# Inhaltsverzeichnis

# 1 Einführungsteil

## 1.1 A multilingual Dictionary of Ophthalmology

Ausgangspunkt dieser Arbeit ist das „Digitale multilinguale Wörterbuch der Augenheilkunde“[1] des Augenarztes Dr. Philipp Franko Zeitz, das seit 2010 existiert und online auf der Webpräsenz der Düsseldorfer Augenklinik Zeitz Franko Zeitz zugänglich ist. Das Wörterbuch umfasst mittlerweile fast 25.000 manuell gepflegte Einträge aus dem Bereich der Augenheilkunde in 13 (anfänglich 8) Sprachen. Die Einträge enthalten Erklärungen, Abkürzungen, sowie Übersetzungs- und Synonymverweise, welche mittels eines einfach gestalteten Suchformulars gefunden werden können.

Seit 2011 sind sowohl die Funktionen, als auch der Inhalt des Wörterbuchs stetig erweitert worden. Ergebnisse sind, neben der nun breiten Sprachunterstützung, eine benutzerfreundliche Schnittstelle mit verbesserten Suchfunktionen und ein Bildatlas mit annähernd 2000 Bildern, die einige Einträge illustrieren.

Die Einträge des Wörterbuchs verteilten sich zum Zeitpunkt des Projektbeginns auf ca. 2500 Konzepte, die jeweils unterschiedliche Ausprägungen eines medizinischen Begriffs zusammenfassen (also alle Einträge des Wörterbuchs mit jeweils derselben Bedeutung, wie Synonyme, Übersetzungen und unterschiedliche Schreibweisen desselben Begriffs). In Anlehnung an den englischsprachigen Thesaurus WordNet werden diese Konzepte - beziehungsweise Mengen von synonymen Termen - im Nachfolgenden *Synsets* genannt. Synsets werden im Wörterbuch intern durch eine eindeutige Kennziffer ausgezeichnet, die das Identifizieren von Bedeutungen sehr einfach macht und damit einen sehr guten Ansatz zum Ausbau der Bedeutungsstruktur innerhalb des Wörterbuchs liefert.

## 1.2 Ideen und Zielsetzung

Ziel des nachfolgend vorgestellten Teamprojekts war es, die Einträge des „Wörterbuchs der Augenheilkunde“ zu erweitern um ihre jeweilige Repräsentation als formalen Begriff, um die Einträge auf diese Weise stärker miteinander zu verknüpfen und schließlich ein *semantisches*

[1] Erreichbar unter: http://www.zeitzfrankozeitz.de/index.php/fachwoerterbuch.html

*Netz* im Wörterbuch zu etablieren. Die Synsets des Wörterbuchs bilden im Sinne der Formalen Begriffsanalyse, die nachfolgend einführend erläutert wird, die Menge an Objekten. Um diese Objekte repräsentieren zu können werden beschreibende Merkmale benötigt. Die Erweiterung des Wörterbuchs beschäftigt sich also vornehmlich mit der Frage, wie diese Merkmale erzeugt und den entsprechenden Objekten, den Synsets, zugewiesen werden. Hierzu bieten sich Methoden der Informationsextraktion an, die im Hauptteil der Arbeit erläutert werden.

Die Synsets im Wörterbuch sind jedoch nicht gleichmäßig über vorhandene Sprachen verteilt. Das Deutsche und das Englische dominieren deutlich. Ein Grund dafür könnte unzureichendes Quellenmaterial sowie natürlich fehlende Fremdsprachenkenntnisse bei der Erstellung des Wörterbuchs gewesen sein. Ein anderer Grund könnte sein, dass es bestimmte Begriffe in der einen Sprache gibt und in anderen Sprachen nicht. Solcherlei Lücken, lexical gaps genannt, gibt es im „Wörterbuch der Augenheilkunde“ ebenfalls. Dieses Phänomen kann umso ausgeprägter werden, je spezifischer die Domäne ist, so wie beispielsweise die Augenheilkunde. Einträge mittels formaler Begriffe auszudrücken und damit über deren Merkmale prinzipiell übersetzbar zu machen in Sprachen, in denen es die Bezeichnung nicht gibt, ist eine Idee, die auf die Arbeit von Janssen, 2002 zurückgeht.

Weitere Probleme, die man mit dem vorgestellten Ansatz gut lösen kann, sind die der Homonymie und Synonymie. Ein Satz oder eine Phrase in zwei unterschiedlichen Sprachen hat (im Idealfall) in beiden Sprachen die selbe Bedeutung, verwendet aber andere Wörter um ihn auszudrücken. Für die wortweise Übersetzung zieht man, so vorhanden, ein Wörterbuch heran. Allerdings löst ein Wörterbuch nicht von allein das Problem der Homonymie, also der Frage welcher der Begriffe, die sich eine Bezeichnung teilen, im Kontext verwendet wurde. Auch ist nicht zwingend jede Bezeichnung, die für ein und denselben Begriff steht, Bestandteil eines gegebenen Wörterbuchs (Synonymie). In keinem der Fälle lässt sich dann zuverlässig ableiten, welches die korrekte Bezeichnung ist, es sei denn man zieht den Kontext, sofern erkennbar, heran. Formale Begriffe erlauben es, diesen Kontext herzustellen und sprachübergreifend zu verwenden, weil Begriffe dadurch in termini der ihnen zugewiesenen Merkmale ausgedrückt werden.

Ein weiterer ausschlaggebender Grund für den Einsatz von formalen Begriffen als Repräsentation der Synsets ist die Identifizierung der semantischen Nähe (Bedeutungsähnlichkeit) aller Synsets zueinander anhand gemeinsamer Merkmale. Der Anteil an gemeinsamen Merkmalen in Bezug auf die Gesamtanzahl zweier Synsets kann als Messwert für ihre semantische Nähe herangezogen werden, sprich je mehr Merkmale sich zwei Synsets teilen relativ gesehen zu der Summe ihrer Merkmale, desto ähnlicher sind sie sich in ihrer Bedeutung. In Kombination mit einer geeigneten Darstellung der Begriffsverbände wird den Nutzern des Wörterbuchs so die Möglichkeit gegeben, ausgehend von einem bekannten Begriff, in ihrer Bedeutung ähnliche Einträge des Wörterbuchs zu entdecken, um so spezifischere medizinische Begriffe ausfindig zu

machen oder einfach neue Begriffe zu entdecken.

## 1.3 Erste Überlegungen zur Umsetzung

Teil der Aufgabenstellung war es, die Formale Begriffsanalyse einzusetzen, um eine Zwischensprache aus Konzepten (Interlingua) zu entwickeln, über welche wiederum die im „Wörterbuch der Augenheilkunde“ vorhandenen Wörter unterschiedlicher Sprachen miteinander verbunden werden konnten.

### 1.3.1 Begriffe der Formalen Begriffsanalyse

Die Formale Begriffsanalyse ist ein Teil der mathematischen Ordnungslehre. Sie wurde in den 1980er Jahren von Rudolf Wille, Bernhard Ganter und Peter Burmeister entwickelt. In der Formalen Begriffsanalyse setzt sich ein Begriff zusammen aus dem Begriffsumfang und dem Begriffsinhalt. Ein formaler Begriff $(A, B)$ hat einen Kontext $(G, M, I)$ bestehend aus einer Menge von Gegenständen $G$, einer Menge von Merkmalen $M$ und einer Inzidenzrelation $I$ als Gegenstand-Merkmal-Beziehung (Ganter & Wille, 1996, S.58). Diese lässt sich in Form einer Kreuztabelle genannten Darstellung, typischerweise mit den Merkmalen als Spalten und den Gegenständen als Zeilen, anschaulich beschreiben. Die Kreuze in den Zellen dieser Tabelle bedeuten dann das Vorhandensein des Merkmals für den jeweiligen Gegenstand. Für eine beliebige Menge $A$ von Gegenständen aus einem formalen Kontext ist ihre Ableitung $A'$ die Menge der gemeinsamen Merkmale der Gegenstände aus $A$.

$$A' := \{m \in M \mid \forall g \in A : (g, m) \in I\}$$

Für eine beliebige Menge $B$ von Merkmalen aus einem formalen Kontext ist ihre Ableitung $B'$ die Menge der gemeinsamen Gegenstände der Merkmale aus $B$.

$$B' := \{g \in G \mid \forall m \in B : (g, m) \in I\}$$

Ist $A$ Umfang und $B$ Inhalt, so heißt $(A, B)$ formaler Begriff des Kontextes $(G, M, I)$ wenn *i.*) $A$ eine echte Teilmenge von $G$, und *ii.*) $B$ eine echte Teilmenge von $M$ ist, und *iii.*) die Menge $A'$ der gemeinsamen Merkmale der Gegenstände aus $A$ gleich $B$, und *iv.*) die Menge $B'$ der gemeinsamen Gegenstände der Merkmale aus $B$ gleich $A$.

Die Menge der Begriffe eines Kontextes sind damit alle Begriffe, die sich auf oben beschriebene Weise aus dem Kontext ermitteln lassen. Ein Unterbegriff ist dann ein Begriff, welcher alle Merkmale eines gegebenen Begriffs teilt und mindestens ein weiteres Merkmal hat. Damit definiert

sich eine Ordnung auf die Menge der Begriffe eines gegebenen Kontextes. Dies zusammen nennt man Begriffsverband zu dem Kontext.

| | pferd | männlich | weiblich | erwachsen | jung |
|---|---|---|---|---|---|
| Pferd | × | | | | |
| Hengst | × | × | | × | |
| Stute | × | | × | × | |
| Fohlen | × | | | | × |
| Stutfohlen | × | | × | | × |
| Hengstfohlen | × | × | | | × |

*Tabelle 1.1:* Beispiel für eine Kreuztabelle. Übernommen aus Janssen (2002). Ins Deutsche übertragen. Hervorgehoben sind die Kreuzungspunkte aus den Gegenständen und Merkmalen des Kontextes, der den Begriff *pferd, weiblich* ergibt.

### 1.3.2 SIMuLLDA

Die Structured Interlingua MultiLingual Lexical Database Application (SIMuLLDA) ist im Rahmen der Dissertation von Maarten Janssen aus dem Jahr 2002 entstanden. Statt von einer Sprache direkt in eine gegebene andere Sprache zu übersetzen, verwendet dieser Ansatz eine Zwischensprache (Interlingua). Die natürlichen Sprachen sind repräsentiert durch ihre Worte (Lexeme).

Grundlage der Zwischensprache sind Kreuztabellen, wie sie unter *Begriffe der Formalen Begriffsanalyse* auf Seite 5 vorgestellt wurden. Aus den daraus gebildeten Begriffsverbänden sind die Lexeme der jeweiligen Sprachen mit den entsprechenden formalen Begriffen verbunden. Immer dann, wenn einem formalen Begriff eine Bedeutung in beiden Sprachen zukommt, können die entsprechenden Lexeme verbunden werden. Auf diese Weise kann überbrückt werden, wenn es kein entsprechendes Lexem in der zu übersetzenden Sprache gibt, weil die Bezeichnungen der Merkmale in der zu übersetzenden Sprache trotzdem existieren. Der Ansatz der Arbeit von Janssen diente unserem Projekt als Ausgangspunkt mit Hilfe der Formalen Begriffsanalyse Zuweisungen innerhalb des Wörterbuchs der Augenheilkunde bewerkstelligen zu können.

### 1.3.3 Ordnungsdiagramme

Ordnungsdiagramme sind graphische Darstellungen mathematischer Halbordnungen. Geläufige Namen für Ordnungsdiagramme sind außerdem noch *Liniendiagramm* und *Hasse-Diagramm*.[2] Das nachfolgende Ordnungsdiagramm ist eine vollständige Darstellung der Kreuztabelle unter *Begriffe der Formalen Begriffsanalyse* auf Seite 5. Die Ordnung ist aufgebaut auf der Grundlage der Potenzmenge, der Menge aller Teilmengen, der Merkmalsmenge.[3] Diese Darstellung ist so zu lesen, dass jeder Knoten im Diagramm einem Begriff entspricht. Die Merkmale stehen jeweils oberhalb der Knoten, die natürlichsprachigen Entsprechungen bestimmter Begriffe stehen unterhalb mancher Knoten. Die Knoten selbst haben keine Beschriftung, sondern sie sind definiert durch die Merkmale, die entlang der Linien zwischen einem gegebenen Knoten und dem Knoten zuoberst stehen. So ist der Knoten, über dem das Merkmal *weiblich* steht, definiert durch die beiden Merkmale *{pferd, weiblich}*. Der Knoten unter dem das Wort STUTE steht ist definiert durch die Merkmale *{pferd, weiblich, erwachsen}*. Damit ist auch deutlich, dass der eine Knoten Oberbegriff des anderen Knotens ist, weil sie alle Merkmale bis auf eines teilen. Die Beschriftung unterhalb der Knoten mit den natürlichsprachigen Ausdrücken des jeweiligen Begriffs für die deutsche Sprache, verdeutlicht recht anschaulich das Konzept der Zwischensprache. In anderen Sprachen mögen an einigen Knoten keine Worte verfügbar sein, um dem Begriff Ausdruck zu verleihen, die Ordnung in Bezug auf die verwendeten Merkmale ist aber immer gleich.

### 1.3.4 Grenzen der Formalen Begriffsanalyse

Die Formale Begriffsanalyse erzeugt semantische Strukturen ausschließlich auf Grundlage der vorhandenen Merkmale sowie der Zuordnung der Merkmale zu den untersuchten Objekten. Naturgemäß ist diese Art der Inhaltserschließung an Grenzen gebunden. Zunächst ist festzuhalten, dass nur binäre Zuweisungen von Merkmalen in der Formalen Begriffsanalyse möglich sind: ein Merkmal ist einem Objekt entweder zugewiesen oder nicht. Größere Wertebereiche können somit nicht abgebildet werden, wie beispielsweise ein gegebener Messwert für den Augeninnendruck von 17 mmHg[4] als Ausprägung des Merkmals Augeninnendruck. Man könnte zwar das Merkmal 17 mmHg zuweisen, dies würde jedoch allenfalls Objekte, die mit dem Merkmal 17 mmHg versehen wurden, untereinander vergleichbar machen, da es die Formale Begriffsanalyse nicht

[2] Wir halten *Ordnungsdiagramm* für die geeignete Vorzugsbezeichnung und verwenden sie durchgehend.

[3] Zu einer gewissen Verwirrung im Umgang mit Ordnungsdiagrammen trägt anfangs bei, dass der graphisch zuunterst liegende Knoten die Menge aller Merkmale repräsentiert (also am spezifischsten ist), aber *obere Schranke* genannt wird, während der zuoberst dargestellte Knoten die leere Menge repräsentiert, aber *untere Schranke* genannt wird. Erschwerend kommt hinzu, dass das hier gezeigte Ordnungsdiagramm keine untere Schranke hat, denn der zuoberst dargestellte Knoten ist repräsentiert durch das Merkmal *pferd*, das allen Objekten des Verbands zugrundeliegt.

[4] Die Maßeinheit mmHg beschreibt den Druck in Flüssigkeiten in mm Aufstieg einer Quecksilbersäule.

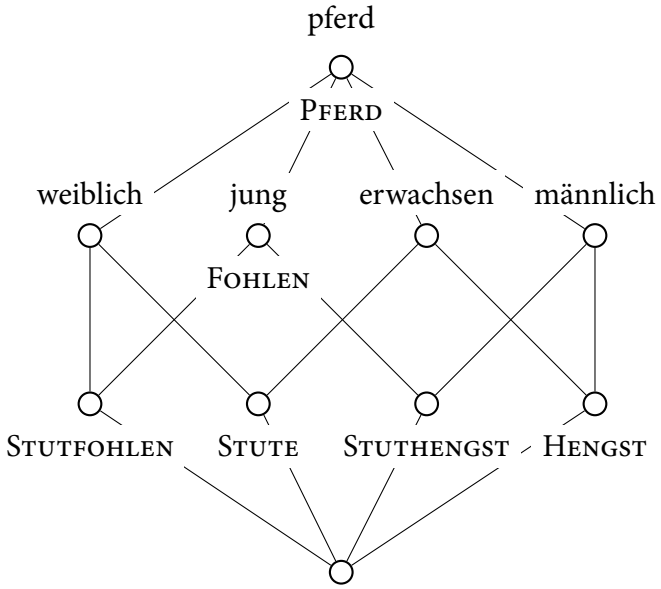


*Abbildung 1.1:* Ordnungsdiagramm der Kreuztabelle aus Janssen, 2002. Ins Deutsche übertragen.

vorsieht, Beziehungen zwischen Merkmalen zu kennzeichnen oder Merkmale als Zahlenwerte zu interpretieren. Ein Objekt mit dem Merkmal 16 mmHg wäre somit genauso wenig bedeutungsgleich zu einem Objekt mit dem Merkmal 17 mmHg wie eines mit dem Merkmal 6 mmHg, es sei denn die Bedeutungsähnlichkeit würde durch weitere Merkmale deutlich gemacht werden.

Das deutlich größere Problem besteht jedoch in der Abhängigkeit vom Merkmalsraum. Herkömmlich gebrauchte hierarchisch strukturierte Wissensordnungen, wie ein Thesaurus oder eine Klassifikation, haben oft durch ihren eingrenzenden und definierenden Charakter den Anspruch auf Vollständigkeit in ihrem Nutzungskontext. Die ICD-10 beispielsweise klassifiziert alle bekannten Krankheiten, die Internationalen Regeln für die Zoologische Nomenklatur (ICZN) das Tierreich. In der Formalen Begriffsanalyse kann dieser Anspruch für einen gegebenen Begriffsverband nur geltend gemacht werden, wenn alle relevanten Merkmale im Merkmalsraum vorhanden sind, sodass eine vollständige und gültige mathematische Ordnung entsteht. In der Praxis ist das für konkrete Probleme aus wissenschaftlichen Disziplinen – gerade dann, wenn sie nicht aus der Naturwissenschaft stammen – kaum möglich. Bei den genannten Beispielen sind alle relevanten klassifizierenden Merkmale im Vorhinein definiert in Form von Klassen. Für Kontexte mit geringer Komplexität ist der Aufbau eines hinreichenden Merkmalsraums einfach, wie beispielsweise bei der Untersuchung der Menge der natürlichen Zahlen daraufhin, ob sie gerade, ungerade und/oder eine Primzahl sind (die entsprechenden Merkmale sind *gerade*, *ungerade* und *Primzahl*). Die Domäne der Augenheilkunde ist jedoch sehr umfangreich, so dass eine Festlegung auf eine sowohl minimale als auch vollständige Merkmalsmenge, die die Augenheilkunde abdecken kann nicht so eindeutig lösbar ist, wie bei den natürlichen Zahlen.

Um einen dennoch annähernd abdeckenden Merkmalsraum generieren zu können, ist unserer Auffassung nach Domänenwissen erforderlich.

### 1.3.5 Chancen der Formalen Begriffsanalyse

Neben den geschilderten Nachteilen weist die Formale Begriffsanalyse trotzdem entscheidende Vorteile auf, in denen wir viel Potential für die Erschließung des Wörterbuchs der Augenheilkunde sehen. Der Umstand, dass lediglich geprüft werden muss, ob ein Merkmal einem untersuchten Objekt zugewiesen werden kann oder nicht, bedingt sehr einfache Datenstrukturen. Ein Synset erschließt sich daher aus seiner Menge an Merkmalen und die Ähnlichkeit jedes Synsets zu einem anderen beliebigen Synset kann leicht mit der Teilmenge gemeinsamer Merkmale bestimmt werden. Die resultierenden Begriffsverbände lassen sich in der Regel gut mit Ordnungsdiagrammen darstellen. Einzelne Knoten dieser Diagramme können sowohl Synsets, als auch Mengen von repräsentativen Merkmalen darstellen.

Gegen Ende unseres Berichts stellen wir vor, wie wir Nutzern des Wörterbuchs die Möglichkeit geben, sich einzelne Begriffsverbände visualisieren zu lassen und interaktiv zu steuern. Das zuvor angesprochene Problem der fehlenden Beziehungen von Merkmalen untereinander kann teilweise dadurch gelöst werden, dass während der Merkmalszuweisung bekannte semantische Relationen berücksichtigt werden. Beispielsweise können jedem Objekt, dem das Merkmal *Gonioskopie* zugewiesen wird, auch die Merkmale *Untersuchung* sowie *diagnostisches Verfahren* zugewiesen werden.

# 2 Bericht

## 2.1 Automatisierter Aufbau einer Wissensdatenbank

Um mit der Formalen Begriffsanalyse beginnen zu können, wurde zunächst eine Menge von Merkmalen benötigt, die den Begriffen im „Wörterbuch der Augenheilkunde“ zugewiesen werden konnten. Ein Begriffsverband als Ergebnis der Formalen Begriffsanalyse stellt die Menge an Begriffen in Bezug zu einer so vordefinierten Menge an Merkmalen. Im „Wörterbuch der Augenheilkunde“ befanden sich zu diesem Zeitpunkt 20000 Einträge repräsentiert durch 2500 Synsets. Ein gegebenes Synset ist dabei die Menge aller im Vorhinein als synonym bestimmten Wörterbucheinträge, insbesondere auch über Sprachen hinweg.

Zu Beginn stand keine Quelle zur Verfügung aus der sich geeignete Merkmale hätten entlehnen lassen. Eine derart große Menge an Konzepten lässt sich nur mit großem intellektuellen Aufwand erschließen, zumal hierzu ein hohes Maß an vertieftem Domänenwissen innerhalb des Bereichs der Augenheilkunde nötig wäre. Aus diesem Grund erschien die Automatisierung dieses Schrittes als die einzig vielversprechende Methode, Merkmale für formale Begriffsverbände aus den Synsets des Wörterbuchs zu generieren, weshalb diese auch eine der ersten Zielsetzungen für das Projekt darstellte. Gerade im Hinblick auf zukünftige Projekte, die sich mit der Anwendung Formaler Begriffsanalyse auf große Begriffsmengen befassen, werden die Erkenntnisse, die bei der Merkmalsextraktion gewonnen wurden, von uns als nützlich eingeschätzt. Wichtig ist hierbei die Wahl geeigneter Wissensquellen, aus denen die Merkmale gewonnen werden, und die Herangehensweise bei der Verwertung der Quellen, also der Art und Weise, wie Merkmale extrahiert werden. Es stellte sich zunächst die Frage, welche Eigenschaften ein Wissensbestand aufweisen sollte, damit er als Quelle für die Merkmalsextraktion in Betracht gezogen werden kann. Im Nachfolgenden werden wir auf die von uns herangezogenen Quellen und deren Erschließung eingehen.

Für einen kompletten formalen Kontext, wie ihn die Formale Begriffsanalyse beschreibt, sind eine Menge an Begriffen und eine Menge an Merkmalen aber nicht ausreichend: der Bezug der Elemente der einen Menge zu den Elementen der anderen Menge fehlt. In der Formalen Begriffsanalyse wird dieser durch eine Relation mit binärem Wertebereich, der Inzidenzrelation, angegeben, welche die Zuweisungen von Merkmalen zu Objekten beschreibt. Ein Merkmal bekommt einen Begriff entweder zugewiesen oder nicht. Diese Beziehung zwischen der Gegenstands- und

der Merkmalsmenge bildet die Grundlage der Analyse. Die Merkmalszuweisung wird im Zuge der automatisierten Merkmalsextraktion vollzogen, da die untersuchten Synsets ausschließlich auf Eigenschaften hin untersucht werden, die auf zutreffende Merkmale schließen lassen, also auf solche, die dem untersuchten Synset zugewiesen werden können. Diese Methode erschien am leichtesten umsetzbar und ist vergleichbar mit Verschlagwortung, bei der wissensrelevanten Elementen Schlagworte zugewiesen werden, die dieses Element beschreiben.

Ein formaler Kontext für eine gegebene Menge an Synsets besteht dementsprechend aus der Menge dieser Synsets, der Vereinigungsmenge der zugewiesenen Merkmale aller betrachteten Synsets, sowie der daraus resultierenden Zuweisungsrelation zwischen beiden Mengen (ist ein Merkmal aus der Menge einem Synset nicht zugewiesen, so hat das Synset für dieses Merkmal den Wert 0, sonst 1). Die Menge an Merkmalen ergibt sich nach dieser Methode also nicht aus vordefinierten Überlegungen, wonach ein formaler Kontext analysiert werden soll, sondern aus der Semantik jedes einzelnen Synsets, das in den formalen Kontext mit aufgenommen wird. Somit existiert in einem formalen Kontext kein einziges Merkmal, das keinem der betrachteten Synsets zugewiesen wurde.

### 2.1.1 Die ICD-10 als Wissensquelle

Das „Wörterbuch der Augenheilkunde“ umfasst Begriffe aus vielen Bereichen der Medizin. Wir konnten während unserer Arbeit am Wörterbuch folgende Kategorien identifizieren:

1. Anatomie (Aufbau des Auges)
2. Pathologie (Erkrankungen des Auges)
3. externe Schädigungen des Auges
4. Behandlungs- und diagnostische Verfahren
5. Instrumente und Apparate in der Augenheilkunde
6. Arzneimittel

Da vor allem Erkrankungen des Auges einen Großteil der Synsets ausmachen, schien es vielversprechend entsprechende externe Quellen nach Merkmalen zu durchsuchen. Ausgehend von dieser Überlegung wurde zunächst die deutsche Ausgabe der „Internationalen Klassifikation für Krankheiten“ (vollständig: International Statistical Classification of Diseases and Related Health Problems, nachfolgend ICD-10) in der Version des Jahres 2011 zur Merkmalsextraktion herangezogen. Die Tatsache, dass die komplette Klassifikation online in HTML-Form zugänglich

ist und Einträge der Klassifikation hierarchisch nach anatomischer Lage geordnet sind, lässt die ICD-10 als ideale Wissensquelle für unsere Zwecke erscheinen.

Die automatisierte Merkmalsextraktion setzt am Aufbau der ICD-10-Notationen an. Diese sind nach einem strikt hierarchischen Klassensystem aufgebaut. So steht beispielsweise die Klasse *H40* für ein Glaukom, während *H40.1* für ein primäres Weitwinkelglaukom steht. Die Synsets des Wörterbuchs sollten nun auf die Einträge in der ICD-10 abgebildet werden. Dies geschah durch Zeichenkettenvergleich von Synset und der Überschrift des ICD-10-Eintrags. Die Klassen der Notation wiederum wurden auf entsprechende Merkmale abgebildet, beispielsweise: *Glaukom* auf *H40*. Identifiziert wurden einzelne Einträge der Klassifikation, einschließlich der Überschrift und der Notation, mittels regulären Ausdrücken, welche einzelne Abschnitte der ICD-10 automatisch erkennen und extrahieren.

Mithilfe der aus den Notationen erkannten Klassenbezeichnungen sollten dann anschließend jedem Synset, dem eine spezifische Notation zugeteilt wurde, zusätzlich alle Merkmale zugewiesen werden, die auch dem Synset der Oberklasse zugeteilt wurden. Obwohl den Synsets häufig mehrere synonyme Bezeichnungen vergeben wurden, beispielsweise unterschiedliche Schreibweisen, konnten lediglich 43 Synsets aus dem Wörterbuch automatisiert ICD-10-Notationen zugeordnet werden und nur etwa 50 Merkmale (bzw. Notationen) erkannt und vergeben werden, von denen die meisten auch nur ein einziges Mal vergeben wurden. Hauptursache sind die Zusatzinformationen, die bei vielen Krankheiten im Titel stehen und ohne manuelles Eingreifen nur schwer zu identifizieren sind, sowie abweichende Wortstellungen bei komplexeren Begriffen. Die niedrige Zahl der Merkmalszuweisungen ergibt sich aus der hierarchischen Struktur der ICD-10, da durch die Einteilung in Klassen Überlappungen von Informationen vermieden werden. Folglich ist dieses Ergebnis als Grundlage für aussagekräftige formale Verbände in keiner Weise hinreichend und der Ansatz der automatisierten Verarbeitung der ICD-10-Klassifikation wurde nicht weiter verfolgt.

### 2.1.2 Wortmuster als Merkmalsidentifizierer

Die Durchführung einer (teilweise) automatisierten Merkmalsextraktion und -zuweisung erwies sich als sehr wirksam. Diese wurde anhand von Wortteilmustern und ihrer Bedeutung als Affix, Wortwurzel beziehungsweise Teil eines Kompositums oder Derivats erzielt, wie sie gehäuft in medizinischen Fachbegriffen vorkommen. Viele dieser Fachbegriffe beinhalten Wortteile aus dem Lateinischen oder Altgriechischen oder sind fast vollständig Lehnwörter daraus. Immer wieder

auftretende Affixe des Lateinischen und des Altgriechischen[1][2] deuten meist auf einen hinreichend eindeutig gekennzeichneten Umstand hin, so dass dieser auf entsprechende Merkmale übertragen werden kann. Die Merkmalsextraktion findet hierbei manuell statt, die Merkmalszuweisung automatisch.

Einen großen Mehrwert erhält man zudem durch Berücksichtigung der semantischen Beziehungen zwischen den zugewiesenen Merkmalen. Beispielsweise kann jedem Synset, dem das Merkmal *Entzündung* zugewiesen wurde, auch das Merkmal *Krankheit* als Oberbegriff zugewiesen werden. Neben Hyperonymen kann dieser Folgerungsschritt auch auf Meronyme angewendet werden. So kann einem Synset, dem das Merkmal *Macula* anhand des Teilmusters */macul/* zugewiesen wurde, auch das Merkmal *Retina* zugewiesen werden, da die Makula auf der Netzhaut liegt. Diesem Prinzip folgend wurden initial zunächst etwa 20 Wortteilmuster manuell aus der Menge an Begriffen im Wörterbuch identifiziert, die sehr häufig auftraten und eine eindeutige Semantik als Morphem aufweisen. Beispiele hierfür sind

- */itis/* für Entzündung,
- */graphy/* für bildgebende Verfahren und
- */retin/* für Retina oder Netzhaut.

Sofern ein Teilmuster Bestandteil eines Begriffes aus dem Wörterbuch ist wird dem zugehörigen Synset das Merkmal zugewiesen, das aus dem identifizierten Wortteilmuster abgeleitet werden kann. Da keine externen Informationsquellen für diesen Prozess benötigt werden, können die Merkmalszuweisungen direkt auf Datenbankebene per SQL-Befehl ausgeführt werden. Der folgende beispielhafte Programmcode skizziert die Zuweisung des Merkmals mit der ID 101 (Entzündung) an alle Synsets aus dem Wörterbuch, von denen eine ihrer Bezeichnungen den Wortbestandteil */itis/* enthält:

*Listing 2.1:* Merkmalszuweisung der ID 101 (Entzündung) an Synsets mit dem Wortbestandteil */itis/*

```
1 INSERT INTO Merkmalszuweisungen (SynsetID, MerkmalID)
2 SELECT SynsetID, 101
3 FROM Woerterbuch
4 WHERE Bezeichnung LIKE '%itis%'
```

Durch automatischen Abgleich der Synonyme in der Datenbank mit den erkannten Teilmustern wurden so fast 2000 Merkmalszuweisungen vorgenommen, die etwas mehr als 20% aller Synsets des Wörterbuchs abdecken.

[1] Wir stellen im Fließtext Präfixe durch vorangestellten Schrägstrich, Affixe durch finalen Schrägstrich und Infixe durch beidseitige Schrägstriche dar.

[2] Beispielhaft: die englische Endung */oplasty*, die einen rekonstruktiven Eingriff bezeichnet, stammt aus dem griechischen und bedeutet *formen*. Die griechische Endung */itis* wiederum bezeichnet meist eine entzündliche Krankheit

### 2.1.3 Einsatz einer N-Gramm-Frequenzliste

Die Wahl geeigneter Wortmuster durch manuelle Untersuchung der englischen Begriffsliste - der Menge aller englischsprachigen Einträge des Wörterbuchs - ist als zu unvollständig bezüglich der Abdeckung der 2500 Synsets einzustufen, wie die bisherigen Ergebnisse der Wortmusteranalyse zeigen. Eine Verbesserung der Abdeckung könnte zu erreichen sein, indem weitere geeignete Wortteilmuster anhand ihrer absoluten Häufigkeit identifiziert werden. Aus dieser Überlegung entstand die Idee bereits im Vorhinein eine Frequenzliste aller existierenden Zeichen-n-Gramme aus der Menge der Wörter in den Synsets zu generieren. Anhand diesen sollten geeignete Wortmuster mit semantischem Inhalt ausgewählt und dann in einem weiteren Durchlauf der Wortmusteranalyse berücksichtigt werden. Hierfür musste eine nach Häufigkeiten der vorkommenden n-Gramme sortierte Liste automatisiert erstellt werden.

N-Gramme wurden erst ab einer Länge von mindestens 4 Zeichen gebildet. Durch Versuche mit n-Grammen unterschiedlicher Länge zeigte sich, dass unterhalb dieser Grenze kaum noch brauchbare Wortteilmuster identifiziert werden konnten. Im späteren Verlauf wurden trotzdem noch die Trigramme berücksichtigt, die geeignete Wortteilmuster darstellten. Bei diesen Trigrammen handelt es sich um:

1. */dys/* für *illness*
2. */cyt/* für *cell*
3. */myo/* für *muscle*
4. */exo/* für *outside*

Anschließend wurde die Liste ausgehend vom am häufigsten vorkommenden n-Gramm bis zu einer Häufigkeit von 6 manuell abgearbeitet und jedes n-Gramm jeweils nach semantischem Inhalt sowie semantischer Eindeutigkeit überprüft. Im medizinischen Kontext muss die Bedeutung des vorliegenden Gramms nicht nur offensichtlich, sondern innerhalb dieses Kontextes auch eindeutig sein, also keine Mehrdeutigkeiten aufweisen.

Auf diese Weise ausgesuchte n-Gramme qualifizieren sich als Wortteilmuster und werden mit den zugehörigen Merkmalen, die sich aus der Semantik des Musters ergeben, gekennzeichnet. Dies muss manuell geschehen, da hierbei Domänenwissen absolut notwendig ist. Ursprung dieses Wissens sind fachbezogene Quellen und die Wikipedia-Enzyklopädie, aber insbesondere der englische Wikipedia-Artikel „List of medical roots, suffixes and prefixes“[3].

Synsets, die eine oder mehrere Synonymbezeichnungen besitzen, welche das gekennzeichnete n-Gramm als Teilkette enthalten, werden wie zuvor mit den zugewiesenen Merkmalen versehen. Zum Beispiel wird das Wortteilmuster */dystroph/* mit den Merkmalen *degeneration* und *disease* gekennzeichnet. Daraus folgt, dass Begriffe, die */dystroph/* als Wortteilmuster enthalten, diese zwei Merkmale automatisch zugewiesen bekommen.

[3] Erreichbar unter http://en.wikipedia.org/wiki/List_of_medical_roots,_suffixes_and_prefixes.

Das Ergebnis dieser erweiterten Extraktionsmethode ist ein aussagekräftiger formaler Kontext aus 119 Merkmalen, die 4467 mal zugewiesen wurden. Gut die Hälfte aller Synsets aus dem Wörterbuch wurden mithilfe der Methode mit Merkmalen versehen. Besonders ermutigend sind dabei der hohe Grad der Qualität an Merkmalen und Zuweisungen aufgrund der eindeutigen Beziehung zwischen Wortbestandteil und Semantik, sowie die hohe Vernetzung der Synsets untereinander durch Merkmale, die sehr häufig mehreren verschiedenen Synsets zugewiesen werden konnten.

## 2.2 Wikipedia als Wissensquelle

### 2.2.1 Die Kategorie-Tags in der Wikipedia

#### Vorüberlegungen

Die Beschränkung auf Erkrankungen des Auges und damit auf den betreffenden Bereich der ICD-10 erbrachte kaum Merkmalszuweisungen. Zwar sind die Mehrzahl der Einträge im „Lexikon der Augenheilkunde“ solche, die Krankheiten betreffen, aber es zeigte sich, dass die ICD-10 ungeeignet als Quelle für Merkmale ist. Dies liegt im Wesentlichen in der Art der dort aufgefundenen Information begründet.

In der ICD-10 beinhaltet die Bezeichnung einer Klasse keine Erklärung für den Begriff. Der Begriff wird nicht in den Kontext anderer vorhandener Klassen gesetzt und es erfolgt keine Umschreibung in termini seiner Ober- oder Unterklassen oder verwandter Begriffe. Eine einfache Verschlagwortung mittels eines kontrollierten Vokabulars erzeugte vermutlich mehr Querverweise, die für den Aufbau einer Merkmalsmenge nützlich sein könnten, als die aus der ICD-10 gewonnenen Zuweisungen. Der Grund dafür ist, dass der Zweck einer Klassifikation genau ist, möglichst wenig Kongruenz im Begriffsumfang mit benachbarten, vor allem aber Ober- und Unterbegriffen zu beinhalten.

Aufgrund der großen Menge an bereitgestellten Informationen haben wir uns dazu entschieden, die Online-Enzyklopädie Wikipedia als Wissensquelle für das Wörterbuch der Augenheilkunde zu berücksichtigen. Auch wenn anfänglich sowohl die englische, als auch die deutsche Version der Wikipedia für die Informationsextraktion herangezogen wurden, entschied man sich schließlich für die englische Wikipedia als Wissensquelle, um einerseits den Merkmalsraum einsprachig zu halten und andererseit aus dem Grund, dass die englische Wikipedia deutlich mehr Informationen aus dem Bereich der Augenheilkunde enthält (siehe Anzahl erkannter Artikel am Ende des Abschnitts), und weil die englische Sprache weniger Flexionen verwendet als die deutsche, was den linguistischen Abgleich von Informationen der Wikipedia mit den Einträgen aus dem Wörterbuch vereinfacht.

In der Wikipedia wird durch Zuweisung eines Artikels einer oder mehreren Kategorien eine Klassifikation erzeugt. In erster Linie dient dies dazu, thematische Bereiche zu erstellen und weniger eine Begriffsordnung zu schaffen. Daher gibt es, neben einer Begriffsordnung eher zurechenbaren Kategorien wie „Medizin“, „Medizinisches Fachgebiet“ oder „Augenheilkunde“ auch Sammelkategorien wie „Liste Nationalparks“. Die Wikipedia hat damit Eigenschaften einer hierarchischen Klassifikation und solche einer Facettenklassifikation. Die Wahl einer geeigneten Kategoriebezeichnung für einen Artikel ist dem jeweiligen Autor freigestellt, er kann auch eine neue Kategoriebezeichnung wählen.

In der Praxis gibt es aber Empfehlungen für die Eigenschaften und die korrekte Wahl einer geeigneten Kategoriebezeichnung und über die Zeit bildet sich dadurch eine brauchbare (das heißt: nicht zu diverse) Begriffsordnung heraus, weil darauf geachtet wird, dass es beispielsweise nur eine Kategorie „Liste Nationalparks“ gibt, und nicht zusätzlich noch „Liste der Nationalparks“, „Liste aller Nationalparks“ und ähnlichen Abweichungen. Eines der Probleme dabei ist, dass nicht sichergestellt ist, dass eine Seite nicht einer Kategorie und einer Oberkategorie dieser Kategorie gleichzeitig eingeordnet wird. Dadurch wird es für die automatische Erschließung eines Kategoriebaums allerdings erforderlich, auf diese Weise doppelt gesammelte Artikel zu eliminieren. Insbesondere ist darauf zu achten, dass wenn den Kategorien-Tags automatisiert gefolgt wird um weitere Artikel zu erschließen, bereits besuchte Kategorien zu ignorieren, um Schleifenbildung zu vermeiden. Dies berücksichtigend schien es aber vielversprechend den Bereich „Augenheilkunde“ sowohl der englischen wie der deutschen Wikipedia bezüglich der vergebenen Kategorien zu untersuchen.

Zu diesem Zweck wurde von beiden eine Volltextkopie der Wikipedia Abstracts als XML-annotierte Datei heruntergeladen und innerhalb der *title*-Tags nach Übereinstimmungen mit Begriffen oder Phrasen aus dem „Wörterbuch der Augenheilkunde“ gesucht. Die Abstracts sind automatisch erstellte Kurzfassungen der entsprechenden Artikel und bestehen, wie sich herausstellte, im Wesentlichen aus dem ersten Absatz, oder, falls eine bestimmte, nicht exakt erkennbare Untergrenze an Wörtern nicht überschritten wurde, aus den ersten Absätzen. Weiterhin sind die Kategorien vermerkt, denen der Artikel angehört. Die Abstracts wurden aus dem Quelltext gewonnen und der Text ist mit dem Markup der Wikipedia ausgezeichnet.

Da es unterschiedliche Konventionen bezüglich der Formatierung im Quelltext gibt, sind einige Abstracts unbrauchbar gewesen, da sie im Wesentlichen aus dem Markup einer dem Artikel vorangestellten Infobox (ein Beispiel einer solchen Infobox kann man auf der Seite jedes chemischen Elements betrachten) bestanden und keinerlei Fließtext mehr enthielten. Aus der deutschen Wikipedia konnten 433 Artikel für die weitere Verwendung extrahiert werden und aus der englischen Wikipedia 731 Artikel, die Mitglieder der Kategorie selbst oder einer Unterkategorie von „Augenheilkunde“ oder „Ophthalmology“ waren.

### Vorgehensweise

Das Vorgehen hierzu umfasste zahlreiche nur manuell durchzuführende Schritte, um die Ergebnismenge möglichst frei von Artikeln oder Kategorien zu halten, die nicht zur eigentlichen Fragestellung zählten. Die Vorgehensweise war wie folgt:

1. Download der Abstracts. Für die deutsche Wikipedia 1.19 GB und für die englische Wikipedia 3.15 GB Dateigröße.

2. Ermitteln der Artikelnamen in den Unterkategorien der Kategorien „Augenheilkunde“ bzw. „Ophthalmology“. Dieser Schritt wurde noch manuell unter Verwendung des Kategorie-Browsers auf der Wikipedia-Hauptseite durchgeführt.

3. Extraktion der Abstracts durch Suche nach den Artikelnamen. Da es sich um ein XML-Dokument handelt, wurde versucht mittels XSLT (Extensible Stylesheet Language Transformation, eine Sprache unter anderem zur Extraktion von Teilbäumen eines XML-Dokuments) und regulären Ausdrücken, die Abstracts zu gewinnen. Verwendet wurde dazu das Kommandozeilenprogramm *xsltproc*. Der Aufwand damit eine fehlerfreie Menge an Abstracts zu extrahieren zeigte sich aber als zu groß, insbesondere aufgrund der spärlichen Dokumentation bezüglich der Leistungsfähigkeit der regulären Ausdrücke[4], aber auch weil die Abstracts keine vollkommen einheitliche Struktur hatten, und die XSL-Templates immer wieder angepasst werden mussten um die Treffermenge zu erhöhen. Daher wurde die Extraktion mit den üblichen dafür geeigneten Programmen der Unix-Shell (*grep, sed, awk*) und *Perl* bewerkstelligt. Für wiederholte Aufgaben dieser Art ist eine Einarbeitung in XSLT aber zu empfehlen, weil die Verarbeitung sehr großer Dateien, selbst im Vergleich zu den eben genannten Programmen, noch einmal erheblich schneller ist und auch dem eigentlichen Zweck der Verwendung von XML zur Datenauszeichnung entspricht.

4. In den Abstracts befinden sich die Kategorie-Tags, die wiederum dem Artikelnamen zugewiesen werden. Allerdings fanden sich auch mehrere hundert Kategorien, die lediglich der Verwaltung der Wikipedia dienen, und die vorher nicht aufgefallen waren, weil sie bei der Anzeige im Browser unterdrückt werden. Ein allgemeines Beispiel dafür ist die Kategorie „Articles needing attention“, die selbst wiederum mehrere hundert Unterkategorien enthält (unter anderem zu jener Zeit sich selbst: „Medical Articles needing attention“). Außerdem gibt es Sammelkategorien der Art „Accuracy disputes from August 2009“. Je nachdem wie lange der Dissens bezüglich eines Themas schwelt, sind genau so viele Kategorien dieses Typs einem

[4] Die technische Dokumentation des W3C überlässt Teile davon der jeweiligen Implementation und war damit auch keine Hilfe.

Artikel zugewiesen, wie Monate vergangen sind. Obwohl viele dieser Kategorien, einmal identifiziert, durch reguläre Ausdrücke gut zu filtern waren, war es doch ein erheblicher manueller Aufwand, der aber im Wiederholungsfall automatisierbar ist.

#### Ergebnisse

Die so ermittelten Kategorien wurden als Schlagwörter der dazugehörigen Artikel betrachtet und versucht, mit den Begriffen aus dem „Wörterbuch der Augenheilkunde“ in Verbindung zu bringen. Die resultierenden Merkmalszuweisungen blieben aber hinter den Erwartungen zurück. Viele Kategorien gehören nur der Ausgangskategorie („Augenheilkunde“, „ophthalmology“) an.

Aufgrund der Mischung von hierarchischer Ordnung und Facettenordnung ist mindestens die Menge der Kategoriebegriffe zu klein, um eine erkennbare Vererbung der Kategoriebegriffe auszunutzen; wahrscheinlich ist aber, dass selbst bei einer wesentlich größeren Zahl an Kategorien in der selben Domäne keine wirkliche Verbesserung der Situation einträte, weil dann immer noch vererbte Eigenschaften mit assoziierten Eigenschaften der Kategorie vermischt wären.

Der wesentliche Grund für die geringen Zuweisungen ist daher die geringe Menge an Überlappungen der Kategorien gewesen. Im Grunde handelt es sich hierbei um das selbe Problem, wie bei dem Versuch mit der ICD-10 und scheiterte aus dem selben Grund: der Zweck der Klassifikation ist ja die Elimination der Überlappung der Begriffsumfänge und die Einordnung in hierarchische Strukturen unter Ausnutzung der Vererbung von Merkmalen. Der Facettenaspekt in der Klassifikation der Wikipedia-Kategorien brachte hier also kaum Verbesserung.

### 2.2.2 Die Volltext-Abstracts aus der Wikipedia

#### Vorüberlegungen

Zu Beginn jedes Lexikonartikels steht eine einfache Definition und Einordnung des Begriffs, der Gegenstand des Artikels ist. In der Wikipedia gibt es sehr umfangreiche und präzise Anleitungen, wie der Aufbau eines Artikels zu sein hat. Insbesondere für Artikel, die schon etwas älter sind und die viele Bearbeiter, insbesondere Domänenspezialisten, hatten, ist die Wahrung der Formvorgaben mit hoher Sicherheit als gegeben zu betrachten. Gerade der erste Satz folgt damit so gut wie immer dem Schema, welches exemplarisch am nachfolgenden Exzerpt aus der Wikipedia veranschaulicht werden soll:

> Die **Augenheilkunde** (Augenmedizin, fachsprachl.: **Ophthalmologie, Ophthalmiatrie**; von gr. οφθαλμος, *Auge*) ist die Lehre von den Erkrankungen und Funktionsstörungen des Sehorgans, deren Anhangsorgane, sowie des Sehsinnes und deren medizinischer Behandlung.

Die Überlegungen zur Erzeugung von Merkmalszuweisungen wurden daher darauf konzentriert, aus den ersten beiden Sätzen des Abstracts eine Zerlegung in Wort-n-Gramme der Längen 2 und 3 zu erstellen, um mit diesen mittels Zeichenkettenvergleich weitere Titel von Wikipedia-Artikeln aus der Volltextkopie zu identifizieren. Auf diese Weise identifizierte Begriffe, zu denen ein Wikipedia-Artikel existiert, sollten die Merkmale bilden, die dem Synset, das dem Wikipedia-Artikel entspricht, zugewiesen werden. Wenn beispielsweise die Wortfolge *Diabetes mellitus* im Abstract des Artikels „Diabetische Retinopathie“ gefunden wurde, dann wird das Merkmal *Diabetes mellitus* dem Synset „diabetische Retinopathie“ zugeordnet.

Der Grund aus dem die ersten beiden Sätze und nicht nur der erste Satz herangezogen wurden ist, dass bei manchen Artikeln der erste Satz nur aus einer etymologischen Einordnung besteht (ähnlich dem Inhalt der Klammer im Beispiel oben) ohne weitere Erklärung des Begriffs. Wortfolgen, zu denen erfolgreich ein Wikipedia-Eintrag zugewiesen werden konnten, werden als sinnvolles Merkmal für die Zwecke der Formalen Begriffsanalyse betrachtet, da sie anscheinend als bedeutungstragend genug erachtet werden, um in der Wikipedia-Enzyklopädie aufgenommen und erklärt zu werden. Anschließend wurden die Worte der Abstracts vorher mit Stemming auf ihren Wortstamm zurückgeführt (im Rahmen der Möglichkeiten des verwendeten Stemmers), um eine bessere Übereinstimmung mit vorhandenen Titeln von Wikipedia-Artikeln zu ermöglichen.

Hierzu wurde die ANSI-C-Implementation des Porter-Stemmers verwendet, jeweils in Standardeinstellung für die englische Sprache und mit angepassten Mustern für die deutsche Sprache (die angepassten Muster sind Teil des Programms). Die resultierende Liste von n-Grammen je Artikel wurde zusammen mit dem Artikelnamen in einer Pseudo-XML-Notation abgespeichert: die n-Gramme wurden in Tags namens *bigram* und *trigram* gespeichert ohne sich weiter mit einer geeigneten Document Type Definition aufzuhalten. Die Ketten sollten aus der Datei nur gut extrahierbar sein und die Datei trotzdem für Menschen lesbar.

### Vorgehensweise

1. Unix Shell-Skripte (vor allem mit *sed, awk, Perl*) zur Extraktion der ersten beiden Sätze. Es musste sowohl das Wiki-Markup, als auch Schriftzeichen anderer Sprachen (insbesondere Altgriechisch, Arabisch und Indisch) herausgefiltert werden, zum Teil mit manuellem Eingriff und Korrektur[5].

2. Shell-Skript für den Einsatz des Porter-Stemmers und anschließendes Speichern. Dieser Schritt

[5] Auch hier gilt, was an anderer Stelle schon bemerkt wurde: manche der manuell ausgeführten Arbeiten sind prinzipiell automatisierbar, allerdings erschien zum relevanten Zeitpunkt der Aufwand dafür, gegeben dass es sich um eine einmalige Aufgabe handelte, als nicht geboten. In jedem Fall wurde aber darauf geachtet, die Schritte so zu wählen, dass sie automatisierbar sein können.

geht vollautomatisch und bedarf keiner weiteren Überprüfung oder Korrektur.

3. PHP-Skript für den Import der XML-Daten in die Datenbank, sowie Abgleich der Wort-Gramme mit den Titeln der Artikel und Speichern der Ergebnisse.

4. Anlegen und Zuweisung der Merkmale, sowie Matching der Artikeltitel und der Synsetbezeichnungen via SQL.

### Ergebnisse

Das Ergebnis zeigte sich bezüglich der Ausbeute an Merkmalen und der Anzahl an Zuweisungen je Begriff als unbefriedigend. Zwar wurden insgesamt ca. 5000 Merkmale produziert, jedoch befindet sich darunter jede Menge Ballast in Form von irrelevanten oder aus dem Kontext heraus falsch extrahierten Daten (beispielsweise *is a* als Merkmal, welches als Wikipedia-Eintrag vorhanden ist). Zudem ist bei insgesamt ca. 7000 Merkmalszuweisungen jedes Merkmal im Schnitt 1,4 mal zugewiesen worden, was sehr wenig erscheint. Es gibt nur ca. 500 Merkmale, die öfters als dreimal zugewiesen wurden. Auch das Matching der Artikelnamen auf die Synsets bleibt mit einem Wert von 266 erkannten Synsets bei 626 Lexikonartikeln hinter den Erwartungen zurück.

Daher wurde ein zweiter Durchlauf der selben Daten allerdings nach vorheriger Bereinigung um Stoppworte durchgeführt. Die Quelle der verwendeten Stoppwortliste ist im Anhang aufgeführt. Auch hier blieb das Ergebnis hinter den Erwartungen zurück. Die Merkmalsanzahl blieb nahezu unverändert, Zuweisungen haben sich um gut 1000 reduziert. Dank großzügigerem Matching unter Nichtbeachtung von Groß- und Kleinschreibung konnten ca. 30 Synsets mehr auf Artikel abgebildet werden, aber auch das ändert die Einschätzung des Endresultates nicht. Positiv herauszustellen ist jedoch die Erzeugung eines bereinigteren Begriffsverbandes, da Merkmale wie *is a* nicht mehr zustande kommen aufgrund der Stoppwortfilterung.

## 2.3 Inhalte der Wikipedia via API und Erkennung der Wortarten

### Vorüberlegungen

Die Gründe aus denen die Volltextkopie der Wikipedia-Abstracts verwendet worden waren sind: 1. es musste kein *Crawler* geschrieben werden, der immer wieder neu auf die Seiten angewandt wurde 2. es war anfangs auch nicht zu überblicken, ob und welche Kategorien noch hinzugenommen werden sollten, und 3. die Hoffnung, dass der reine Text mit Wiki-Markup, statt dem daraus generierten (X)HTML der dann extrahierten Seiten, leichter zu erschließen sein würde.

Insbesondere das Wiki-Markup stellte sich aber als sehr schwierig vom restlichen Text zu trennen dar und erforderte übermäßig viel manuelles Eingreifen. Im Frühjahr 2012 wurde das API für Anfragen der Wikipedia stark überarbeitet und in der Funktionalität ausgebaut. Insbesondere die Ausgabe der Inhalte als reiner Text, frei von Wiki-Markup, stellte eine erhebliche Verbesserung der Möglichkeiten der maschinellen Weiterverarbeitung dar.

Das API bietet seitdem sogar die Möglichkeit eine beliebige Anzahl von Sätzen, gezählt von Beginn des Artikels an, auszugeben. Allerdings scheitert das gelegentlich daran, dass das Kriterium für das Ende eines Satzes allein das Zeichen für den Punkt ist, so dass Sätze bereits nach „Dr.“, „Prof.“ und „et al.“ enden. Für diese Fälle musste der restliche erste Satz manuell geholt werden. Da die Ansätze lediglich die Kategorien und später die gestemmten, stoppwortbereinigten und zu Wort-n-Grammen reduzierten ersten Sätze zusammen mit den Kategorien zu verwenden keinen Erfolg brachten, sollte untersucht werden, ob möglicherweise schon die „merkmalsenthaltenden“ Teile des ersten Satzes genügen, um bessere Merkmale zu erhalten. Dazu sollten Nomen und Adjektive des ersten Satzes gesammelt und als Merkmale zugewiesen werden. Die Nomen sollten in erster Linie dazu dienen, Überschneidungen mit anderen Begriffen zu begünstigen. Von den Adjektiven wurde sich erhofft, auf einer kleinen Menge beschränkt zu bleiben, die nur

- Lagebezeichnungen ähnlich jenen in der Anatomie, wie zum Beispiel *vorne, hinter, über, in, teilt, beinhaltet* und
- Bezeichnungen aus dem näheren Umfeld von Krankheiten, wie zum Beispiel *fiebrig, geschwollen, eiternd*

und ähnlichem umfasst. Die Erwartung war, in größerer Menge die Überschneidungen in der Merkmalsmenge zu bekommen, die erforderlich sind, um überhaupt mittels Formaler Begriffsanalyse analysieren zu können.

Wie bereits erwähnt, ist der erste Satz eines enzyklopädischen Artikels im Grunde immer identisch aufgebaut. Es findet sich darin der Begriff, der Gegenstand des Artikels ist, selbst. Weiterhin adjektivische Elemente, die den Begriff einordnen in den Rahmen der durch die eine hierarchische Relation implizierende Wendungen *ist ein*, *ist der/die/das* und ähnliche gesteckt wird. Ein Beispiel dafür ist der erste Satz des Artikels „Hornhaut“:

> Die **Hornhaut** (lateinisch **Cornea**, eingedeutscht auch **Kornea**, griechisch *keras* = Horn, *keratoeides chiton* = Hornhaut) ist der glasklare, von Tränenflüssigkeit benetzte, gewölbte vordere Teil der äußeren Augenhaut.

Da es mit einem Part-of-Speech-Tagger vergleichsweise einfach ist, die entsprechenden Teile zu extrahieren, sollte der Versuch unternommen werden dem Merkmalscharakter möglichst

nahe zu kommen, indem einfach die Wörter, die unmittelbar verwendet werden um den in Frage stehenden Begriff zu beschreiben, als Merkmale zu nutzen. Die Wahrscheinlichkeit, einerseits möglichst treffende Begriffe zu finden und andererseits nicht zu viele, nicht in unmittelbaren Zusammenhang stehende Wörter zuzuweisen, erschien durch die Beschränkung auf den ersten Satz eines enzyklopädischen Artikels als maximal.

### Vorgehensweise

Anders als bei der Extraktion aus der Volltextkopie wurden die Seiten von einem Shell-Skript aus mit *wget* geholt und mit den geeigneten Teilen der zuvor verwendeten Skripte sofort weiterverarbeitet. Für das Part-of-Speech-Tagging wurde der *TreeTagger* von Helmut Schmid (Schmid, 1994) verwendet. Er enthält sogenannte „parameter files" für eine Vielzahl von Sprachen. Vor allem aber ist das Programm ohne große Konfiguration recht schnell und verarbeitet Standarddatenströme der Unix-Shell, so dass es sich gut in Shell-Skripten verwenden lässt. Die genaue Vorgehensweise war:

1. Ausgehend von der Kategorie „Eye" rekursiv alle ersten Sätze aller Artikel in allen Unterkategorien holen. Bei Verwendung des API ist die Standardeinstellung, dass die oben erwähnten Verwaltungskategorien nicht beachtet werden. Man erhält also sofort eine wesentliche fehlerfreiere Menge von Artikeln. Auf diese Weise wurden 118 Kategorien und 2415 von 2841 möglichen Artikeln geholt. Die Abweichung ergibt sich zu etwa zwei Dritteln aus Verbindungsabbrüchen nach fünf Versuchen. Das Programm *wget* hat sehr elaborierte Möglichkeiten, um sein Retrievalverhalten wie das einer natürlichen Person, die mittels Web-Browser die Seiten besucht, aussehen zu lassen. Das ist notwendig, weil auch die Wikipedia-Server zu hochfrequente Verbindungsanfragen drosseln (und bei drastischem Überschreiten IP-Adressen sperren), basierend auf der IP-Adresse und der absoluten Frequenz der Anfragen. Um die Server zu schonen und die Zugriffe nicht so auffällig zu machen, wurden die Zugriffe unter anderem auf einen Zufallswert zwischen einer und 60 Sekunden eingestellt, die Download-Rate auf 5 Kb/s gedrosselt und die Anzahl der Verbindungsversuche je Artikel auf fünf Versuche beschränkt.

   Etwa 1300 der 2415 Artikel sind Namen natürlicher Personen. Der Grund dafür ist, dass ein ganzer Bereich unterhalb der Kategorie „Eye" berühmte Augenärzte und Wissenschaftler der Augenheilkunde beinhaltet. Ein anderer Bereich hingegen gehört zu Unterkategorien der Kategorie „blindness"[6] und dort tauchen sämtliche Personenartikel in der Wikipedia auf, deren Gegenstand ein blinder Mensch ist. Die Artikel dieser Kategorie wiederum umfassen Kategorien wie „blind academics", „fictional blind characters" oder „sportspeople with a vision

[6] In dieser Kategorie finden sich tatsächlich auch Leute, die nicht blind sind. Das sind zum Beispiel Personen, die sich um die Belange von Blinden verdient gemacht haben.

impairment“. Insbesondere der Teil der Personen der Wissenschaftler und Ärzte umfasst, ist vermutlich nicht unmittelbar als Ballast anzusehen. Formal sind diese Artikel aber alle Mitglieder der Oberkategorie „blindness“ via der Kategorie „blind people“. Dies ist ein weiterer Hinweis auf die Probleme der Mischung aus facettierter und streng hierarchischer Klassifikation in der Wikipedia.

2. Nach kleineren Filterschritten durch die üblichen Programme (siehe oben), musste bei 58 Artikeln manuell der erste Satz hinzugefügt werden, weil der erste Punkt im Satz der Punkt nach einer Abkürzung war. Später ist bei genauerer Betrachtung aufgefallen, dass es Artikel gab, die trotzdem vollständig waren, obwohl vor dem Satzende ein Punkt war. Das deutet darauf hin, dass diese Fehlerrate auch absolut höher hätte sein können und die Vorgehensweise vermutlich nicht ohne manuelle Kontrolle auskommt. Allerdings waren die Artikel relativ leicht zu finden, indem die durchschnittliche Satzlänge aller Artikel berechnet wurde und die Artikel ausgegeben wurden, deren Satzlänge $\geq 2$ Standardabweichungen vom Mittelwert der Satzlänge aller Sätze abwich. Diese Vorgehensweise ist zwar auch nicht vollständig reliabel, aber gut als Schritt bei der Nachbearbeitung einzubauen. Die Ergebnisse wurden an TreeTagger übergeben und die Ausgabe von TreeTagger gefiltert, so dass nur die Stammformen der Nomen und Adjektive gespeichert wurden[7].

3. Nach weiteren Filterschritten werden die Ergebnisse zusammen mit dem Artikelnamen und den Kategorien, die in einem separaten Schritt geholt wurden, gespeichert. Auf das finale Speichern der Part-of-Speech-Information wurde verzichtet.

Das Ergebnis wurde wie folgt gespeichert: die Zahl am Anfang ist eine durchlaufende Nummer je Begriff, die mit dem Sternsymbol hervorgehobenen Merkmale sind Kategorien. Jeweils vorangestellt ist der ursprüngliche Satz. Das erste Beispiel ist ein eher erfolgreiches Resultat des Ansatzes, das zweite Beispiel ist ein typischer Fall für die weniger brauchbaren Ergebnisse.

> **Keratoprosthesis** is a surgical procedure where a severely damaged or diseased cornea is replaced with an artificial cornea.

*Listing 2.2:* Beispiel für eine Ergebnisliste nach Filterung mit Wortartenerkennung. Das in Sternen eingefasste Wort ist die Kategorie.

```
69,Keratoprosthesis,surgical
69,Keratoprosthesis,procedure
69,Keratoprosthesis,damaged
```

[7] TreeTagger verwendet die Nomenklatur des Penn Treebank Project. Gespeichert wurden alle Vorkommen von JJ*, also JJ - Adjektiv, JJR - Adjektiv komparativ und JJS - Adjektiv Superlativ und NN*, also NN - Nomen Singular, NNS - Nomen Plural, NNP und NNPS für Eigennamen respektive.

```
69,Keratoprosthesis,diseased
69,Keratoprosthesis,artificial
69,Keratoprosthesis,*Ophthalmology*
```

Die zusätzliche Kategorie „Ophthalmology“ bringt hier keine besonderen Zugewinne, weil sie eine der neun unmittelbaren Unterkategorien von „eye“ ist und damit allein zu allgemein ist.

> A **scleral lens** is a large lens that rests on the sclera and creates a tear-filled vault over the cornea.

*Listing 2.3:* Beispiel für eine Ergebnisliste nach Filterung mit Wortartenerkennung. Die in Sternen eingefassten Wörter sind Kategorien.

```
185,Scleral lens,large
185,Scleral lens,vault
185,Scleral lens,*contact lenses*
185,Scleral lens,*corrective lenses*
```

Dadurch, dass bei der Verarbeitung in Schritt 3 die Bestandteile des Artikelnamens *scleral* und *lens* herausgefiltert werden, bleibt nicht mehr viel Nützliches übrig. Zu überlegen ist, ob man bei einer Unterschreitung einer Anzahl Ergebniszeilen, den oder die Begriffe aus denen sich der Artikel zusammensetzt, hinzunimmt. Allerdings geben die beiden Kategorien hinreichend Information um „scleral lens“ zumindest einordnen zu können.

### Ergebnisse

Insgesamt wurden 1379 Wikipedia-Abstracts analysiert, aus denen 2439 Merkmale extrahiert werden konnten, die in 9797 Fällen entsprechenden Artikeln zugeordnet wurden. Nach Matching der Artikelnamen auf Synsets aus dem Wörterbuch verbleiben (nur) 335 gematchte Synsets, auf die insgesamt 931 Merkmale 2582 mal zugewiesen wurden. Im Gegensatz zu den ersten Durchläufen der Wikipedia-Extraktion ist die Ausbeute an Merkmalen damit deutlich geringer, dafür konnten aber mehr Synsets identifiziert und die Wiederverwendungsrate der Merkmale in Relation zu ihren Zuweisungen deutlich erhöht werden (ungefähr um den Faktor 2). Zudem ist der resultierende Begriffsverband deutlich kohärenter als zuvor. Damit kann festgehalten werden, dass diese Methode allein keinen hinreichend aussagekräftigen Begriffsverband für das Wörterbuch generiert hat, jedoch als sinnvolle Ergänzung zu den bereits angewandten Methoden gesehen werden kann.

## 2.4 Visualisierung des Begriffsverbands

Der Aufbau eines formalen Begriffsverbandes für das Wörterbuch hat für die Nutzer nur dann einen Mehrwert, wenn eine geeignete Schnittstelle zur Betrachtung und Steuerung des Verbands

vorliegt. Dazu gehört eine übersichtliche Visualisierung formaler Kontexte, die den Nutzer nicht überfordert und einen guten Überblick über das semantische Umfeld eines gewählten Synsets liefert. Hierzu bietet sich die Verwendung von Ordnungsdiagrammen an.

Unser Script zur Visualisierung eines formalen Kontextes erzeugt mithilfe des Programms *GraphViz* und der Skriptsprache PHP ein Ordnungsdiagramm ausgehend von einem ausgewählten Synset. Der große Vorteil beim Rendern eines Graphen mithilfe von GraphViz ist das Wegfallen eines erheblichen Aufwands beim Platzieren der Knoten und Kanten, da GraphViz dies beim Rendern automatisch übernimmt, ohne jegliche Platzierungsinformationen zu fordern. Die Menge der dem gewählten Synset zugewiesenen Merkmale bildet die obere Schranke. Die zugehörige Potenzmenge dieser Menge bildet, wie im Kapitel Ordnungsdiagramme beschrieben, alle Knoten des erzeugten Diagramms, einschließlich der unteren Schranke, die der leeren Menge entspricht. Ein Ordnungsdiagramm wird dadurch lediglich aus einer Teilmenge aller Merkmale des Datenbestandes erzeugt, weil ein aus der kompletten Menge an Merkmalen generiertes Ordnungsdiagramm schlicht viel zu groß und damit zu unübersichtlich wäre, als dass ein Nutzer damit interagieren könnte. Um diese Platzproblematik zu lösen und die Ordnungsdiagramme so klein und übersichtlich wie möglich zu halten, werden zusätzlich alle Teilmengen der Potenzmenge des gewählten Synsets, die dieselbe Menge an Synsets beschreiben, zu einem Knoten zusammengefasst. Ebenso wurde in einem späteren Entwicklungsstadium die Darstellung der unteren Schranke abgeschafft, da sie für jedes erzeugte Diagramm stets gleich ist und keinen Mehrwert zur Exploration des Kontextes bietet.

Die erzeugten Darstellungen sind interaktiv gestaltet, indem Knoten des Diagramms klickbar sind. Bei Klick auf einen Knoten öffnet sich ein Menü, das sowohl die Merkmale, als auch die Vorzugsbezeichnungen der zugehörigen Synsets anzeigt. Die Synsets sind ebenso klickbar, sodass sich nach einem Klick der entsprechende formale Kontext für das geklickte Synset neu aufbaut, unter Berücksichtigung aller zugewiesenen Merkmale des Synsets. Auf diese Weise ist es möglich, alle Synsets des Wörterbuchs zu erforschen und benachbarte Bedeutungen zu entdecken, vorausgesetzt sie wurden durch Merkmale erschlossen.

# 3 Schlußteil

## 3.1 Diskussion der Ergebnisse

Bereits bestehende Klassifikationen, wie zum Beispiel die ICD-10, aber auch die kurzzeitig erwogene und wieder verworfene Einbeziehung des Operationen- und Prozedurenschlüssels (OPS)[1], zeigten sich für die Merkmalsextraktion als wenig bis gar nicht geeignet. Die Klassen der ICD-10 ließen sich praktisch überhaupt nicht mit Begriffen aus dem „Wörterbuch der Augenheilkunde“ in Verbindung bringen. Erfolgreicher sind die Ergebnisse aus der Extraktion von Teilen des ersten Satzes aus Artikeln der Wikipedia in der Kategorie „Augenheilkunde“ oder Unterkategorien davon, da diese Methode deutlich mehr Wissen und Begrifflichkeiten hergibt, sowie die Extraktion von Merkmalen und Zuweisungen anhand von Wortteilmustern und n-Gramm-Frequenzstatistiken, weil diese Methode leicht und direkt zu überblicken und zu automatisieren ist.

## 3.2 Diskussion der Probleme

Im Idealfall ist die Menge der verwendeten Merkmale genau so groß, wie sie sein muss, um alle formalen Begriffe bilden zu können und kein weiteres Merkmal zusätzlich. Insbesondere für allgemeine Übersetzungsaufgaben scheint es kaum vorstellbar eine Merkmalsmenge auf die hier vorgestellte Art zu verkleinern, ohne manuelle Korrekturen vorzunehmen. Bei der Beschränkung auf eine Domäne, so unsere initiale Annahme, müsste aber die automatisiert gewonnene Merkmalsmenge klein zu halten sein und dennoch ausreichend Zuweisungen erlauben. Aus unserer Sicht gab es vor allem zwei Problemstellen.

1. Die Überdeckung der Wikipedia-Artikel mit den Einträgen des Wörterbuchs der Augenheilkunde. Je mehr Artikel aus der Wikipedia sich zu Einträgen aus dem Wörterbuch finden lassen, desto mehr unmittelbar passende Merkmale lassen sich aus den Artikeln extrahieren.

2. Die Anzahl der unmittelbaren Übereinstimmungen in den Begriffen zwischen beiden betrug

[1] Der OPS ist eine Modifikation der Internationalen Klassifikation der Prozeduren in der Medizin. Es handelt sich dabei primär um ein Werkzeug des Controllings von Krankenhäusern und ist aufgrund seiner seriellen Struktur ohne Nutzen für uns gewesen.

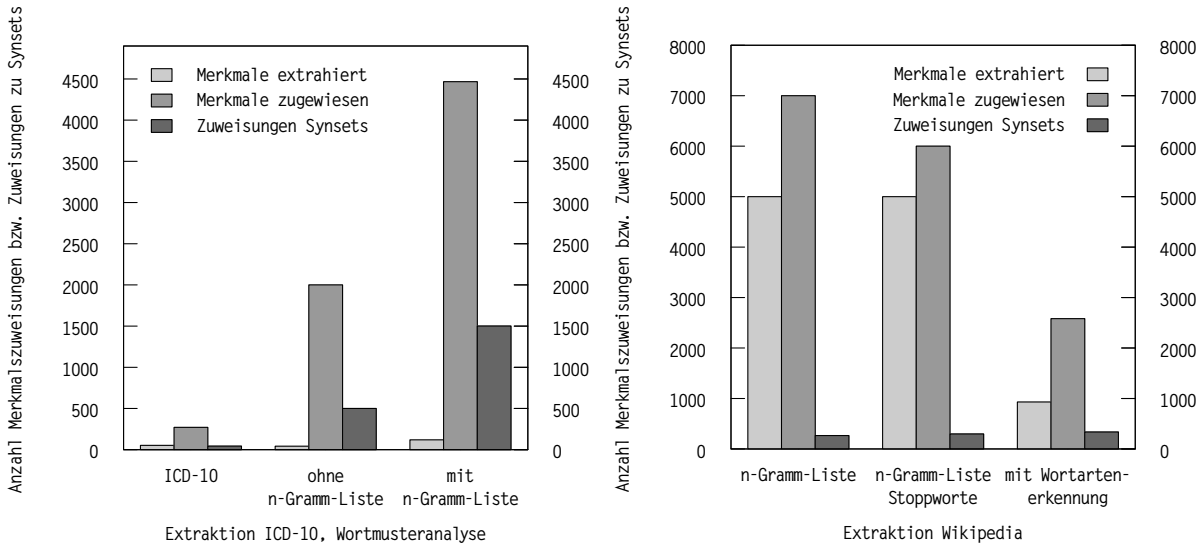


*Abbildung 3.1:* Anzahl der Merkmalszuweisungen zu Synsets im „Wörterbuch der Augenheilkunde“ in Abhängigkeit unterschiedlicher Vorgehensweisen. *Abbildung links*: Extraktion aus der ICD-10, sowie Wortmusteranalyse einmal mit und einmal ohne n-Gramm-Liste und *Abbildung rechts*: Extraktion aus den Daten der Wikipedia einmal mit n-Gramm-Liste und einmal zusätzlich mit Stoppwort-Liste und schließlich nach Wortartenerkennung.

aber nur 335 Artikel-Eintrag-Kombinationen. Weil die Anzahl der Artikel in der englischsprachigen Wikipedia maximal ist[2], verwendeten wir die Einträge daraus. Da die Auswahl geeigneter Merkmale im Wesentlichen einer Verschlagwortung der untersuchten Domäne gleichkommt, wurden unsere Ergebnisse auch besser, je mehr typische Schritte zur Schlagwortgewinnung wir durchführten (Stemming, Zerlegen in n-Gramme).

## 3.3 Auf diese Arbeit aufbauende weitere Ansätze

### 3.3.1 Evaluation durch Einschränkung der Merkmalsmenge

Die Ergebnisse bezüglich der gewonnenen Menge an Merkmalen sind schwer zu beurteilen. Es ist unbekannt, ob es eine optimale Menge an Merkmalen gäbe, die ein Maximum an Zuweisungen erzielte. Mit einiger Sicherheit kann man aber sagen, dass eine solche optimale Merkmalsmenge kleiner wäre, als die von uns verwendete Menge. Unsere finale Menge von Merkmalen wurde nicht weiter versucht zu verkleinern. Sie besteht aus allen Worten, die in den einführenden Sätzen der Wikipedia-Artikel vorkamen und die nach den geschilderten Schritten, insbesondere Stemming, verblieben sind, sowie den zugewiesenen Merkmalen aus der Wortmusterextraktion. Mit steigender Zahl der Merkmale wird es schwieriger, kohärente Begriffsverbände zu erzeugen, weil zuviele Begriffe zu wenige gemeinsame Merkmale haben, oder nur solche von sehr allgemeiner

[2] sowohl absolut als auch unterhalb der Kategorie „Ophthalmology“

Natur.

Eine Möglichkeit die Merkmalsmenge zu reduzieren und dieses Problem unmittelbar anzugehen, ist zu versuchen Synonyme in der Merkmalsmenge zu identifizieren, einen der synonymen Begriffe auszuwählen und alle anderen durch diesen zu ersetzen. Dies könnte in Teilen mittels eines Wörterbuchs automatisch erledigt werden, weil viele Bezeichnungen der Merkmale entweder nicht spezifisch aus der Domäne der Augenheilkunde sind, sondern Teil der normalen Sprache, oder allgemeinere medizinische Bezeichnungen sind, für die es beispielsweise ein englisches Wörterbuch mit Synonymen gibt. Dennoch: ein verbleibender Teil müsste unserer Ansicht nach per Hand erledigt werden. Mit dieser neuen Merkmalsmenge könnte man dann erneut versuchen Begriffsverbände zu bilden und Zuweisungen zu den Einträgen zu produzieren, um zu vergleichen, ob eine größere Zahl Zuweisungen erreicht wurde und ob die Anzahl gemeinsamer Merkmale gesteigert werden konnte.

Auf diesen Gedanken aufbauend, scheint es uns vielversprechend, mit Teilmengen unterschiedlicher Größe der ursprünglichen Merkmalsmenge zu experimentieren. Den bei den „echten" Synonymen verfolgten Ansatz könnte man hierzu erweitern und beispielsweise sämtliche Merkmale als Synonyme von einem aus 30, 50 oder 70[3] zuvor ausgewählten Merkmalen zu definieren. Das heißt, dass man alle verbliebenen Merkmale als Synonym je einem der ausgewählten Merkmale zuweisen müsste. Dem Grunde nach ist das eine Verschlagwortung mittels eines kontrollierten Vokabulars und man müsste sehr umsichtig sein, welche Bezeichnungen man unter einer Merkmalsbezeichnung zusammenfasst, um nicht Ungenauigkeiten oder sogar Fehler in den resultierenden Begriffsverbänden zu bekommen. Ein solcherlei erstelltes „Synonymwörterbuch" hätte aber den Vorteil, dass Merkmale, die aus neuen Quellen extrahiert wurden, ohne weiteres ihrem zu verwendenden Synonym zugewiesen werden können. Unterschiedlich große Mengen von Merkmalen, die auf diese Weise geschaffen wurden, erlauben es dann den Erfolg bei den Zuweisungen im „Wörterbuch der Augenheilkunde" zu evaluieren und ausgehend davon Modifikationen vorzunehmen.

Mit dem hierzu zu betreibenden manuellen Aufwand verabschiedete man sich aber, bis auf die Zuweisung im Wörterbuch und die eigentliche Extraktion, vollständig von jeglicher Automatisierbarkeit. Man könnte daher untersuchen, ob eine bestimmte Teilmenge der Menge aller Merkmale ausreicht, um die (in unserem Fall) Artikel aus der Wikipedia-Extraktion mit je mindestens zwei, drei oder mehr Merkmalen abzudecken. Sofern dies gelingt, könnte man wiederum die Anzahl der damit erzielten Zuweisungen im „Wörterbuch der Augenheilkunde" evaluieren und ausgehend davon Modifikationen vornehmen. Welche Merkmale dafür in Frage

[3] Diese Zahlen sind zwar beliebig gewählt, insbesondere das Intervall, sind aber insofern fundiert, als dass wir auf mindestens 30 Merkmale als untere Grenze kommen, um sowohl Aspekte der anatomischen Lage, der Pathologie, der diagnostischen und operativen Verfahren, sowie der Instrumente abzudecken.

kommen, könnte man über die Häufigkeitsverteilung der gefundenen Merkmale über alle Artikel aus der Wikipedia-Extraktion herausfinden. Interessant wären hier die Merkmale, die nicht mit absolut größter Häufigkeit vorkommen, und die nicht schon zu der sehr großen Menge Merkmalen zählt, die wenige oder nur ein Mal vorkommen. Dieser Bereich der Verteilung der Häufigkeit der Merkmale (insgesamt sind es 2443) ist etwa 100 bis 150 Merkmale groß mit einer Häufigkeit zwischen je 20 und 100. Er deckt die Artikel der Wikipedia-Extraktion vollständig mit mindestens einem Merkmal je Artikel ab[4].

Wir halten diese Ansätze für vielversprechend, weil sie erlauben würden zu ermitteln, welche Merkmalsmenge unbedingt notwendig (versus ausreichend versus optimal) ist, um die Domäne für die Formale Begriffsanalyse zu erschließen. Darauf aufbauend wäre dann aus unserer Sicht besser zu entscheiden, wie man eine so reduzierte Menge an Merkmalen maschinell erschließt (bzw. ob es überhaupt mit vertretbarem Aufwand möglich ist). Im Hinblick auf die praktische Verwendung ist der erstgenannte Ansatz mit dem manuellen Synonymwörterbuch daher nicht geeignet, weil fast alles, was die Formale Begriffsanalyse später überhaupt leistet, im Vorfeld schon manuell erledigt worden ist, indem man die Synonymlisten überhaupt geschaffen hat.

### 3.3.2 Vorgeben des Merkmalsraums

Die Merkmalsmenge einzuschränken und mit unterschiedlich großen Mengen zu experimentieren, kann weitere Erkenntnisse erbringen. Dieser Ansatz ist aber noch weiter zu verschärfen, indem man die zu verwendenden Merkmale selbst vorgibt und den Zuweisungsschritt auf das Zuweisen von extrahierten Merkmalen zu vorher festgelegten verschiebt.

Wie bereits geschildert wurde, ist die automatisierte Erzeugung von Merkmalen äußerst problematisch gewesen und war dadurch gekennzeichnet, dass an vielen Stellen manuell eingegriffen werden musste, um einen aussagekräftigen formalen Begriffsverband zu erzeugen. Dadurch ist der Merkmalsraum in seiner Struktur nicht so, wie man ihn aus der Literatur kennt, wo die Merkmale nach bestimmten Aspekten ausgesucht wurden, mit denen man die Objekte erschließen wollte.

Im Beispiel nach Janssen aus Tabelle 1.1 auf Seite 6 sind die Merkmale so gewählt, dass sie zur Kennzeichnung der Bedeutungen der einzelnen Bezeichnungen für Pferde herangezogen werden können und daher auch Sinn ergeben, nämlich Geschlecht und Alter. Bei der automatisierten Extraktion von Merkmalen aus externen Wissensquellen, wie der Wikipedia, entstehen mit hoher Wahrscheinlichkeit zu allgemeine Merkmale wie *Tier, Reittier* oder *Huf*. Im schlimmsten Fall sind es Merkmale, die gar nicht unmittelbar in das Bedeutungsumfeld der untersuchten Objekte

[4] Allerdings zeigt sich auch hier: lediglich einige Artikel sind dadurch mit drei oder mehr Merkmalen abgedeckt, die allermeisten nur mit einem Merkmal (häufig aber zuzüglich eines der Merkmale aus den ersten zehn Rangplätzen der Verteilung der Merkmale).

passen oder nichts mit ihnen gemeinsam haben, weil in der Wissensquelle bei der Erklärung des Begriffs zu weit ausgeholt wird (beispielsweise, wenn im ersten Satz im Wesentlichen die Etymologie geklärt wird.).

Daher schlagen wir vor, formale Kontexte zusätzlich auch unter Verwendung eines vorgegebenen Merkmalsraums zu bilden, um anschließend Methoden für die automatisierte Merkmalszuweisung zu erproben. Den Aufbau eines solchen Merkmalsraumes sollte unbedingt ein Fachkundiger - in unserem Fall ein Augenarzt - durchführen oder zumindest überwachen. Hierfür sollten möglichst allgemeine, aber dennoch günstig die Domäne abgrenzende, Merkmale verwendet werden. Solcherlei Synsets sollten nach Aspekten unterscheiden beziehungsweise gruppieren, die sehr gehäuft in der Domäne der Augenheilkunde auftreten und in der Betrachtung eines Fachkundigen Relevanz haben (zum Beispiel *Krankheit, Organ, Behandlungsmethode, diabetisch* und ähnliche). Gründe dafür sind:

- die automatisierte Merkmalsgenerierung ist problematisch, da oft viel Ballast zustande kommt (gerade bei Methoden zur linguistischen Verarbeitung von Texten)
- es ist unklar ob nach einem gegebenen Merkmal unterschieden werden soll
- je größer der Merkmalsraum, desto schwieriger ist es, diesen automatisiert mit Merkmalszuweisungen zu füllen, um einen aussagekräftigen formalen Kontext zu generieren
- bei Erstellung des Merkmalsraums mit intellektuellem Aufwand und mithilfe von Fachwissen hat man deutlich mehr Kontrolle über die Bildung des Begriffsverbandes

Da sich ein gewisser intellektueller Aufwand beim Aufbau einer Merkmalsmenge nicht vermeiden lässt, erscheint es akzeptabel, im Vorhinein die Merkmale selbst zu bestimmen. Es ist zu erwarten, dass sich auf diese Weise einige der geschilderten Probleme mindern lassen, wie die Aufnahme nicht relevanter Merkmale oder mangelnde Wiederverwendung der Merkmale auf Objekte aus dem Begriffsverband.

### 3.3.3 Nutzung der Systematisierten Nomenklatur der Medizin

Die „Systematized Nomenclature of Human and Veterinary Medicine“ ist, anders als die ICD-10, die eine Klassifikation ist, eine Nomenklatur medizinischer Terminologie mit dem Ziel den klinischen Alltag, das heißt Aufnahme, Diagnose, Behandlung und Abrechnung, auf eine einheitliche Kodierung zu bringen. Während die ICD-10 für die Erhebung statistischer und insbesondere epidemiologischer Daten geschaffen wurde, handelt es sich bei der SNOMED und ihren Nachfolgern um eine an der klinischen Praxis und ihren Bedürfnissen orientierte Begriffsordnung. Die aktuelle SNOMED 3 genannte Version enthält elf orthogonal zueinander stehende Achsen zur Kodierung

der Begriffe. Die Achsen bilden Aspekte medizinischer Terminologie ab. Beispielsweise enthält die Achse „Topography“[5] anatomische Begriffe wie die Bezeichnungen für Organe, Knochen, Muskeln, Sehnen und anderen. Die Achse „Morphology“ enthält Bezeichnungen für Veränderungen in Geweben, innerhalb von Körperzellen und Organen. Darunter fallen sowohl entzündliche Veränderungen, als auch degenerative oder traumatische Veränderungen. Weitere Achsen sind „Function“ für Symptome und Auffälligkeiten zur Beschreibung einer Diagnose (zum Beispiel „Fieber, Schüttelfrost, Übelkeit“) oder „Procedure“, welche sowohl medizinische als auch die Krankenhausverwaltung betreffende Maßnahmen enthält. Mit der Achse „Diagnosis“ gibt es auch eine weitreichende Verbindung in die ICD-10 hinein, da diese Achse die Bezeichnungen für Krankheiten, aber auch Verletzungen und andere Störungen umfasst. Die Kodierung erfolgt mittels Aneinanderreihung der Bezeichner aus der entsprechenden Achse, beginnend mit der ersten passenden Achse. Eine Appendizitis wird kodiert mit *T66000 M40000*, eine perforierende Appendizitis mit *T66000 M46300*, Krankheiten der Appendix allgemein mit *D62700* und eine Appendektomie mit *T66000 P11000*.

Aufgrund der Beschränkung auf elf Achsen und insbesondere aufgrund der Unterscheidung in Anatomie, Ursachen von Krankheiten, Diagnose und Prozeduren (und einigen anderen Dimensionen noch) scheint es uns lohnenswert, die Begriffe des Wörterbuchs der Augenheilkunde zur Passung mit SNOMED zu bringen beziehungsweise sie mittels der dortigen Notation auszudrücken. Viele der Extrakte aus der Wikipedia enthalten Wörter, die als Konzept in entsprechenden Achsen der SNOMED unmittelbar Verwendung finden, so dass sich aus unserer Sicht hier die Chance ergibt, die Verschlagwortung der Synsets mittels eines von Fachleuten erstellten Systems zu betreiben.[6]

## 3.4 Schlussbemerkungen

Bei oberflächlicher Inaugenscheinnahme der Ergebnisse kann man feststellen, dass wir die ursprünglich formulierten Ziele erreicht haben. Allerdings sind dabei viele Einschränkungen zu machen. So sind die gebildeten Begriffsverbände nicht minimal. Während es zwar fraglich ist, ob die kleinstmögliche Merkmalsmenge, die noch genügt um alle Begriffe des Wörterbuchs zu bilden, auch optimal ist, kann man sicher festhalten, dass unsere Merkmalsmenge zu groß ist. Für die meiste Zeit des Projekts waren wir aber beschäftigt eine möglichst große Merkmalsmenge zu

[5] Nachfolgend werden die englischen Achsenbezeichner verwendet, da damit die resultierende Kodierung deutlicher wird.

[6] Stichprobenartige Versuche ausgewählte erste Sätze der Wikipedia-Extraktion in den „SNOMED CT Categorizer“ http://eagl.unige.ch/SNOCat/index.jsp der Universität Genf zu geben, zeigen ermutigende Ergebnisse im Hinblick auf die Abdeckung. Dieses System soll eigentlich Formulierungen wie sie in Arztbriefen vorkommen in SNOMED-Kategorien übersetzen

finden, ohne auf die Passung mit dem „Wörterbuch der Augenheilkunde“ Rücksicht zu nehmen. Die Begriffe des Wörterbuchs dann mit den Titeln der Wikipedia-Artikel zur Passung zu bringen und im Erfolgsfall Merkmale zuzuweisen, zeigte sich als deutliche Einschränkung. Vieles, das insbesondere aus der Wikipedia-Extraktion kam, fand keine Verwendung, weil es keine Entsprechung im Wörterbuch hatte. Das Wörterbuch ist aber weit entfernt davon vollständig zu sein. Wir wissen also eigentlich gar nicht, was die Bedingungen sind, die man im Hinblick auf die Merkmalsmenge braucht, um die Formale Begriffsanalyse besser in einem Umfeld wie dem „Wörterbuch der Augenheilkunde“ nutzbar zu machen, weil wir zu diesem Punkt nicht gekommen sind. Um das quantifizieren zu können, hätten wir für die Evaluation geeignete Teilmengen der Merkmalsmenge benötigt. Aus unserer Sicht wäre es günstig gewesen diese Tests an den Wikipedia-Artikeln, aus denen die Merkmale ursprünglich stammten, vorzunehmen und das „Wörterbuch der Augenheilkunde“ anfangs außen vor zu lassen, bis die Frage der geeigneten Merkmalsmenge geklärt ist.

Wenngleich noch in eher experimentellem Stadium, so ist die Visualisierung der Ergebnisse auf den Seiten des Wörterbuchs der Augenheilkunde eine nützliche und wertvolle Bereicherung. Anders als Tagclouds und ähnliche „siehe auch“-Navigationselemente, fußt der gezeigte Begriffsverband auf hierarchisch korrekter Information (sofern die Extraktion der Information korrekt ist). Wir sind der Ansicht, dass es sich lohnt, diese Art der Anreicherung von Wörterbucheinträgen mit hierarchischer Information weiter zu verfolgen.

Im Hinblick auf die Projektplanung wäre unsere Wunschvorstellung bezüglich der Aufgabenstellung, nunmehr im Lichte unserer Erfahrungen, folgendermaßen gewesen:

1. Identifizieren einer geeigneten Quelle zur Merkmalsextraktion (liefe wahrscheinlich auf Wikipedia hinaus)

2. Testen und Verbessern der Möglichkeiten der vollautomatischen Merkmalsextraktion

3. Innerhalb der Wikipediadaten: Bilden von Begriffsverbänden und Testen und Evaluieren unterschiedlich großer Merkmalsmengen

4. Am Schluß wieder der Versuch einer Anbindung an das „Wörterbuch der Augenheilkunde“, gegebenenfalls durch Ergänzen mit in der Wikipedia vorhandenen Einträgen und Umbenennen von Einträgen für eine bessere Passung mit den Einträgen aus der Wikipedia.

# 4 Anhang

## 4.1 Beispiel für Verarbeitung der Extrakte aus der Wikipedia

Basierend auf den als relevant erachteten Kategorien der Wikipedia, wurden alle Artikel, denen ein gegebenes Kategorie-Tag zugewiesen war, geholt (siehe auch *Die Kategorie-Tags in der Wikipedia* auf Seite 15). Der typische Aufbau eines solchen Artikels im Quelltext ist nachfolgend dargestellt. Von Interesse sind hier der Titel des Dokuments, die Kategorien und das Extrakt selbst. Die Kategorien wurden zum Abgleich mit der ursprünglichen Kategorieliste extrahiert. Dabei gab es keine Abweichungen, der Schritt ist daher in Zukunft vermeidbar.

*Listing 4.1:* XML-Quelltext eines typischen Artikels der Wikipedia: „Parinaud's syndrome“

```
1  <?xml version="1.0"?>
2  <api>
3  <query>
4  <pages>
5  <page pageid="3055944" ns="0" title="Parinaud&#039;s syndrome">
6  <categories>
7  <cl ns="14" title="Category:Diseases of the eye and adnexa" />
8  <cl ns="14" title="Category:Medical signs" />
9  </categories>
10 <extract xml:space="preserve">
11 Parinauds Syndrome, also known as dorsal midbrain syndrome is a group of abnormalities
   of eye movement and pupil dysfunction.
12 </extract>
13 </page>
14 </pages>
15 </query>
16 </api>
```

*Listing 4.2:* Extraktion des Titels aus Zeile 5 des Artikels „Parinaud's syndrome“. HTML-Entities wie ' für hochgestellte einzelne Anführungszeichen in Zeile 5 wurden gesammelt und in der Ergebnisdatei ersetzt.

```
1  sed -n 's/<page.*title="\([^"]*\)".*$/\1/p' Parinauds_syndrome.xml
```

*Listing 4.3:* Extraktion des Textes aus der der Zeile 10 nachfolgenden Zeile des Artikels „Parinaud's syndrome“

```
1  sed -n 's/<extract.*>\([^<]*\)\([<$ ].*\)/\1/p' Parinauds_syndrome.xml
```

## 4.2 Erzeugen eines Ordnungsdiagramms

### 4.2.1 Die Auszeichnungssprache DOT

Die Auszeichnungssprache DOT kann verwendet werden zur einfachen Beschreibung von Graphen. Sie wird von unterschiedlichen Programmen interpretiert, die meisten sind aber Teil des Programmpaketes GraphViz. Diese Programme erlauben die Ausgabe des Graphen in einer Vielzahl von Dateitypen für Raster- und Vektorgrafiken. Nachfolgend wird die Syntax am Beispiel eines aus unseren Daten generierten Begriffsverbands gezeigt. Als Ausgabestrom erzeugt *dot* dann SVG-Syntax, die direkt in das Markup der HTML-Seite eingebettet werden kann. Erzeugt wurde das Diagramm aus dem Begriffsverband des Synsets *Macula* mit dem Script „visualize_new.php“. Die Formatierung erlaubt es Knoten und Kanten, sowie deren Attribute deutlich hervorzuheben ohne eine Rastergrafik zu verwenden, weil die Auszeichnung des Diagramms in reinem Text im Quelltext des Dokumentes erfolgt.

*Listing 4.4:* Dot-Syntax für Ordnungsdiagramm des Begriffsverbands für das Synset *Macula*. Vereinfachte Darstellung.

```
1  strict digraph G {
2    0 [ label=190,
3      URL="/visualize_new.php?WID=1895&show=1"];
4    1 [ label=598,
5      URL="/visualize_new.php?WID=1895&show=2"];
6    2 [ label=46,
7      URL="/visualize_new.php?WID=1895&show=4"];
8    3 [ label=103,
9      URL="/visualize_new.php?WID=1895&show=8"];
10   4 [ label=50,
11     URL="/visualize_new.php?WID=1895&show=9"];
12   5 [ label=79,
13     URL="/visualize_new.php?WID=1895&show=10"];
14   1 -> 0;
15   1 -> 5;
16   4 -> 2;
17   0 -> 4;
18   5 -> 4;
19   3 -> 5;
20 }
```

### 4.2.2 Berechnen der Potenzmenge der Attribute

Anhand der folgenden Funktion wird die Potenzmenge der Attribute errechnet, mit der die Graphstruktur aufgebaut wird.

*Listing 4.5:* Potenzmenge der Attribute

```
1 function getPowerset($array) {
2   $power = array(array());
3   foreach ($array as $element)
4   foreach ($power as $combi)
5     array_push($power, array_merge(array($element),$combi));
6   return $power;
7 }
```

### 4.2.3 Bestimmung von Untermengen und transitiven Kanten

Im Folgenden sieht man die Methoden, mit denen Untermengen von Attributmengen und transitive Kanten identifiziert werden. Die Zuweisung von Ober- und Untermengen ist für die Erzeugung der Kanten und somit des kompletten hierarchischen Aufbaus des Graphen notwendig. Die Erkennung von transitiven Kanten (Kanten, zwischen denen ein oder mehrere Knoten liegen) ist wichtig, um überflüssige Kanten ausfindig zu machen, da in Ordnungsdiagrammen Kanten nur Verbindungen zu direkten Ober- bzw. Unterknoten herstellen und zusätzliche Kanten, die Zwischenknoten überspringen, überflüssig sind.

*Listing 4.6:* Überprüfung, ob `$a1` Untermenge von `$a2` ist

```
1 function is_subset($a1, $a2) {
2   foreach ($a1 as $word)
3     if (!in_array($word, $a2)) return false;
4   return true;
5 }
```

*Listing 4.7:* Überprüfung, ob eine Kante überflüssig ist. Wenn der Ausgangsknoten beider Kanten gleich ist, wird überprüft, ob ein Alternativpfad existiert.

```
1 function is_transitive($edge) {
2   global $edges;
3   foreach ($edges as $ecomp)
4     if ($edge != $ecomp && $edge[0] == $ecomp[0]) {
5       if (sameRoute($edge, $ecomp)) return true;
6     }
7   return false;
8 }
```

*Listing 4.8:* Überprüfung, ob `$ecomp` über andere ausgehende Kanten zum Zielknoten von `$ebasic` führt.

```
1 function sameRoute($ebasic, $ecomp) {
2   global $edges;
3
```

```
4  // true-Bedingung: Zielknoten sind identisch = Kante überflüssig (wg. Alternativroute)
5
6    if ($ebasic[1] == $ecomp[1]) return true;
7
8  // wenn ausgehende Kante...
9
10   foreach ($edges as $next) if ($ecomp[1] == $next[0]) {
11
12 //... dann rekursiv mit dieser Kante Überprüfung fortsetzen.
13
14     if (sameRoute($ebasic, $next)) return true;
15
16   }
17   return false;
18 }
```

## 4.3 Beispiele der Visualisierung der Begriffsverbände

**Attributes**
keratitis
**Synsets**
herpes keratitis || keratitis ||
scrofulous keratitis || keratitis
punctata || flash
keratoconjunctivitis ||
neurotrophic keratitis ||
peripheral ulcerative keratitis ||
thygeson's disease || diffuse
lamellar keratitis || filamentary
keratitis || dendritic keratitis ||
marginal keratitis || nummular
keratitis || disciform keratitis ||
acanthamoeba keratitis

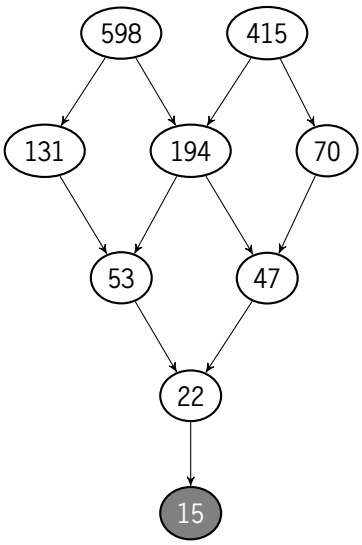


*Abbildung 4.1:* Nachbildung der Visualisierung eines Begriffsverbands, wie das Skript ihn als Ausgabe produziert.

**Attributes**
Anterior chamber || illness || optic
**Synsets**
anterior ischaemic optic neuropathy || arteritic anterior ischemic optic neuropathy || non-arteritic anterior ischemic optic neuropathy ||

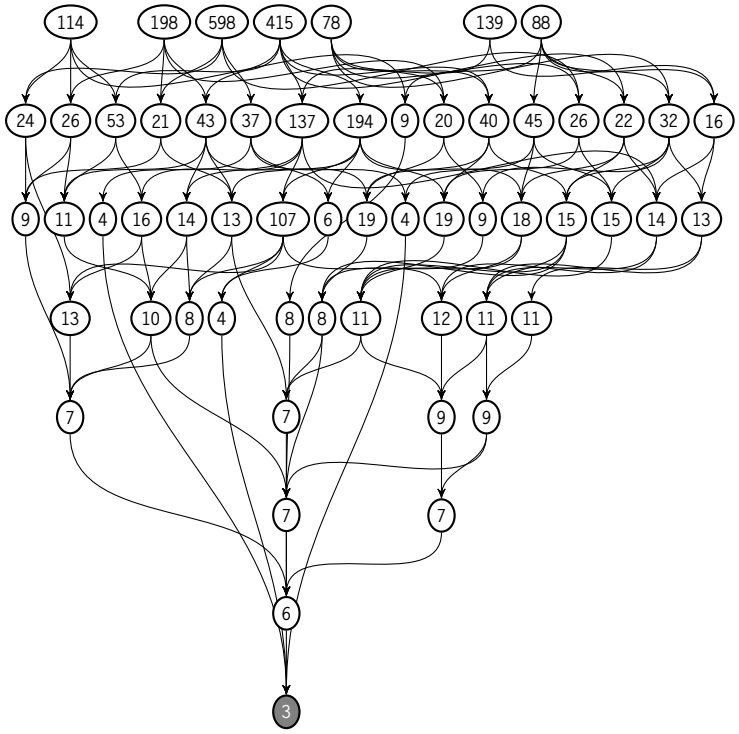


*Abbildung 4.2:* Nachbildung der Visualisierung eines Begriffsverbands, wie das Skript ihn als Ausgabe produziert.