit einem vorangestellten Adjektiv liegt das Kompositum in substantivischer Form vor. Beispielsweise erzeugt das Wortmuster ERAK – „büste der personifiziert kardinaltugend“ oder KCAK – „apollobrunnen im klein rathaushof“. Wie an den Beispielen ersichtlich wird das Kompositum durch das vorangestellte Adjektiv näher spezifiziert. Anhand der Auftrittshäufigkeiten dieser Muster zeigt sich, dass bereits viele Komposita als Fachbegriffe in den Wörterbüchern aufgenommen wurden und somit durch die Wortklasse E extrahiert werden. Beispielsweise wurden durch das Muster ERAE 1.647 MWGs extrahiert, durch das Wortmuster KRAE dagegen 681. Auch bei den Wortmustern ECAE mit 1.162 erzeugten MWGs, gegenüber des Musters KCAE=598, lässt sich dieser Sachverhalt feststellen.
Im Folgenden sind weitere Beispiele für positive sequences und daraus erzeugter MWGs aufgeführt, die die oben genannten Merkmale aufweisen. Diese erfüllen des Weiteren die Kriterien von fachterminologischen MWGs, die durch die Verbindung von Fachtermen bzw. gemeinsprachlichen Begriffen erzeugt werden.

[76] Beispiele werden in ihrer Grundform angegeben, wie sie vom sequencer identifiziert und extrahiert werden.

ERAE

„buchillustration der deutsch romantik“

„theorie der schön kunst“

„symbol der jungfräulich geburt“

ECAE

„abbildung von neun erzengel“

„abschnittsbefestigung durch äußere abschnittswall“

„abwehrzauber gegen böse dämon“

KRAE

„doppelfunktion der architektonisch wirkung“

„edelknabe des römisch könig“

„edelmetallkunst der katholisch kirche“

ERAK

„entwicklung des abendländisch stufenportal“

„abbildung des jüdisch tempelbau“

„abbildung einer sonderbar naturerscheinung“

KCAE

„abendmahlstisch im protestantisch kultus“

„bilderkreis auf alttestamentlich szene“

„elfenbeinrelief in silbervergoldet rahmung“

KCAK

„bildzusammenhang mit biblisch historienbild“

„grundrißgestaltung an rheinisch bettelordensbasiliken“

„kunstlandschaft im romanisch kleinkirchenbau“

ERKE

„darstellung der geburtsgeschichte christi“

„darstellung der himmelserscheinung maria“

„personifikation des wegweisend stern“

Auch an den getesteten Wortmustern AECE, AKRK und AKUE, welche mit einem Adjektiv beginnen, zeigen sich die positiven Ergebnisse. Durch ein Funktionswort

(Präposition, Artikel oder Konjunktion), welches eine Adjektiv - Substantiv - Verbindung (AE oder AK) und einen einzelnen Fachterm (E oder K) miteinander in Beziehung setzen, können fachterminologische MWGs extrahiert werden. Die sequence AECE erzeugt z.B. die MWG „merowingisch scheibenfibel mit grubenemail“, AKRK - „historisch wahrheitsgehalt der debutadeslegende“ oder AKUE – „toskanisch pilasterordnung und volutengiebel“. Durch die AE bzw. AK-Verbindung wird eine fachterminologische Wortfolge durch einen zusätzlichen Begriff, erzeugt durch die Wortklasse E bzw. K, präzisiert.

Für die Muster AERAK und AERAE trifft dieser Sachverhalt ebenfalls zu. Allerdings stehen hier zwei Adjektiv–Substantiv-Verbindungen (AE oder AK) miteinander in Beziehung, verbunden durch einen Artikel. MWGs, die aus diesen sequences resultieren sind z.B. AERAK – „bildlich darstellung des apostolisch glaubensbekenntnis“ oder AERAE – „exakt nachbildung des korinthisch kapitell“. Zwei fachsprachliche Verbindungen werden durch den Artikel miteinander in Beziehung gesetzt und entsprechen damit dem Kriterium einer kunsthistorischen MWG. Im Folgenden sind weitere Beispiele aufgeführt.

AECE

„zwölfjährig jesus im tempel“

„adorierend engel auf giebel“

„allegorisch auslegung auf sündenfall“

AKRK

„deutsch goldschmiedekunst der spätgotik“

„früh holzschnittdarstellung einer schulszene“

„spiritualistisch lichtmetaphysik des neuplatonismus“

AKUE

„weiß leinentuch als altarbekleidung“

„zweitürig reliquienschrank als vorbild“

„malerisch bildgestaltung und farbgebung“

AERAK

„künstlerisch möglichkeit des optisch farbaufbau“

„lehrhaft erklärung der antik götterbild“

„linke portal der nördlich querschiffsfassade“

AERAE

„elsässisch form der spätromanisch zeit“

„episch dichtung der höfisch zeit“

„eucharistisch bedeutung des gekreuzigt christus“

In qualitativer Hinsicht lässt sich kein Unterschied feststellen, ob und durch welche der eingesetzten Wortklassen Gemeinsprache, z.B. „früh“ „deutsch“ oder „entwicklung“, extrahiert wird. Dies wird an den positiv erzeugten MWGs (s.o.) sichtbar, solange eine MWG nicht nur aus gemeinsprachlichen Termen besteht.[77]
In allen Beispielen entfalten die genutzten Funktionswörter innerhalb einer MWG ihr Potenzial und verbinden die gewählten Kombinationen zu in sich abgeschlossenen fachterminologischen Mehrwortgruppen.

5.2 Analyse Spezifizierung von Personen-und Ortsnamen

Im Folgenden werden sequences analysiert, die sich vor allem für die Spezifizierung von Personen und Ortschaftsbezeichnungen bzw. Geografika auszeichnen.

Sequences EEREE / ECREE / KCREE / AECEE / ECEE

5.2.1 Spezifizierung von Namensformen

Personennamen werden durch die hintere EE-Wortklassenkombination in der Form Vor- und Nachname extrahiert. Durch zusätzliche Sachbegriffe, produziert durch die Wortklassen E, K und Verbindungen von AK, AE oder EE am Anfang einer sequence, wird die Namensform näher spezifiziert. Dies lässt sich z.B. an dem Muster AECEE – „ausgezeichnet abendmahlskelch von christoph knittel“ erkennen. Die hintere EE-Verbindung erzeugt den Namen Christoph Knittel, welcher durch die AE-Wortfolge, welche kunsthistorische Fachterminologie erzeugt, näher spezifiziert wird.

[77] Siehe negative Ergebnisse Kapitel 5.4.3, und Problematik einer Abgrenzung zwischen Gemeinsprache / Fachsprache Kapitel 4.3.1

EEREE

„freising hochaltar des jakob kaschauer“

„burgkmair totenbild des conrad celtis“

„historia scholastica des petrus comestor“

“hypnerotomachia poliphili des francesco Colonna”

ECEE

„altar von georg raphael“

„aquarell von william blake“

Es treten auch Varianten auf, in denen nur eine Namensform, entweder Vor- oder Nachname, durch einen zusätzlichen Sachbegriff definiert werden. Die hintere EE-Wortklassenkombination erzeugt eine Namensform mit zusätzlicher Spezifizierung durch einen Fachterm, z.B. EE – „kunst michelangelo“. Durch die Verbindung mit einem bzw. zwei eingesetzten Funktionswörtern (Wortklasse C oder R), wird die Namensspezifizierung mit einen zusätzlichen Begriff (Wortklasse E oder K) oder einer erweiterten AE-Verbindung in Beziehung gesetzt, z.B. ECREE – „berufung auf die kunst michelangelo“ oder AECEE – „schmiedeeisern chorgitter von johannes eberle“. Im Folgenden sind weitere sequences aufgeführt, die dem Beschriebenen entsprechen.

ECREE

„romgedanke in der kunst bernini“

„blut aus der seitenwunde christi“

„christus in der protestant kunst“

AECEE

„figürlich darstellung zu fuß christi“

„blind bettler beim einzug christi“

KCREE

„bildzyklen aus dem leben jesu“

5.2.2 Spezifizierung von Ortsnamen bzw. Geografika

Die gewählte EE–Kombination am Ende einer sequence erzeugt Spezifizierungen von Ortsnamen bzw. Geografika, welche durch vorangestellte AE oder EE-Verbindungen weiter präzisiert werden, sodass MWGs mit kunsthistorischen Inhalten erzeugt werden. Folgende Beispiele zeigen diese Merkmale.

KCREE

„christenverfolgung in den wüsten ägypten“
„wandgemälde in der capella greca“
„benediktusaltar in der abteikirche ettal“

AECEE

„dorisch portal am schloß aschaffenburg“

EEREE

„decretum gratiani der paris bibliothek“

ECEE

„altar im wien stephansdom“

ECREE

„altartuch aus dem zisterzienserinnenkloster zehdenick“

Erneut lässt sich feststellen, dass auch bei Spezifizierungen von Personen und Geografika die gezielt eingesetzten Funktionswörter ihren Zweck erfüllen. Die Terme der MWGs werden miteinander in Beziehung gesetzt, sodass inhaltliche Zusammenhänge abgeschlossen dargestellt werden.
Die analysierten Muster erzeugen natürlich nicht ausschließlich MWGs, die Personen oder Orte näher spezifizieren, wie an den nächsten Beispielen deutlich wird. Die AE oder EE-Wortfolgen sowie das alleinstehende E extrahieren auch kunsthistorische Sachbegriffe, die durch ein Funktionswort miteinander in Beziehung gesetzt werden.

AECEE

„deutsch ausgabe von ripa iconologia“

ECREE

„aktdarstellung in der christlich kunst“

ECEE

„arzt im hortus sanitatis“

5.3 Fazit positiver Ergebnisse

Die Extraktion von MWGs mit Einbeziehung von Funktionswörtern innerhalb kunsthistorischer Fachtexte erweist sich als sehr lohnenswert, wie die untersuchten sequences und die daraus resultierenden MWGs aufzeigen.
Funktionswörter stellen die nötige Verbindung zwischen den extrahierten Begriffen her, sodass kunsthistorischer Inhalt zusammenhängend dargestellt wird. Vier- und fünfteilige MWGs sind dafür besonders geeignet, da diese einer nötigen Länge entsprechen, in denen mehrere Fachbegriffe durch ein Funktionswort verbunden werden können und zudem in sich abgeschlossen sind.[78]
MWGs können auch aus zwei Bestandteilen bestehen, z.B. in der Kombination Adjektiv-Fachbegriff AE oder durch die Verbindung zweier Fachbegriffe EE. Die AE Kombination extrahiert z.B. „figürlich darstellung“ und EE „fuß christi“. Zwei MWGs die für sich alleinstehend noch keine hohe Aussagekraft besitzen. Figürliche Darstellungen gibt es im kunsthistorischen Bereich viele. „fuß christi“ ist nichtssagend, da ohne einen direkten Bezug nicht klar ist, wer oder was zu Fuße Christi ist. Anhand der extrahierten MWG „figürlich darstellung zu fuß christi“ durch das Wortmuster AECEE wird deutlich, dass ein Funktionswort, in diesem Fall eine Präposition, zwei einzelne MWGs verbindet und in Beziehung setzt. Diese, durch ein Funktionswort erweiterte Kombination, erzeugt ein ausdrucksstarkes Ergebnis, das die zwei einzelnen MWGs nicht vollbringen.

[78] Zum Unterschied siehe Negativbeispiele Kapitel 5.4.2, welche aus drei und sechs Teilen bestehen.

5.4 Negative Ergebnisse

In den folgenden Ausschnitten (s. Abbildung 10) ist zu erkennen, dass sowohl vier-, als auch fünfteilige Wortmuster, welche fachterminologische Inhalte extrahieren, immer auch schlechte Ergebnisse produzieren.[79] Was unter negativen Ergebnissen zu verstehen ist, und warum diese erzeugt werden wird in den folgenden Abschnitten erläutert.

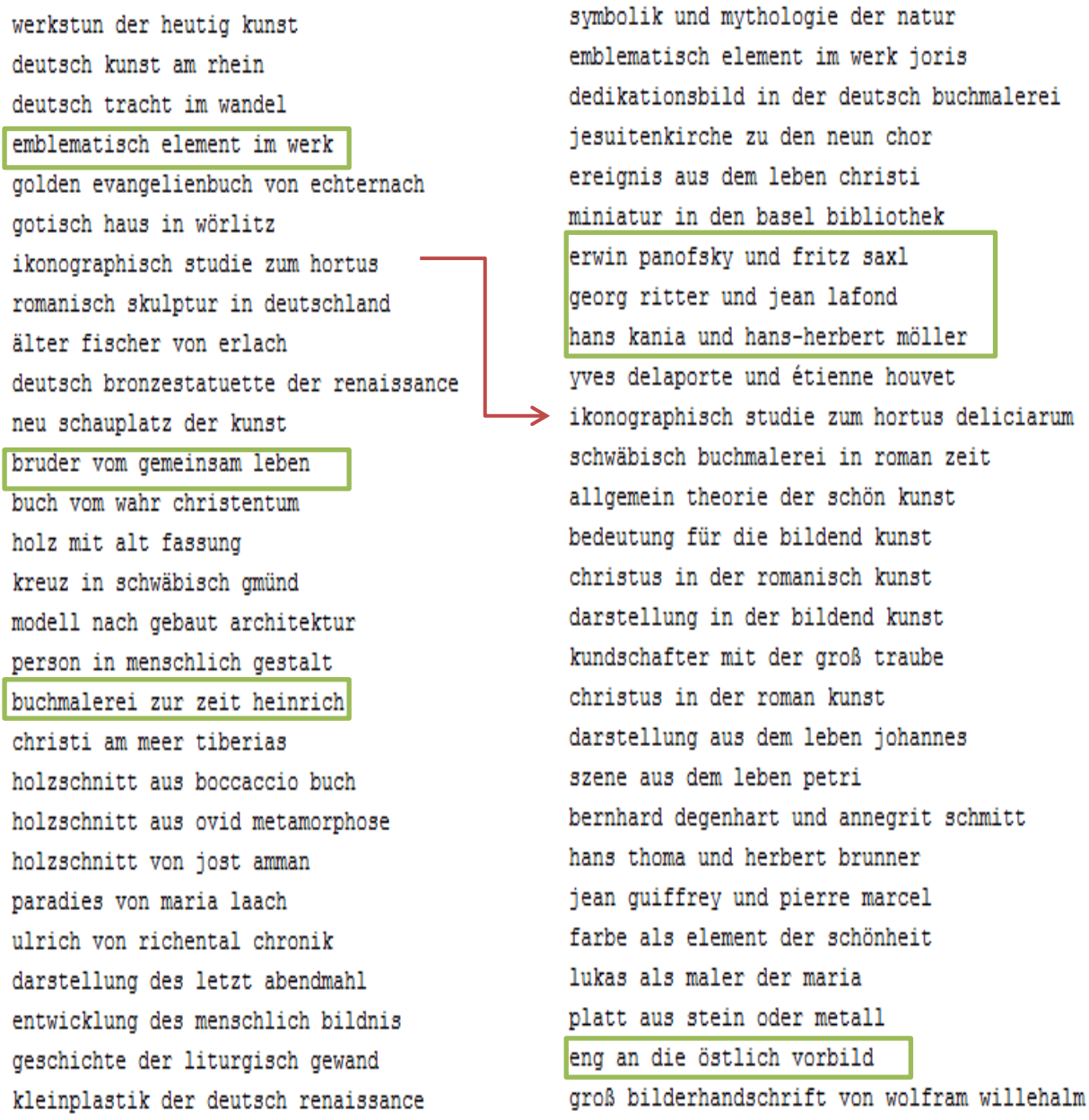


Abbildung 10: zeigt Ausschnitte von vier- und fünfteiligen Mehrwortgruppen

[79] Markierte Beispiele

5.4.1 Unvollständigkeit bei vier- und fünfteiligen Mustern

Unvollständige MWGs sind ein Problem, welches nicht vermieden werden kann, da die Wortmuster teils zu kurz sind, um einen abgeschlossenen Inhalt zu erzeugen. Verdeutlicht wird dies an den Beispielen „buchmalerei zur zeit heinrich“ – ECEE oder „ikonographisch studie zum hortus“ – AECE. Ein abgeschlossener, fachterminologischer Inhalt bildet sich erst durch einen weiteren Bestandteil ab, der beispielsweise durch das fünfteilige Wortmuster AECEE erzeugt wurde - „ikonographisch studie zum hortus deliciarum“ (s. Abbildung 10). Eine andere Form von Unvollständigkeit ist die Erzeugung von Satzfragmenten, da Wortmuster auch MWGs erzeugen, die zu lang sind, also aus zu vielen Bestandteilen bestehen. Ein abgeschlossener Inhalt würde sich erst durch einen kompletten Satz abbilden lassen. Z.B. zeigt die MWG „konrad peutinger die römisch inschrift“ deutlich, dass Terme zur Vervollständigung einer MWG fehlen. Es wird ersichtlich, dass sich unter Umständen ein kompletter Satz abbilden ließe, was nicht im Sinne dieser Arbeit ist, um einen Inhalt vollständig zu präsentieren - „1505 veröffentlichte Konrad Peutinger die römischen Inschriften…“[80].

5.4.2 Drei- und sechsteilige Muster

Die Problematik unvollständiger MWGs oder Erzeugung von Satzfragmenten zeigte sich vor allem bei den getesteten Mustern, welche sich aus drei und sechs Bestandteilen zusammensetzen. Die erzeugten MWGs im Falle von dreiteiligen Mustern sind zu kurz, wohingegen die Extraktion durch sechsteilige Wortmuster zu lang ist, um fachterminologische Inhalte zu repräsentieren.

Beispiele dreiteiliger MWGs:

„deutsch in bildnis“
„gleichzeitigkeit von miniatur“
„schweiz architekt georg“
„gnadenstuhl zwischen stifter“
„gottesdienst durch gewand“

[80] Artikel „Epitaphienbuch“, Band 5, Spalte 933

Beispiele sechsteiliger MWGs:

„abgesehen von kloster und wallfahrtskirche können“

„viele haus in schaffhausen oder stein“

„sachsen verwendet in der früh gotik“

„wohl dem rudolfinisch kunstkreis angehörend tafelbild“

„hieroglyphe und emblem in den drucker“

5.4.3 Allgemeinsprache

MWGs, die sich aus allgemeinen Begriffen zusammensetzen (s. Abbildung 10), bilden ebenfalls schlechte Ergebnisse ab[81], wie „bruder vom gemeinsam leben“ – ECAE oder „eng an die östlich vorbild“ – ACRAE zeigen. Diese Beispiele erfüllen nicht die Kriterien einer fachterminologischen MWG. Es sollte demnach mindestens ein Term extrahiert werden und Bestandteil einer MWG sein, der Fachterminologie repräsentiert, um kunsthistorische Inhalte abzubilden.

5.4.4 Personennamen

Sequences, die lediglich Personennamen erzeugen und miteinander verbinden, können ebenfalls nicht als fachterminologische MWGs angesehen werden. Das Muster EEUEE unter gezieltem Einsatz einer Konjunktion erzeugte Kombinationen dieser Art, wie z.B. „erwin panofsky und fritz saxl“, „georg ritter und jean lafond“ und „hans kania und hans-herbert möller“. Deshalb kann diese sequence bei der Extrahierung von kunsthistorischen MWGs vernachlässigt werden. Allerdings lässt sich anhand dieser Ergebnisse erkennen, dass der Einsatz eines Funktionswortes die Qualität einer MWG bestimmen kann. Das Wortmuster EEREE erzeugte durch Nutzung des Artikels, wie bereits an dem Bespiel ersichtlich – „historia scholastica des petrus comestor“, im Gegensatz zu EEUEE, fachterminologische Ergebnisse.

[81] Beschriebene Problematik der Trennung Gemein- und Fachsprache innerhalb der Klasse E, sowie Wortklasse K und A, welche generell sowohl Fach- als auch Gemeinsprache kennzeichnen.

5.4.5 Optimierung negativer Ergebnisse

Die Extraktion von MWGs, die aus den genannten Gründen, nicht den Kriterien einer fachterminologischen MWG entsprechen, lassen sich durch ein algorithmisches Verfahren, welches auf Wortmustern basiert, nicht vermeiden. Optimierungspotenzial besteht immer durch die Abänderung von Einträgen in den Wörterbüchern und die damit verbundene Wortklassenkennzeichnung. Der Wörterbucheintrag „haus“, welcher mit der Wortklasse E gekennzeichnet wurde, könnte beispielsweise durch die Wortklasse S[82] gekennzeichnet werden. Somit wäre eine Abgrenzung zwischen Fach- und Gemeinsprache möglich. Extrahierte MWGs, wie „rot haus in monschau“ könnten dadurch vermieden werden. Personennamen könnten eine andere Kennzeichnung durch eine andere Wortklasse erhalten, damit deren Extrahierung gezielter durch die Einbindung in ein Wortmuster gesteuert wird. Weiteres Optimierungspotenzial besteht in der Abänderung der Suffixliste.

Die erzeugten MWGs, welche durch eine Indexierung mit dem *sequencer* extrahiert werden, kann als Empfehlung angesehen werden, um diese nach einer intellektuellen Sichtung in ein Mehrwortgruppenwörterbuch zu übernehmen. So besteht die Möglichkeit eine Indexierung nur mit kunsthistorischen MWGs durchzuführen und diese als potentielle Indexterme den jeweiligen Datensätzen zuzuordnen.

5.5 Indexierungsergebnisse mit dem multiworder

Im Folgenden werden Ergebnisse präsentiert, die durch eine Indexierung mit dem *mutiworder* erzeugt wurden.[83]

Nach der Umarbeitung von MWGs in eine grammatikalisch korrekte Form, die vorab durch den *sequencer* extrahiert wurden, sind MWG-Einträge wie im Folgenden Ausschnitt (s. Abbildung 11) ersichtlich, in das neu erstellte Mehrwortgruppenwörterbuch ba-mul aufgenommen wurden. Die darin enthaltenen MWGs weisen die Merkmale fachterminologischer MWGs auf.[84]

[82] Wortklasse S könnte gemeinsprachliche Substantive kennzeichnen.
[83] Vorgehen bei der Indexierung mit dem multiworder s. Kapitel 4.5
[84] Die aus allen durchgeführten Indexierungsdurchläufen extrahierten MWGs wurden stichprobenartig gesichtet und beispielhaft in das WB ba-mul aufgenommen.

wort und bild in der graphik
zusammengekommenes wissen über engel und dämon
ärztlicher schutzpatron in der bildenden kunst
äußerer steinring in einer geschlossenen mauer
östlicher bildtypus in der abendländischen kunst
darstellung der geburt christi
darstellung der kreuzigung christi
system der zeichnenden kunst
päpstlicher hof in avignon
barocke freskomalerei in deutschland
galluspforte des basel münster
karolingische schule von tours
meister des registrum gregorii
neues palais in potsdam
neun chor der engel
opfer des alten bundes
werkstun der heutigen kunst
ägypter im roten meer
buch vom wahren christentum
buchmalereifarbe auf pflanzlicher basis
christi am meer tiberias
darstellung des letzten abendmahls
entwicklung des menschlichen bildnis
goldenes evangelienbuch von echternach
kaisersaal der würzburg residenz
person in menschlicher gestalt
theorie der schönen kunst
werk der bildendenden kunst
auferstehung und himmelfahrt christi
bildende kunst der antike
bilderkreis des griechischen physiologus
bildprogramm zum leben jesu
chorgestühl des ulmer münster
darstellung der auferstehung christi
einband des codex aureus
fest der beschneidung christi
frau am grab christi
freude und schmerzen marias
gemalte anbetung der könige
germanischer gott und held
goldener adler in innsbruck
idee des königtum christi
karikatur der europäischen völker
karolingische buchmalerei von tours
klassisches altertum in rom
kreuzkirche in schwäbisch gmünd
kunst der alt christen

Abbildung 11: Ausschnitt aus ba-mul

Abbildung 12 zeigt die generierten Indexterme, die nach einer Indexierung mit dem Mehrwortgruppenwörterbuch ba-mul gebildet wurden und wie sie den einzelnen Datensätzen zugeordnet sind

```
01-1445*beweis für die jungfräuliche geburt|symbol der jungfräulichen geburt
01-1449*bilderschrift der renaissance|hieroglyphenkunde des humanismus
01-1457-1*darstellung der kreuzigung christi
01-1503*kreuzkirche in schwäbisch gmünd|renaissance in deutschland
01-1519*speculum humanae salvationis
02-0063-1*darstellung des sündenfall|sündenfall und erlösung
02-0065*darstellung des sündenfall
02-0069*christus am kreuz
02-0071*christus am kreuz|speculum humanae salvationis|symbol der jungfräulichen geburt
02-0073*baum des lebens
02-0105*baum des lebens
02-0131*renaissance in deutschland
02-0179*darstellung des sündenfall
02-0225-1*speculum humanae salvationis
02-0265*klosteranlage des frühen mittelalter
02-0327*renaissance in deutschland
02-0371*germanischer gott und held
02-0467*speculum humanae salvationis
02-0475*ikonographische studie zum hortus deliciarum
02-0501*speculum humanae salvationis
02-0505*speculum humanae salvationis
02-0507*speculum humanae salvationis
02-0515*speculum humanae salvationis
02-0517*kunst der alt christen
02-0517-1*bilderschrift der renaissance|hieroglyphenkunde des humanismus
```

Abbildung 12: Ergebnisdatei basetall.mul nach der Indexierung mit Mehrwortgruppenwörterbuch

Nach der Indexierung wurden die Datensätze mit MWGs angereichert, teils auch mit mehreren. Somit wäre das Auffinden der Dokumente durch eine Suche mit MWGs möglich, wenn diese in einem weiteren Arbeitsschritt in die Datenbank seite.dbm integriert werden.[85]

[85] Die Integration neuer MWGs zur Datenbank seite.dbm wurde nicht durchgeführt.

Der folgende Beispielausschnitt zeigt das Auftreten der extrahierten Mehrwortgruppen in einem RDK Lexikonartikel.[86]

„[…] Im ausgehenden Mittelalter benutzte der Dominikaner Franz von Retz († 1421) im "Defensorium inviolatae virginitatis beatae Mariae" den Naturgesetzen widersprechende Eigenschaften von Bäumen als Symbole der jungfräulichen Geburt [...]
*Die Legenda aurea berichtet von dem wundertätigen B. Persidis, der sich, als die Heilige Familie auf der Flucht nach Ägypten kam, anbetend zur Erde neigte; Darstellungen z. B. von Schongauer (B. 7; H. Schenck, Martin Schongauers Drachenbaum, Naturwiss. Wochenschrift N.F. 19, 1920) und Hans Baldung im Freiburger Hochaltar; s. *Flucht nach Ägypten. Als Typus für die Aufrichtung des Kreuzes und ebenso für Christus am Kreuz gilt im Speculum humanae salvationis der B., den Nebukadnezar im Traum sah (Dan. 4, 7ff.; Lutz und Perdrizet Taf. 132 XXIV, 2 u. Taf. 47). […]“*

[86] Ausschnitt RDK, Artikel „Baum“, Band 2, Spalte 71

6 Fazit

Die positiven Ergebnisse der Analyse zeigen, dass das vorgenommene algorithmische Verfahren zur Mehrwortgruppenerkennung und Extraktion zweckmäßig ist. Durch die eingesetzte Indexierungssoftware Lingo und dessen Programmmodul *sequencer* konnten fachterminologische MWGs erzeugt werden. Kunsthistorische Inhalte werden durch die erzeugten Ergebnisse vermittelt, welche die fachterminologischen Kriterien einer Mehrwortgruppe erfüllen (s. Kapitel 4.1 und 5). Wie sich gezeigt hat, können die in der Vergangenheit erstellten fachspezifischen Wörterbücher zur Kunstgeschichte für weitere Indexierungsvorhaben in diesem Fachbereich genutzt werden und flexibel auf den eigenen Bedarf hin abgeändert werden. Relevant für die Erzeugung fachterminologischer MWGs sind die gezielt erstellten Wortmuster und die zusätzliche Einbindung von Funktionswörtern. Es konnte bewiesen werden, dass sich das Einbinden von Funktionswörtern innerhalb einer Mehrwortgruppe lohnt. Relevante Begriffe, die durch die Funktionswörter miteinander in Beziehung gesetzt werden, repräsentieren zusammenhängende, abgeschlossene Inhalte. In der vorliegenden Arbeit wurde nur ein Testset potentieller Wortmuster erstellt und analysiert. Das Bilden weiterer Muster anhand der gebildeten Kriterien hat das Potential, zusätzliche fachterminologische MWGs zu erzeugen. In weiteren Indexierungen können sequences festgelegt werden, die fachterminologische MWGs in Fachtexten identifizieren, aber negative Ergebnisse von vornherein reduzieren. Das angewendete algorithmische Verfahren zeigt, dass es fächerübergreifend Verwendung finden kann. Sowohl im mathematischen[87] als auch kunsthistorischen Bereich dieser Arbeit. An Datenquellen anderer Fachbereiche kann untersucht werden, ob sich die positiven Ergebnisse auf Grund der Verwendung von Funktionswörtern wiederholen.

Unabhängig vom Fachbereich können beim automatischen Indexieren mit der Software Lingo identifizierte MWGs für eine hohe Anzahl an elektronisch vorliegenden Daten als Ausgangsbasis einer intellektuellen Erschließung dienen. Wenn rein automatisch arbeitende Verfahren nicht ausschließlich für eine inhaltliche Erschließung genutzt werden, können die erzeugten MWGs des Programmmoduls *sequencer* als Empfehlung einer intellektuellen Erschließung von Nutzen sein. Die MWGs können allerdings auch direkt zum Aufbau eines Mehrwortgruppenwörterbuchs genutzt werden. So werden sie durch eine zusätzliche Indexierung mit dem Programmmodul *multiworder* von Lingo und der entsprechenden Indexierungskonfiguration als Indexterme den jeweiligen Dokumenten zugeordnet. Die erzeugten Mehrwortgruppen der

[87] Vgl.: Gödert (2012a)

vorliegenden Arbeit könnten beispielsweise in einem weiteren Arbeitsschritt als potentielle Indexterme im Information Retrieval weitere Verwendung finden, wenn sie in die entsprechende Datenbank integriert werden.

I Literaturverzeichnis

Arens, F. V. (1964): Epitaphienbuch. In: Reallexikon zur Deutschen Kunstgeschichte. Bg. von Otto Schmitt, hrsg. von L. H. Heydenreich … , Bd. 5, Stuttgart 1964. Spalte 932-936

Arntz, R., Picht H., Mayer, F. (2004): Einführung in die Terminologiearbeit. 5. Aufl., Hildesheim [u.a.]: Olms. (Sudien zu Sprache und Technik ; 2)
ISBN: 978-3-487-11553-5

Augustyn, W. (2004): Das „Reallexikon zur Deutschen Kunstgeschichte“. IN: RDK-multimedial: Erstellung einer Multimedia-CD-ROM des „Reallexikons zur Deutschen Kunstgeschichte“. Hrsg. von Winfried Gödert …Köln 2004. (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft ; 45)
URL: http://www.fbi.fh-koeln.de/institut/papers/kabi/volltexte/band045.pdf [letzter Aufruf: 26.08.2013]

Bachmann, K.-W., Jászai, G., Kobler, F. (2003): Flügelretabel. In: Reallexikon zur Deutschen Kunstgeschichte. Hrsg. von Zentralinstitut für Kunstgeschichte, Bd. 9, München 2003, Spalte 1450-1478

Dengel, A. (2012): Semantik in der Informationsextraktion. IN: Semantische Technologien: Grundlagen – Konzepte – Anwendungen. Heidelberg: Spektrum. S. 205-229
ISBN: 978-3-8274-2664-2

Exner, Matthias (1993): Flammenschwert. In: Reallexikon zur Deutschen Kunstgeschichte. Hrsg. von Zentralinstitut für Kunstgeschichte, Bd. 9, München 1993, Spalte 693-740

Fandrych, C., Thurmair-Mumelter, M. L. (2011): Textsorten im Deutschen: Linguistische Analysen aus sprachdidaktischer Sicht. Tübingen: Stauffenburg. (Stauffenburg Linguistik ; 1)
ISBN: 3-860571958

Gödert, W. (2012a): Detecting multiword phrases in mathematical text corpora.
URL: http://arxiv.org/ftp/arxiv/papers/1210/1210.0852.pdf [letzter Aufruf: 26.08.2013]

Gödert W., Lepsky, K., Nagelschmidt, M. (2012b): Informationserschließung und Automatisches Indexieren: ein Lehr- und Arbeitsbuch. Berlin: Springer.
ISBN: 978-3-642-23512-2

Hoberg, R., Hoberg, U. (2004): Deutsche Grammatik. 3. überarb. Aufl., Mannheim [u.a.]: Dudenverl. (Der kleine Duden ; 4)
ISBN: 978-3-411-05573-1

Huo, W. (2012): Automatic Multi-word Term Extraction and its Application to Web-page Summarization. Guelph Master-Thesis.
URL:
https://atrium.lib.uoguelph.ca/xmlui/bitstream/handle/10214/4959/Thesis_21.pdf?sequence=1 [letzter Aufruf 26.08.2013]

Lepsky, K. (2006): Automatische Indexierung des Reallexikons zur Deutschen Kunstgeschichte. IN: Information und Sprache: Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern. Festschrift für Harald H. Zimmermann. Hrsg. von Ilse Harms … München (2006), S. 169-178.

Neuß, Wilhelm (1953): Christus. In: Reallexikon zur Deutschen Kunstgeschichte. Bg. von Otto Schmitt, hrsg. von Ernst Gall … , Bd. 3, Stuttgart 1953. Spalte 609-630

Pittner, K., Berman, J. (2008): Deutsche Syntax: ein Arbeitsbuch. 3. aktualisierte Aufl., Tübingen: Narr.
ISBN: 978-3-8233-6450-4

Roelcke, T. (2010): Fachsprachen. 3. Aufl., Berlin: Erich Schmidt. (Grundlagen der Germanistik, 37)
ISBN: 978-3-503-12221-9

Srikumar, V., Roth, D. (2013): Modeling Semantic Relations Expressed by Prepositions. IN: Transactions of the Association for Computational Linguistics. Vol. 1 (2013) S. 231-242
URL: http://www.transacl.org/wp-content/uploads/2013/05/paper231.pdf [letzter Aufruf: 26.08.2013]

Stauch, L., Föhl, W. (1938): Baum. In: Reallexikon zur Deutschen Kunstgeschichte. Hg. von Otto Schmitt, Bd. 2, Stuttgart 1938. Spalte 63-82

Über das RDK – RDK Web: Homepage.
URL: http://rdk.zikg.net/gsdl/cgi-bin/library.exe?e=p-01000-00---off-0rdkZz-web%2e1--00-1--0-10-0---0---0prompt-10---4-------0-1l--11-de-Zz-1---20-about---01-3-1-00-0-0-11-1-0utfZz-8-00&a=p&p=impressum [letzter Aufruf 26.08.2013]

Zentralinstitut für Kunstgeschichte: Forschungsstelle Realienkunde / Reallexikon zur Deutschen Kunstgeschichte. Homepage.
URL: http://www.zikg.eu/main/rdk/rdk.htm [letzter Aufruf 26.08.2013]

II Anhang

1 Indexierungskonfiguraton ba-rdk-ling.cfg

```
# Lingo-Konfiguration
#
---
meeting:

  attendees:

    ########################################
    # Text bereitstellen
    #

    # Angegebene Datei zeilenweise einlesen und verarbeiten
    - text_reader:     { files: $(files), progress: true }


    ########################################
    # Inhalte verarbeiten
    #

    # Zeile in einzelnen Sinnbestandteile (Token) zerlegen
    - tokenizer:       { }

    # Abkürzungen erkennen und auflösen
#   - abbreviator:     { source: sys-abk }

    # Verbleibende Token im Wörterbuch suchen
    - word_searcher:   { source: brd-dic fkt-dic dic-sys, mode: first }

    # Schreibweisen variieren und erneut suchen
#   - variator:        { source: sys-dic }

    # Bindestrichergänzungen rekonstruieren
#   - dehyphenizer:    { source: sys-dic }

    # Wortstämme für nicht erkannte Wörter einfügen
#   - stemmer:         { }

    # Nicht erkannte Wörter auf Kompositum testen
    - decomposer:      { source: brd-dic dic-sys }

    # Mehrwortgruppen im Strom erkennen
    # multi_worder:    { source: ba-mul, out: syn }

    # Wortsequenzen anhand von Regeln identifizieren
    - sequencer:       { stopper: PUNC,OTHR, out: syn }

    # Relationierungen einfügen
    # synonymer:       { skip: '?,t', source: sys-syn, out: syn }


    ########################################
    # Datenstrom anzeigen
    #
#   - debugger:        { eval: 'true', ceval: 'cmd!="EOL"', prompt: 'lex:) ' }
```

```
########################################
# Ergebnisse ausgeben
#

# Erstelle Datei mit Endung .log für Datenstrom
- vector_filter:   { in: syn, debug: 'true', prompt: 'lex:) ' }
- text_writer:     { ext: log, sep: "\n" }

# Erstelle Datei mit Endung .non für nicht erkannte Wörter
- noneword_filter: { in: syn }
- text_writer:     { ext: non, sep: "\n" }

# Erstelle Datei mit Endung .ste für Wortstämme
- vector_filter:   { in: syn, lexicals: z }
- text_writer:     { ext: ste, sep: "\n" }

# Erstelle Datei mit Endung .vec für erkannte Indexterme
- vector_filter:   { in: syn, lexicals: '^[ksavebtwopnzxmcru]$' }
- text_writer:     { ext: vec, sep: "\n" }

# Erstelle Datei mit Endung .ven für erkannte Indexterme mit absoluter Häufigkeit
- vector_filter:   { in: syn, lexicals: '^[q]$', sort: term_abs }
- text_writer:     { ext: ven, sep: "\n" }

# Erstelle Datei mit Endung .ver für erkannte Indexterme mit relativer Häufigkeit
- vector_filter:   { in: syn, lexicals: '^[q]$', sort: term_rel }
- text_writer:     { ext: ver, sep: "\n" }

# Erstelle Datei mit Endung .mul für erkannte Mehrwortgruppen
- vector_filter:   { in: syn, lexicals: m }
- text_writer:     { ext: mul, sep: "\n" }

# Erstelle Datei mit Endung .seq für erkannte Wortsequenzen
- vector_filter:   { in: syn, lexicals: q }
- text_writer:     { ext: seq, sep: "\n" }

# Erstelle Datei mit Endung .syn für erkannte Synonyme
- vector_filter:   { in: syn, lexicals: y, sort: term_abs }
- text_writer:     { ext: syn, sep: "\n" }
```

2 Indexierungskonfiguration ba-rdk-lir.cfg

```
# Lingo-Konfiguration für den Test mit einer LIR-Datei
# Gebräuchliche Patterns sind
#   "^\021(\d+\-\d+)\022"
#   "^\[(\d+)\.\]"
meeting:
attendees:
   ##########################################
   # Text bereitstellen
   #
   # Angegebene Datei zeilenweise einlesen und verarbeiten
   - text_reader:     { files: $(files), records: '^\[([\d-]+)\.\]', progress: true }

   ##########################################
   # Inhalte verarbeiten
   #

   # Zeile in einzelnen Sinnbestandteile (Token) zerlegen
   - tokenizer:       { }

   # Verbleibende Token im Wörterbuch suchen
   - word_searcher:   { source: ba-dic, mode: first }

   # Nicht erkannte Wörter auf Kompositum testen
   - decomposer:      { source: ba-dic }

   # Mehrwortgruppen im Strom erkennen
   - multi_worder:    { source: ba-mul, out: syn }

   # Wortsequenzen anhand von Regeln identifizieren
   # sequencer:       { stopper: PUNC,OTHR, out: syn }

   # Relationierungen einfügen
   # synonymer:       { skip: '?,t', source: rdk-syn, out: syn }
   ##########################################
   # Datenstrom anzeigen
   #
#   - debugger:        { eval: 'true', ceval: 'cmd!="EOL"', prompt: 'lex:) ' }
   ##########################################
   # Ergebnisse ausgeben
   #
   # Erstelle Datei mit Endung .log für Datenstrom
   - vector_filter:   { in: syn, debug: 'true', prompt: 'lex:) ' }
   - text_writer:     { ext: log, sep: "\n" }

   # Erstelle Datei mit Endung .non für nicht erkannte Wörter
   - noneword_filter: { in: syn }
   - text_writer:     { ext: non, sep: '|' }

   # Erstelle Datei mit Endung .vec für erkannte Indexterme
   - vector_filter:   { in: syn, lexicals: '^[ksavetmcru]$' }
   - text_writer:     { ext: vec, sep: '|' }

   # Erstelle Datei mit Endung .mul für erkannte Mehrwortgruppen
   - vector_filter:   { in: syn, lexicals: m }
   - text_writer:     { ext: mul, sep: '|' }

   # Erstelle Datei mit Endung .seq für erkannte Wortsequenzen
   - vector_filter:   { in: syn, lexicals: q, sort: term_abs }
   - text_writer:     { ext: seq, sep: '|' }

   # Erstelle Datei mit Endung .syn für erkannte Synonyme
   - vector_filter:   { in: syn, lexicals: y, sort: term_abs }
   - text_writer:     { ext: syn, sep: '|' }
```

3 Wörterbuchkonfiguration de.lang mit eigenen Einstellungen (Ausschnitt)

```
#    Funktionswörterbuch
     fkt-dic: { name: de/funkw-dic.txt, txt-format: WordClass, separator: '=' }

     #    Mehrwortgruppenwörterbuch
     ba-mul: { name: de/bamul-dic.txt, txt-format: SingleWord, use-lex: 'ba-dic', def-wc: m }

#    RDK-Wörterbücher
     brd-dic: { name: de/bardk-dic.txt, txt-format: WordClass, separator: '=' }
     ba-dic: { name: de/bardkall-dic.txt, txt-format: WordClass, separator: '=' }
     dic-sys: { name: de/rdk_dic.txt, txt-format: WordClass }

     rdk-usr: { name: de/rdk-usr.txt, txt-format: WordClass }
     rdk-us3: { name: de/rdk-us3.txt, txt-format: WordClass }
     per-big: { name: de/rdk_per_big.txt, txt-format: WordClass }
     rdk-ort: { name: de/rdk_ort.txt, txt-format: WordClass }

suffix:
     #    Suffixliste, Stand: 30-06-2005
     #    Suffixklasse:   s = Substantiv, a = Adjektiv, v = Verb, e = Fachbegriff, f = Fugung
     #    Suffixe je Klasse:   "<suffix>['/'<ersetzung>][ <suffix>['/'<ersetzung>]]"
     #    neue Klassen: c=präposition, r=artikel, u=konjunktion
     #    weitere Suffix-/Wortklassen: w=wortform t=stopwort
     - [s, "e en er ern es n s se sen ses"]
     - [a, "este ste ster sten stes ester estes esten e em en er ere eren erer eres es erem"]
     - [v, "e/en en/en est/en et/en st/en t/en te/en ten/en eten/en ete/en etest/en s"]
     - [e, "s - e en er ern nen es n s se sen ses ' 's s- en- n- isch ischen ische ischer isches ischem schem
sches sche scher schen ner"]
     - [f, "s n e en es er ch/che /en"]

sequencer:
     sequences: [ [EEE, "EEE 1 2 3"], [KEE, "KEE 1 2 3"], [EEK, "EEK 1 2 3"], [EKK, "EKK 1 2 3"], [ARE,
"ARE 1 2 3"]
```

4 Indexierungskonfiguration lingo.cfg (Standardkonfiguration)

```
meeting:

  attendees:
    ########################################
    # Text bereitstellen
    #
    # Angegebene Datei zeilenweise einlesen und verarbeiten
    - text_reader:     { files: $(files), progress: true }
    ########################################
    # Inhalte verarbeiten
    #
    # Zeile in einzelnen Sinnbestandteile (Token) zerlegen
    - tokenizer:       { }
    # Abkürzungen erkennen und auflösen
#   - abbreviator:     { source: sys-abk }
    # Verbleibende Token im Wörterbuch suchen
    - word_searcher:   { source: sys-dic, mode: first }
    # Schreibweisen variieren und erneut suchen
#   - variator:        { source: sys-dic }
    # Bindestrichergänzungen rekonstruieren
#   - dehyphenizer:    { source: sys-dic }
    # Wortstämme für nicht erkannte Wörter einfügen
#   - stemmer:         { }
    # Nicht erkannte Wörter auf Kompositum testen
    - decomposer:      { source: sys-dic }
    # Mehrwortgruppen im Strom erkennen
    - multi_worder:    { source: sys-mul }
    # Wortsequenzen anhand von Regeln identifizieren
    - sequencer:       { stopper: PUNC,OTHR }
    # Relationierungen einfügen
    - synonymer:       { skip: '?,t', source: sys-syn, out: syn }
    ########################################
    # Datenstrom anzeigen
    #   - debugger:        { eval: 'true', ceval: 'cmd!="EOL"', prompt: 'lex:) ' }
    ########################################
    # Ergebnisse ausgeben
    #
    # Erstelle Datei mit Endung .log für Datenstrom
    - vector_filter:   { in: syn, debug: 'true', prompt: 'lex:) ' }
    - text_writer:     { ext: log, sep: "\n" }
    # Erstelle Datei mit Endung .non für nicht erkannte Wörter
    - noneword_filter: { in: syn }
    - text_writer:     { ext: non, sep: "\n" }
    # Erstelle Datei mit Endung .ste für Wortstämme
    - vector_filter:   { in: syn, lexicals: z }
    - text_writer:     { ext: ste, sep: "\n" }
    # Erstelle Datei mit Endung .vec für erkannte Indexterme
    - vector_filter:   { in: syn, lexicals: '^[ksavem]$' }
    - text_writer:     { ext: vec, sep: "\n" }
    # Erstelle Datei mit Endung .ven für erkannte Indexterme mit absoluter Häufigkeit
    - vector_filter:   { in: syn, lexicals: '^[ksavem]$', sort: term_abs }
    - text_writer:     { ext: ven, sep: "\n" }
    # Erstelle Datei mit Endung .ver für erkannte Indexterme mit relativer Häufigkeit
    - vector_filter:   { in: syn, lexicals: '^[ksavem]$', sort: term_rel }
    - text_writer:     { ext: ver, sep: "\n" }
    # Erstelle Datei mit Endung .mul für erkannte Mehrwortgruppen
    - vector_filter:   { in: syn, lexicals: m }
    - text_writer:     { ext: mul, sep: "\n" }
    # Erstelle Datei mit Endung .seq für erkannte Wortsequenzen
    - vector_filter:   { in: syn, lexicals: q, sort: term_abs }
    - text_writer:     { ext: seq, sep: "\n" }
    # Erstelle Datei mit Endung .syn für erkannte Synonyme
    - vector_filter:   { in: syn, lexicals: y, sort: term_abs }
    - text_writer:     { ext: syn, sep: "\n" }
```

5 Indexierungskonfiguration lir.cfg (Standardkonfiguration)

```
# Lingo-Konfiguration für den Test mit einer LIR-Datei
#
# Gebräuchliche Patterns sind
#
#   "^\021(\d+\-\d+)\022"
#   "^\[(\d+)\.\]"
#---
meeting:

  attendees:

    ########################################
    # Text bereitstellen
    #

    # Angegebene Datei zeilenweise einlesen und verarbeiten
    - text_reader:     { files: $(files), records: true, progress: true }

    ########################################
    # Inhalte verarbeiten
    #

    # Zeile in einzelnen Sinnbestandteile (Token) zerlegen
    - tokenizer:       { }
    # Verbleibende Token im Wörterbuch suchen
    - word_searcher:   { source: sys-dic, mode: first }
    # Nicht erkannte Wörter auf Kompositum testen
    - decomposer:      { source: sys-dic }
    # Mehrwortgruppen im Strom erkennen
    - multi_worder:    { source: sys-mul }
    # Wortsequenzen anhand von Regeln identifizieren
    - sequencer:       { stopper: PUNC,OTHR }
    # Relationierungen einfügen
    - synonymer:       { skip: '?,t', source: sys-syn, out: syn }


    ########################################
    # Datenstrom anzeigen
    #
#   - debugger:        { eval: 'true', ceval: 'cmd!="EOL"', prompt: 'lex:) ' }


    ########################################
    # Ergebnisse ausgeben
    #
    # Erstelle Datei mit Endung .log für Datenstrom
    - vector_filter:   { in: syn, debug: 'true', prompt: 'lex:) ' }
    - text_writer:     { ext: log, sep: "\n" }
    # Erstelle Datei mit Endung .non für nicht erkannte Wörter
    - noneword_filter: { in: syn }
    - text_writer:     { ext: non, sep: '|' }
    # Erstelle Datei mit Endung .vec für erkannte Indexterme
    - vector_filter:   { in: syn, lexicals: '^[ksavem]$' }
    - text_writer:     { ext: vec, sep: '|' }
    # Erstelle Datei mit Endung .mul für erkannte Mehrwortgruppen
    - vector_filter:   { in: syn, lexicals: m }
    - text_writer:     { ext: mul, sep: '|' }
    # Erstelle Datei mit Endung .seq für erkannte Wortsequenzen
    - vector_filter:   { in: syn, lexicals: q, sort: term_abs }
    - text_writer:     { ext: seq, sep: '|' }
    # Erstelle Datei mit Endung .syn für erkannte Synonyme
    - vector_filter:   { in: syn, lexicals: y, sort: term_abs }
    - text_writer:     { ext: syn, sep: '|' }
```

6 Wörterbuchkonfiguration de.lang (Ausschnitt Standardkonfiguration)

```
#    lingo language definition
---
language:
  name: 'Deutsch'

  dictionary:
    databases:
      #    Systemwörterbücher
      sys-dic: { name: de/lingo-dic.txt, txt-format: WordClass, separator: '=' }
      sys-abk: { name: de/lingo-abk.txt, txt-format: WordClass, separator: '=' }
      sys-syn: { name: de/lingo-syn.txt, txt-format: KeyValue, separator: '=', def-wc: y }
      sys-mul: { name: de/lingo-mul.txt, txt-format: SingleWord, use-lex: 'sys-dic', def-wc: m }
      #    Benutzerwörterbücher
      usr-dic: { name: de/user-dic.txt, txt-format: WordClass, separator: '=' }
      sin-dic: { name: de/sinnfrei-dic.txt, txt-format: WordClass, separator: '=' }
      prs-dic: { name: de/pers-dic.txt, txt-format: WordClass, separator: '=' }
      nme-dic: { name: de/name-dic.txt, txt-format: SingleWord, def-wc: m }

      #    Testwörterbücher
      tst-dic: { name: de/test_dic.txt, txt-format: WordClass }   #  TEST: Lesen von zwei Quellen
      tst-syn: { name: de/test_syn.txt, txt-format: MultiValue, def-wc: 'y' } # TEST: Mehrere Datenquellen
      tst-syn2: { name: de/test_syn2.txt, txt-format: MultiValue, def-wc: 'y', use-lex: 'sys-dic' } # TEST:
Mehrere Datenquellen
      tst-mul: { name: de/test_mul.txt, use-lex: 'sys-dic', def-wc: m } # TEST: Mehrere Multiwörterbücher
      tst-mu2: { name: de/test_mul2.txt, use-lex: 'sys-dic', def-wc: m } # TEST: Mehrere Multiwörterbücher
      tst-sto: { name: de/test_store.txt, txt-format: WordClass } #  TEST: korrespondierende Store-Datei nicht
vorhanden
      tst-cry: { name: de/test_cry.txt, txt-format: WordClass, crypt } #  TEST: Verschlüsselung
      tst-sgw: { name: de/test_singleword.txt, txt-format: SingleWord } #  TEST: SingleWord-Format


    compound:
      min-word-size: "7"
      min-part-size: "3"
      max-parts: "5"
      min-avg-part-size: "4"
      append-wordclass: "+"
      skip-sequences: [ xx ]


    suffix:
      #    Suffixliste, Stand: 30-06-2005
      #    Suffixklasse:   s = Substantiv, a = Adjektiv, v = Verb, e = Eigenwort, f = Fugung
      #    Suffixe je Klasse:   "<suffix>['/'<ersetzung>][ <suffix>['/'<ersetzung>]]"
      - [s, "e en er ern es n s se sen ses"]
      - [a, "este ste ster sten stes ester estes esten e em en er ere eren erer eres es erem"]
      - [v, "e/en en/en est/en et/en st/en t/en te/en ten/en eten/en ete/en etest/en s"]
      - [e, "s"]
      - [f, "s n e en es er ch/che /en"]

sequencer:
      sequences: [ [AS, "1 2"], [AK, "1 2"], [AAK, "1 2 3"], [AAS, "1 2 3"], [EE, "1 2"] ]
```

III Datenverzeichnis

Dateiname	Verwendung	Dateiordner
bamul-dic.txt	neu erstelltes MWG-Wörterbuch ba-mul	RDK_Wörterbücher
bardkall-dic.txt	Rechtschreibwörterbuch zur Identifizierung aller Bestandteile von MWGs ba-dic	RDK_Wörterbücher
bardk-dic.txt	Rechtschreibwörterbuch kunsthistorische Terminologie brd-dic	RDK_Wörterbücher
funkw-dic.txt	Funktionswortwörterbuch fkt-dic	RDK_Wörterbücher
rdk_dic.txt	allgemeines Rechtschreibwörterbuch dic-sys	RDK_Wörterbücher
basetall.txt	RDK-Indexierungsdatei	RDK_Textdatei_Lexikonartikel
seite.dbm	Datenbankdatei der RDK-Lexikonartikel	RDK_DB_Lexikonartikel
ba-rdk-lir.cfg	Indexierungskonfiguration für Indexierung mit multiworder	Konfigurationsdateien
ba-rkd.ling.cfg	Indexierungskonfiguration für Indexierung mit sequencer	Konfigurationsdateien
de.lang	Wörterbuchkonfiguration	Konfigurationsdateien
basetall3.ven	Ergebnisdatei dreiteiliger MWGs mit Häufigkeitsangabe	RDK_Ergebnisdateien
basetall3.seq	Ergebnisdatei dreiteiliger MWGs mit Wortmuster	RDK_Ergebnisdateien
basetall4.ven	Ergebnisdatei vierteiliger MWGs mit Häufigkeitsangabe	RDK_Ergebnisdateien
basetall4.seq	Ergebnisdatei vierteiliger MWGs mit Wortmuster	RDK_Ergebnisdateien
basetall5.ven	Ergebnisdatei fünfteiliger MWGs mit Häufigkeitsangabe	RDK_Ergebnisdateien
basetall5.seq	Ergebnisdatei fünfteiliger MWGs mit Wortmuster	RDK_Ergebnisdateien
basetall6.ven	Ergebnisdatei sechsteiliger MWGs mit Häufigkeitsangabe	RDK_Ergebnisdateien
basetall6.seq	Ergebnisdatei sechsteiliger MWGs mit Wortmuster	RDK_Ergebnisdateien
basetall.mul	Ergebnisdatei MWGs nach Indexierung mit multiworder	RDK_Ergebnisdateien

IV Eingesetzte Software

Lingo: Version 1.8.2
Download unter: http://lex-lingo.blogspot.de/;

Midos 6: Version 1.6a
Download unter: http://www.progris.de/index.html?/midost.htm (Download einer kostenlosen Demoversion)

Notepad++:
Download unter: http://notepad-plus-plus.org/download/v6.4.3.html

Hiermit versichere ich, die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben.

Köln, den 29.08.2013 _____