Automatische Extraktion von Fachterminologie aus kunsthistorischen Volltexten

Juliane Bredack
julianebredack@web.de
Fachhochschule Potsdam, Master Studiengang Informationswissenschaften

Abstract: Mit Hilfe eines algorithmisch arbeitenden Verfahrens können fachterminologische Mehrwortgruppen aus elektronisch vorliegenden Texten identifiziert und extrahiert werden. Inhaltlicher Schwerpunkt stellt die Einbindung von Funktionswörtern des deutschen Sprachgebrauchs in den Extraktionsalgorithmus dar. Als Datengrundlage dieser Arbeit dienten kunsthistorische Lexikonartikel des Reallexikons zur Deutschen Kunstgeschichte. Das automatische Indexierungssystem Lingo wurde in dieser Studie genutzt. 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.

1 Einleitung

Inhaltsrelevante Informationen eines Fachtextes werden durch Fachwörter repräsentiert. Die entsprechende Fachterminologie, wie sie beispielsweise durch Sachbegriffe oder Eigennamen gekennzeichnet ist, liegt nicht nur in Form von Einzelwörtern vor. Konzepte bzw. Inhalte finden sich oft in Mehrwortbegriffen bzw. Fachphrasen wieder. Diese Mehrwortgruppen sind als zusammenhängende lexikalische Einheiten zu sehen, welche als Ganzes eine Bedeutung aufweisen [Le13]. Die miteinander stehenden Begriffe, wie sie z.B. in Adjektiv-Substantiv Verbindungen auftreten, erzeugen eine hohe Aussagekraft und Deutlichkeit, die ein einzelner Begriff oft nicht erreicht. Ein separat betrachteter Term kann verschiedene Inhalte ausdrücken, wohingegen der inhaltsrelevante Sinn oder die eindeutige Zuordnung zu einem Thema innerhalb einer Mehrwortgruppe sichtbar wird. Somit lassen Mehrwortgruppen eine Bedeutungsdifferenzierung auf der Ebene eines einzelnen Wortes zu [GLN12].

In Bezug auf elektronische Dokumente bzw. Dokumentkollektionen und deren inhaltlicher Erschließung, leisten automatische Indexierungsverfahren eine zuverlässige Identifizierung einzelner, inhaltsrelevanter Terme um diese in einem Index aufzunehmen. Ein zuverlässiges Suchen und Finden im Rahmen des Information Retrieval wird gewährleistet [Le13]. Neben Einzeltermen sollten jedoch auch Mehrwortgruppen im Index hinterlegt sein, um die Suchmöglichkeiten auszuweiten. Ein Dokument kann durch den Einsatz einer Mehrwortgruppe gezielt mit dieser gefunden werden, anstelle von mehreren Dokumenten, die die einzelnen Begriffe beinhalten und so unnötig viele Informationen

bereitstellen, die nicht im Sinne eines Suchenden sind. Im Folgenden wird deshalb ein Verfahren beschrieben, das eine Identifizierung von Mehrwortgruppen zulässt, um diese für ein späteres Retrieval zu nutzen.

Das für die Untersuchung eingesetzte Indexierungssystem Lingo basiert auf linguistischen Methoden und leistet eine Identifizierung von Mehrwortgruppen, wenn diese in einem Wörterbuch lexikalisiert, also vorab bekannt sind (z.B. Nutzung von kontrolliertem Fachvokabular beim Wörterbuchaufbau) [GLN12]. Die lexikalische Funktionalität wurde bereits im Digitalisierungs-, Erschließungs- und Indexierungsprojekt RDK-Web zur Erstellung einer Web-Retrievalumgebung des Reallexikon zur Deutschen Kunstgeschichte (RDK) getestet [Le06], welche sich als zweckmäßig erwiesen hat [Rea14]. Die inhaltliche Komplexität und Vielzahl der RDK Lexikonartikel lässt allerdings vermuten, dass die Anzahl unbekannter Mehrwortgruppen als weitaus höher einzuschätzen ist. Neben einer Indexierung mit bereits bekanntem Vokabular (lexikalische Funktion) ist es mit Lingo und dessen Programmmodul sequencer möglich noch unbekannte und potentiell aussagekräftige Mehrwortkonstruktionen auf algorithmischem Weg zu identifizieren.

Die automatische Identifizierung und Extraktion fachterminologischer Mehrwortgruppen erfolgte am Beispiel kunsthistorischer Fachtexte. Als Datenquelle dienten die deutschsprachigen Lexikonartikel des Reallexikon zur Deutschen Kunstgeschichte.

2 Besonderheiten kunsthistorischer Mehrwortbegriffe im RDK

Das RDK ist ein Nachschlagewerk zur Realienkunde der Kunstgeschichte, welches seit 1937 erscheint. Inhaltliche Schwerpunkte der Lexikonartikel liegen auf Architektur, Bildenden Künsten, Kunsthandwerk, Materialien, Technik und Ikonographie des deutschen Sprachgebiets. Herausgeber ist das Zentralinstitut für Kunstgeschichte in München. Die Bände des RDK erscheinen als fortsetzendes Lieferwerk [Zen14].

Im folgenden Ausschnitt des Lexikonartikels „Christus“ (Band 3, Spalte 611, RDK) sind Mehrwortbegriffe, die kunsthistorische Inhalte wiedergeben, kursiv hervorgehoben.

> Obwohl, abgesehen von der Sarkophagplastik, der erhaltene Bestand altchristlicher Monumente mit figuralem Schmuck verhältnismäßig gering ist, so ergibt sich doch bei genauer Durchsicht, daß fast alle in der Kunst des frühen MA dargestellten Szenen aus dem Leben Christi schon irgendwie in der altchristlichen Kunst vorkommen. Diese hat - neben der Verkündigung an Maria (mit apokryphen Zutaten) und der Heimsuchung Elisabeths - dargestellt: die Geburt Christi, die Anbetung der Weisen, die Darbringung im Tempel, die Flucht nach Ägypten (mit apokryphen Zutaten), den Kindermord, den zwölfjährigen Jesus im Tempel, die Taufe Jesu im Jordan, das erste Wunder Christi bei der Hochzeit zu Kana, die Brotvermehrung, die Heilung des Blinden, […]

Es wird deutlich, dass sich kunsthistorische Begriffe in Mehrwortgruppen aus einer Kombination von Sachbegriffen (Darbringung im Tempel), Sachbegriffen und Personennamen (Verkündung an Maria) oder Geografika (Flucht nach Ägypten) zusammen-

setzen. Weiterhin fällt auf, dass sich Mehrwortgruppen nicht ausschließlich aus feststehenden Kombinationen von Adjektiven und Substantiven bilden, sondern auch Funktionswörter enthalten. Unter Funktionswörtern sind in diesem Fall Artikel, Konjunktionen und Präpositionen zu verstehen.

Funktionswörter stellen eine vernachlässigte Wortart im Retrieval dar, da sie selbst keine inhaltlich relevanten Informationen vermitteln und so als Stoppwörter unbeachtet bleiben. Innerhalb einer Mehrwortgruppe erweitern sie allerdings deren syntaktische Struktur und stellen die nötige Verbindung zwischen inhaltsrelevanten Begriffen her. So verbinden Funktionswörter fachterminologische Begriffe zu einer Einheit, sodass komplexere Inhalte vermittelt werden.

Zusammenfassend lassen sich die Besonderheiten von Mehrwortgruppen und Funktionswörtern mit Bezug zur Kunstgeschichte wie folgt festhalten:

- Verknüpfung von Fachtermen durch Funktionswörter zu syntaktisch abgeschlossenen Mehrwortgruppen, z.B. „Szene aus dem Leben Jesu“. Als Verbindung zwischen inhaltsrelevanten Termen kommt neben dem Artikel dem in diesem Beispiel eine Präposition zum Einsatz. Zu beachten ist, dass ein und dieselbe Präposition unterschiedliche Beziehungen ausdrückt. So kann aus auch ein örtliches Verhältnis beschreiben, wie „Altar aus Kloster Buxheim“. Ein bestimmter Sachverhalt lässt sich allerdings auch durch verschiedene Präpositionen ausdrücken, „Darstellungen zum Leben Jesu“.

- Der Artikelausschnitt „Christus“ enthält als Indexterme in Frage kommende Begriffe wie „Leben“, „Szene“, „Tempel“ oder „Kunst“. Durch den Einsatz von Mehrwortgruppen lassen sich eindeutigere Aussagen zum Artikelinhalt treffen, z.B. „dargestellte Szene aus dem Leben Christi“, „zwölfjähriger Jesus im Tempel“, „monumentale Kunst der altchristlichen Zeit“ oder „Szene aus dem Leben Jesu“. Eine inhaltliche Spezifizierung wird deutlich.

- Bedeutungsdifferenzierung von Begriffen, z.B. „Zeit“, welcher als Indexterm zu allgemein ist, jedoch innerhalb einer Mehrwortgruppe „Kunst der altchristlichen Zeit“ seinen Zweck erfüllt.

Als vorrangig wird in dieser Studie die Integration von Funktionswörtern (Artikel, Konjunktionen, Präpositionen) in den Extraktionsalgorithmus angesehen, um unbekannte und komplex strukturiertere Mehrwortbegriffe mit kunsthistorischem Aspekt zu extrahieren. So wird im Folgenden eine Gruppe, bestehend aus mindestens drei Wörtern, als Mehrwortgruppe angesehen.[1]

[1] Potentielle Mehrwortgruppen, die sich auch aus zwei Bestandteilen zusammensetzen könne (z.B. Adjektiv-Substantiv Verbindungen) blieben in dieser Studie unberücksichtigt.

3 Automatische Identifizierung und Extraktion von Mehrwortgruppen

3.1 Das Indexierungssystem Lingo

Das Indexierungssystem Lingo[2] ist wörterbuchgestützt. Die durch eine Indexierung gewünschten Ergebnisse werden flexibel mittels der dafür vorgesehenen Konfigurationsdateien gesteuert. Einzelne Programmmodule realisieren eine Indexierung, die in ihrer Funktionsweise und Ergebnissen aufeinander aufbauen. So stützt sich das Programmmodul sequencer auf die vorab im Indexierungsprozess erfolgte grammatikalische Normierung des Textes. Die korrekte Identifizierung aller Zeichenketten setzt deren Lexikalisierung in einem Grundformwörterbuch voraus [LV06]. Lingo Wörterbücher für eine Grundformerkennung sind folgendermaßen aufgebaut:

```
automatische=automatisch #a
extraktion=extraktion #s
fachterminologischer=fachterminologisch #a
mehrwortgruppe=mehrwortgruppe #s
mit=mit #c
lingo=lingo #s
```

Links vom Gleichheitszeichen befindet sich die Wortform und rechts davon die zugeteilte Grundform. Jeder Eintrag ist zudem mit einer Wortklasse, die Angaben zur Wortart (z.B. Adjektiv #a oder Substantiv #s) enthält, gekennzeichnet. Durch die im Wörterbuch eingetragenen Terme werden die Bestandteile des folgenden Satzes nach dem Indexierungsprozess zuverlässig identifiziert.

„Automatische Extraktion fachterminologischer Mehrwortgruppen mit Lingo.“

Eine Indexierung führt zu diesem Ergebnis:

```
<Automatische=[(automatisch/a)]>
<Extraktion=[(extraktion/s)]>
<fachterminologischer=[(fachterminologisch)/a]>
<Mehrwortgruppen=[(mehrwortgruppe/s)]>
<mit=[(mit/c)]>
<Lingo=[(lingo/s)]>
:./PUNC
```

Nach einer Indexierung sind die Wortklassen der einzelnen Satzbestandteile bekannt, was sich für eine algorithmische Identifizierung von Mehrwortgruppen mit dem sequencer nutzen lässt. Aus dem Wissen heraus, das jeder Term in seiner Bedeutung gekennzeichnet ist, kann durch das Bilden von Wortmustern, basierend auf den Wortklassen eines Grundformwörterbuchs, eine Identifizierung von Wortgruppen erfolgen.

[2] http://lex-lingo.blogspot.de/

Beispielsweise lässt das Muster AS eine Identifizierung von Adjektiv-Substantiv Wortfolgen zu, wie im Folgenden zu sehen ist.

Automatische [a] Extraktion [s] fachterminologischer [a] Mehrwortgruppen [s]

AS automatisch extraktion
AS fachterminologisch mehrwortgruppe

Voraussetzungen für eine zuverlässige Identifizierung von Mehrwortgruppen bilden zum einen Grundformwörterbücher mit den darin enthaltenen Termen, getaggt durch eine Wortklasse. Zum anderen müssen die auf den Wortklassen basierenden Identifizierungsmuster in der dafür vorgesehenen Lingo-Konfiguration eingebunden sein. Als Ergebnis einer erfolgten Identifikation und Extraktion, werden die Mehrwortgruppen als Listen in einer Ergebnisdatei ausgegeben. Deren Größe richtet sich nach der zu indexierenden Datei bzw. Dokumentkollektion. Die Mehrwortbegriffe werden, wie am oben gezeigten Beispiel zu sehen, in der Grundform ausgegeben [GLN12].

3.2 Wörterbuchaufbau und Wortklassen der RDK-Indexierung

Die algorithmische Identifizierung neuer, kunsthistorischer Mehrwortkonstruktionen mittels sequencer setzt voraus, dass Wörterbücher mit geeignetem Vokabular existieren. Für eine Indexierung mit Lingo können die Wörterbücher als einfache Textdateien abgespeichert und mit einem Texteditor bearbeitet werden. Für die Identifizierung fachspezifischer Inhalte wurde auf die, aus der RDK-Web Indexierung hervorgegangenen Wörterbücher und deren fachterminologischen Einträge zurückgegriffen [Le06]. Die RDK-Web Wörterbücher umfassten die für die Kunstgeschichte typischen Terme von Sachbegriffen, Personennamen und Geografika. Für die erneute Indexierung des RDK wurden alle Fachterme in einem einzigen Wörterbuch hinterlegt. In einem extra geschaffenen Wörterbuch wurden Artikel, Konjunktionen und Präpositionen lexikalisiert, um syntaktisch komplexere Mehrwortbegriffe zu erzeugen. Zusätzlich kam ein drittes Wörterbuch zum Einsatz, in dem Begriffe, die der Alltagssprache angehören, identifiziert werden.

Aufbauend auf den neu erstellten Wörterbüchern, wurden deren Einträge um neue Wortklassen erweitert. So wurden kunsthistorische Sachbegriffe, Geografika und Personenamen mit der Wortklasse E getaggt. Es erfolgte eine separate Kennzeichnung aller Funktionswörter, damit diese gezielt in den Extraktionsalgorithmus eingebaut werden konnten. Die differenzierte Kennzeichnung sollte zudem den Nutzen jener speziellen Wortarten innerhalb einer Mehrwortgruppe verdeutlichen. Artikel bekamen ein R, Konjunktionen ein U und Präpositionen die Wortklasse C zugewiesen. Weiterhin wurden für die Musterbildung die Lingo Standardwortklassen A (Adjektive) und K (Komposita) genutzt. Komposita entstehen im Rahmen des Indexierungsprozesses.[3] Ihnen wird automatisch die Wortklasse K zugeteilt (s. Tabelle 1).

[3] Die Wortbildungsmöglichkeiten im deutschen Sprachgebrauch sind vielfältig. Deswegen werden mit Lingo keine Wörterbücher lexikalisierter Komposita aufgebaut. Diese werden im Indexierungsprozess algorithmisch durch einzelne Wortbestandteile identifiziert und extrahiert [GLN 12].

Wortklasse	Bedeutung
E	Kunsthistorische Fachbegriffe
A	Adjektiv
K	Komposita
C	Präposition
R	Artikel
U	Konjunktion

Tabelle 1 zeigt die verwendeten Kennzeichnungen der Wortklassen und deren Bedeutung

3.3 Kriterien zur Bildung der Extraktionsmuster

Als Basis zur Musterbildung dienten zuvor festgelegte Kriterien. Diese folgen typischen Erscheinungsformen zusammengesetzter Terme im Deutschen. Beispielsweise stehen Adjektive als Beiwort vor einem Substantiv (mit Wortklasse E und K gekennzeichnete Terme), da diese durch das Adjektiv näher definiert werden. So werden Adjektiv-Substantiv Verbindungen als vielfach eingesetzte Kombination in den Wortmustern verwendet, z.B. ECAE – „Dreizack auf antiker Darstellung“. Funktionswörter wurden gezielt platziert, damit sie ihren syntaktischen Zweck erfüllen. Zum Beispiel wurden Artikel vorzugsweise vor Substantiven oder Adjektiv-Substantiv Verbindungen (AE, AK) platziert, wie es die Extraktionsmuster ERAE – „Symbol des eucharistischen Christus“ oder AERAE – „berühmte Komposition der italienischen Malerei“ verdeutlichen. Mehrwortgruppen beginnen und enden nicht mit einem Funktionswort. Einerseits würden Wortmuster dieser Art unvollständige Mehrwortgruppen erzeugen, da ein Funktionswort immer in Verbindung mit einer zusätzlichen Wortklasse, wie dem Substantiv, zu sehen ist. Anderseits wird in diesem Fall nur der Inhalt einer verkürzten Mehrwortgruppe repräsentiert, da Funktionswörter keine eigenständige, inhaltstragende Bedeutung besitzen. Demzufolge wurden die Wortklassen von Funktionswörtern nur innerhalb der eingesetzten Extraktionsmuster positioniert. Die Muster setzen sich aus mindestens drei, maximal aus sechs Bestandteilen zusammen, denn simple Wortmuster wie AE oder AK sind in dem Anwendungsfall nicht zielführend. Dieses Vorgehen bot die Möglichkeit eine Vielzahl von Mehrwortgruppen unterschiedlicher Länge, unter Berücksichtigung aller zur Verfügung stehender Wortklassen und deren als sinnvoll zu erachtender Kombinationen, zu bilden und zu testen.

Ein feststehendes Set von Wortmustern wurde nach umfangreichen Testindexierungen festgelegt.[4] Die für die Untersuchung genutzten Wortmuster umfassten 80 Einträge:

[4] Hierbei handelt es sich um ein Testset verwendeter Wortmuster. Es soll deshalb nicht ausgeschlossen werden, dass weitere Muster existieren, die den zuvor festgelegten Kriterien folgen.

Muster mit drei Bestandteilen:

EEE, KEE, EEK, EKK, ARE, ACE, AUE, ERE, KCE, ECK, EUE, EUK, KUE, KRE, ECE, EAE, KAE, AEE, KUK, EKE

Muster mit vier Bestandteilen:

ECEE, EUEE, EUEK, AECE, AKRK, AKUK, ECAE, ERAE, KRAE, ERAK, KCAE, AKUE, AKRE, KUEE, KCAK, EUKE, EAEE, ECKE, ERKE, EARE

Muster mit fünf Bestandteilen:

ACCEE, ACRAE, AERAK, AERAE, EERAE, AECEE, ACEUE, ECRAE, ECCAE, ECCEE, ECREE, EUERE, KUERE, EUEAE, ECEUK, ECAEE, KCREE, KCCAE, EEREE, EEUEE

Muster mit sechs Bestandteilen:

AECCEE, EECCAE, AKCRAE, KRAEAE, EUECAK, EUKRAE, KACRAE, EACRAE, EUECKE, EUERKE, AECEUE, EUECRE, AKCKUE, ECAKRE, ACEUEE, AECEEE, AKCAEE, ECRAEE, EUKRAE, ERAEAE

4 Kriterien zur Bewertung fachterminologischer Mehrwortgruppen

Zur qualitativen Beurteilung der Extraktionsmuster und den daraus resultierenden Mehrwortgruppen wurden Kriterien erstellt, an denen sich die darauffolgende Analyse orientiert.

- Abgeschlossenheit: Mehrwortbegriffe sollen nach der Extraktion als zusammenhängende Einheit (semantisch abgeschlossen) und nicht als unvollständige Satzfragmente vorliegen. Zum Beispiel stellen „abgesehen von Kloster und Wallfahrtskirche können“ oder „abendländisch von Anfang“ keine repräsentativen Mehrwortgruppen dar.

- Kombination aus Personennamen bzw. Geografika: Personennamen oder Ortschaftsbezeichnungen, die sich spezifiziert durch zusätzliche Begriffe zusammensetzen, zum Beispiel „Barocke Freskomalerei in Schlesien“ oder „Holzschnitt von Georg Lemberger“. Die Identifizierung von Personennamen und/oder Geografika erzeugt kunsthistorischen Inhalt.

- Zusammensetzungen kunsthistorischer Sachbegriffe: Mehrwortgruppen, welche sich aus zwei Fachtermen zusammensetzen. Diese repräsentieren bereits auf der Wortebene kunsthistorische Terminologie, wie durch die Mehrwortgruppen

„Wallfahrt und Pfarrkirche Maria“ oder „Sündenfall und Erlösung Christus“ deutlich wird.

- Zusammensetzungen kunsthistorischer Sachbegriffe und Alltagssprache: Mehrwortgruppen werden als positiv erachtet, wenn sich diese aus Begriffen zusammensetzen, die allgemeinerer Natur sind, jedoch durch Fachtermini präzisiert werden, wie „Verständnis der biblischen Symbolsprache“ zeigt.

5 Analyse

Die Indexierung des RDK umfasst ca. 8.000 doppelspaltige Lexikonseiten. Vier- und fünfteilige Erkennungsmuster lieferten die besten Ergebnisse, wohingegen drei- und sechsteilige Muster überwiegend inhaltlich unvollständige, nicht abgeschlossene Mehrwortgruppen identifizierten.

Die nachfolgende Analyse zeigt einen Ausschnitt positiver Ergebnisse, im Sinne der vorab genannten Kriterien, welche aus der Identifizierung von vier- und fünfteiligen Erkennungsmustern resultieren.

Wortmuster, welche mit einem substantivischen Begriff (Wortklasse E oder K) eingeleitet werden, gefolgt von einem Funktionswort (Artikel oder Präposition), welcher den ersten Term mit einer Adjektiv-Substantiv-Wortfolge in Verbindung setzt, erzeugen durchweg positive Ergebnisse.

ERAE	„theorie der schön kunst“
	„symbol der jungfäulichen geburt“
ECAE	„abwehrzauber gegen böse dämon“
	„abbildung von neun erzengel“
KRAE	„edelmetallkunst der katholisch kirche“
	„doppelfunktion der architektonisch wirkung“
ERAK	„entwicklung des abendländisch stufenportal“
	„abbildung des jüdisch tempelbau“
KCAE	„elfenbeinrelief in silbervergoldet rahmung“
	„bilderkreis auf alttestamentlich szene“
KCAK	„kunstlandschaft im romanisch kleinkirchenbau“
	„grundrißgestaltung an rheinisch bettelordensbasiliken“

ERKE „darstellung der himmelserscheinung maria“

„personifikation des wegweisend stern“

Auch anhand der getesteten Wortmuster AECE, AKRK und AKUE, welche mit einem Adjektiv beginnen, zeigen sich die positiven Ergebnisse. Das Funktionswort (Präposition, Artikel oder Konjunktion) verbindet eine Adjektiv-Substantiv-Verbindung (AE oder AK) und einen einzelnen Fachterm (E oder K). Oder im Falle fünfteiliger Muster, durch die Kombination: Adjektiv-Substantiv-Verbindung – Funktionswort – Adjektiv-Substantiv-Verbindung.

AECE „adorierend engel auf giebel“

„allegorisch auslegung auf sündenfall“

AKRK „früh holzschnittdarstellung einer schulszene“

„spiritualistisch lichtmetaphysik des neuplatonismus“

AKUE „weiß leinentuch als altarbekleidung“

„malerisch bildgestaltung und farbgebung“

AERAE „episch dichtung der höfisch zeit“

„eucharistisch bedeutung des gekreuzigt christus“

AERAK „künstlerisch möglichkeit des optisch farbaufbau“

„lehrhaft erklärung der antik götterbild“

Die Relevanz von Personenamen oder geografischen Begriffen in der kunstgeschichtlichen Terminologie muss auch bei der Extraktion entsprechender Mehrwortgruppen berücksichtigt werden, die diese Begriffe als Bestandteile enthalten. Es zeigte sich, dass vor allem Muster bestehend aus EE, Personennamen in der Kombination Vor- und Nachname, extrahieren. Da es sich bei den mit E getaggten Begriffen allerdings auch um Sachbegriffe handeln kann, sind Kombinationen von Vor-oder Nachnamen möglich, die durch einen Sachbegriff präzisiert werden.

EEREE „historia scholastica des petrus comestor“

“burgmair totenbild des Conrad celtis”

ECEE “altar von georg raphael”

„aquarell von william blake“

ECREE „romgedanke in der kunst bernini“

„christus in der protestant kunst“

Die durch ein E extrahierten Geografika werden in Verbindung mit einem Sachbegriff (E) oder Adjektiv-Substantiv-Wortfolgen (AE) spezifiziert.

KCREE	„wandgemälde in der capella greca“
AECEE	„dorisch portal am schloß aschaffenburg“
EEREE	„decretum gratiani der paris bibliothek“
ECEE	“altar im wien stephansdom”
ECREE	„altartuch aus dem zisterzienserinnenkloster zehdenick“

6 Fazit

Die positiven Ergebnisse der Analyse zeigen, dass das vorgenommene algorithmische Verfahren zur Mehrwortgruppenerkennung und Extraktion zweckmäßig ist. Das Vorgehen wurde bereits an Dokumenten mathematischen Inhalts getestet [Gö12]. Nun belegt sich die Tauglichkeit der Untersuchung auch an kunsthistorischen Datenbeständen.

Durch die eingesetzte Indexierungssoftware Lingo und dessen Programmmodul sequencer konnten fachterminologische Mehrwortgruppen mit kunsthistorischem Bezug extrahiert werden. Die Integration von Funktionswörtern in den Extraktionsalgorithmus ermöglicht zudem, syntaktisch komplex strukturiertere Mehrwortkonstruktionen zu identifizieren.

Relevant für die Erzeugung fachterminologischer Mehrwortgruppen sind die gezielt erstellten Erkennungsmuster. In der Untersuchung wurde nur ein Testset potentieller Wortmuster erstellt und analysiert. Das Bilden weiterer Muster anhand der gebildeten Kriterien hat das Potential zusätzliche, fachterminologische Mehrwortgruppen zu erzeugen. In weiteren Indexierungen können Wortmuster festgelegt werden, die fachterminologische Mehrwortgruppen in Fachtexten identifizieren, aber negative Ergebnisse von vornherein reduzieren.

Die erzeugten Mehrwortgruppen, welche durch eine Indexierung mit dem sequencer extrahiert werden, können als Empfehlung angesehen werden, um diese nach einer intellektuellen Sichtung, hinsichtlich ihrer fachspezifischen Qualität weiter zu verarbeiten. So ist es mit Lingo möglich, Wörterbücher mit Mehrwortgruppen aufzubauen, in denen diese in ihrer grammatikalisch korrekten Form (also nicht in der Grundform) lexikalisiert werden. So kann von vornherein eine Indexierung mit Mehrwortgruppen erfolgen, welche ausschließlich fachliche Relevanz besitzen.

Danksagung

Die vorliegende Arbeit entstand auf Basis meiner Bachelorarbeit am Institut für Informationswissenschaften der Fachhochschule Köln im Studiengang Bibliothekswesen. Meinem besonderen Dank gilt dem Betreuer der Arbeit Prof. Dr. Klaus Lepsky für seine Unterstützung und die Beantwortung zahlreicher Fragen, die wesentlich zum Gelingen der Arbeit beitrugen.

Literaturverzeichnis

[GLN12] Gödert, W.; Lepsky, K.; Nagelschmidt, M.: Informationserschließung und Automatisches Indexieren: Ein Lehr- und Arbeitsbuch. Springer, Berlin, 2012.

[Gö12] Gödert, W.: Detecting multiword phrases in mathematical text corpora. Köln, 2012 http://arxiv.org/ftp/arxiv/papers/1210/1210.0852.pdf

[Le06] Lepsky, K.: Automatische Indexierung des Reallexikons zur Deutschen Kunstgeschichte. In (Harms, I., Giessen, H. W., Luckhardt, H-D., Hrsg.): Information und Sprache: Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern. De Gruyter, Berlin, 2006; S. 169-178.

[Le13] Lepsky, K.: Automatische Indexierung. In (Kuhlen, R., Semar, W., Strauch, D., Hrsg.): Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und –praxis. De Gruyter Saur, Berlin, 2013; S. 272-285.

LV06] Lepsky, K., Vorhauer, J.: Lingo: ein open source System für die Automatische Indexierung deutschsprachiger Dokumente. In ABI-Technik. Vol. 1, 2006; S. 18-28.

[REA14] Reallexikon zur Deutschen Kunstgeschichte. http://rdk.zikg.net/gsdl/cgi-bin/library.exe

[ZEN14] Zentralinstitut für Kunstgeschichte: Forschungsstelle Realienkunde. http://www.zikg.eu/forschung/forschungsstelle-realienkunde#Literatur