

Von der Erstellung bis zur Nutzung: Wortnetze als XML Topic Maps

Abstract

Wir beschreiben exemplarisch anhand eines im Projekt „Hypertextualisierung auf textgrammatischer Grundlage“ (HyTex)¹ entwickelten Wortnetzes, wie ein nach dem WordNet-Modell strukturiertes Wortnetz mit texttechnologischen Methoden erstellt, gewartet, als XML Topic Map (XTM) repräsentiert, visualisiert und zum Hypertext-Linking genutzt werden kann. Im Mittelpunkt steht dabei die Präsentation im XTM-Format, die zu anderen technischen Repräsentationsformalismen für Wortnetze in Beziehung gesetzt wird.

1 Einleitung

Mittlerweile gibt es eine große Anzahl von WordNet-Derivaten (im folgenden „Wortnetze“ genannt) für unterschiedliche Sprachen und Anwendungen, denen allen die Struktur aus Lexemen, Konzepten (Synsets), lexikalischen Relationen und konzeptuellen Relationen gemeinsam ist. Sie werden technisch unterschiedlich repräsentiert, z. B. durch Datenbanken, verschiedenartige XML-Repräsentationen, oder RDFS.

Wir stellen eine weitere Repräsentation vor, die sich als Austauschformat für Wortnetze sehr gut eignet: **XML Topic Maps**.

XML Topic Maps können als standardisiertes Format für semantische Netze aufgefasst werden, das einige interessante Eigenschaften aufweist, darunter die Möglichkeit der Anbindung von Dokumenten oder Dokumentteilen an das semantische Netz und Mechanismen zur Vereinigung mehrerer Topic Maps. Da es sich um ein von der ISO standardisiertes, SGML- oder XML-basiertes Austauschformat handelt, ist es platt-

formunabhängig und kann mit unterschiedlicher Software verarbeitet werden.

In Abschnitt 2 beschreiben wir, wie Wortnetze zur Zeit repräsentiert werden. Am Beispiel eines terminologischen Wortnetzes, das wir im Rahmen des DFG-geförderten Projekts „Hypertextualisierung auf textgrammatischer Grundlage“ (HyTex, siehe auch RUNTE ET AL. 2003, in diesem Band) nutzen, zeigen wir in den Abschnitten 3 bis 7, wie ein Wortnetz erstellt, gepflegt, als XML Topic Map repräsentiert, visualisiert und zum Hypertext-Linking genutzt werden kann. In den Abschnitten 8 und 9 gehen wir kurz darauf ein, welche Vorteile die Topic-Map-Modellierung auch für andere Wortnetz-Projekte haben könnte und wie die Repräsentationen von Wortnetzen als XML Topic Maps und als RDF(S) einander ergänzen könnten.

2 Bisherige Wortnetz-Repräsentationen

Es gibt verschiedene Praktiken und Vorschläge zur Repräsentation von Wortnetzen. Lemnitzer und Kunze stellen eine konzeptuelle Modellierung in Form von Entity-Relationship-Diagrammen vor, die als Ausgangsbasis für verschiedene technische Repräsentationen – Textformate, Datenbanken oder XML-Formate – dienen kann (KUNZE & LEMNITZER 2002; LEMNITZER & KUNZE 2003).

Eine solche technische Repräsentation kann verschiedenen Zwecken dienen, z. B. als einfach zu editierendes, menschenles- und schreibbares Format, als intermediäres Format zum Austausch zwischen verschiedenen Verarbeitungsschritten, als Ausgangsformat für die Publikation in verschiedenen Medien, als Austauschformat für die Integration verschiedener Wortnetze,

oder als Speicherformat für den schnellen Zugriff in Information-Retrieval-Anwendungen.

Rein textbasierte Formate sind am wenigsten standardisiert und erfordern die Entwicklung spezieller Parser, um sie auszulesen. Ein Beispiel ist das vom Princeton WordNet verwendete, von Menschen leicht lesbare und editierbare textbasierte Format, das als Ausgangsbasis für die automatische Erzeugung eines Datenbankformats verwendet wird (BECKWITH ET AL. 1993). Ein solches Datenbankformat wiederum ist zwar nur schwer oder überhaupt nicht mehr menschenlesbar (in Abhängigkeit von der verwendeten Datenbank), ermöglicht dafür aber einen schnellen maschinellen Zugriff, wie er für Anwendungen im Information Retrieval unabdingbar ist.

XML-basierte Formate stellen einen Kompromiss zwischen diesen beiden Extremen dar. Sie sind einerseits menschenlesbar, andererseits aber auch für die maschinelle Verarbeitung sehr gut geeignet. Sie stellen einen Schritt in Richtung Standardisierung dar, so dass eine Reihe von Tools zur Verarbeitung bereits zur Verfügung stehen (z. B. XML-Parser, die also für ein spezielles XML-basiertes Format nicht eigens entwickelt zu werden brauchen). Für einen effizienten Zugriff ist eine Konvertierung in ein Datenbankformat möglich. Es wurden mindestens zwei solcher XML-basierten Formate mit jeweils eigenen Dokumentgrammatiken (DTDs) für Wortnetze entwickelt: für GermaNet (KUNZE & LEMNITZER 2002) und für einen WordNet-Editor (PAVELEK & PALA 2002). Das für GermaNet entworfene Format benutzt zudem einen standardisierten Verweismechanismus, XLink (DE-ROSE ET AL. 2001), zur Repräsentation der lexikalischen und der konzeptuellen Relationen.

Zwei in unterschiedlichen XML-Formaten repräsentierte Wortnetze sind jedoch noch lange nicht kompatibel. Während auf der Ebene der Syntax eine Standardisierung erreicht ist, wird nach wie vor spezielle Software benötigt, die die *Semantik* der speziellen XML-Modellierung

auswertet. Mit zwei relativ neuen WWW-Standards kann die Semantik von netzartigen Informationsstrukturen, wie sie in Wortnetzen vorliegen, zumindest teilweise erfasst werden: Resource Description Framework (RDF, LASSILA & SWICK 1999) mit den darauf aufbauenden Standards RDF Schema (RDFS) und Web Ontology Language (OWL) sowie Topic Maps (PEPPER & MOORE 2001).

Bei RDF handelt es sich um ein sehr elementar gehaltenes Modell zur Beschreibung von Metadaten, wobei das Metadaten-Set allerdings beliebig erweiterbar ist, so dass sich nahezu beliebige Sachverhalte ausdrücken lassen. Mit RDFS, der Schemasprache von RDF, ist die Validierung von erweiterten RDF-Dokumenten möglich. OWL ist eine RDF(S)-basierte, sehr mächtige Sprache zur Definition von Ontologien, die im Wesentlichen ebenfalls für die Beschreibung von Metadatenstrukturen gedacht ist. Auf Topic Maps wird im Folgenden noch genauer eingegangen.

Durch die Nutzung von Topic Maps oder RDF für die Repräsentation von Wortnetzen wird ein weiterer Schritt in Richtung Standardisierung vollzogen, so dass die Austauschbarkeit und Verarbeitung durch Standard-Software erleichtert wird. Sowohl für RDF als auch für Topic Maps gibt es – neben anderen – eine XML-Syntax.

Ein Teil des in Princeton entwickelten ursprünglichen WordNet ist bereits in Form von RDF und RDFS aufbereitet worden und unter der Netzadresse <http://www.semanticweb.org/library/> abrufbar. In der dort vorgeschlagenen Modellierung werden jedoch nur Substantive, Glosses, SimilarTo-Relationen und Hyperonymie-Relationen berücksichtigt. Ein weitergehender Vorschlag zur Abbildung von Wortnetzen auf RDF(S)-Strukturen wird in LEMNITZER & KUNZE 2003 beschrieben.

3 Arbeitsschritte von der Erstellung bis zur Nutzung

Im HyTex-Projekt verwenden wir ein nach den Prinzipien von WordNet modelliertes terminologisches Netz auf eine ganz spezifische Weise, nämlich zur Nutzung in einem ontologiebasierten Hypertext (MILES-BOARD ET AL. 2001). Im gesamten Projekt, also auch bei allen Verarbeitungsschritten des Wissensnetzes, verwenden wir standardisierte, XML-basierte texttechnologische Standards. Auf diese Weise können die entstehenden Zwischenprodukte anderen Projekten zur Verfügung gestellt werden, z. B. kann die im XTM-Format vorliegende Topic Map von jeder Art von Software genutzt werden, die XTM verarbeiten kann. Verarbeitungsschritte, die die von uns eingesetzte Software nicht leistet, können wir selbst programmieren, z. B. spezielle Arten von Inferenzen auf dem Wissensnetz. Ein weiterer Vorteil liegt darin, dass die von uns verwendeten Standards alle plattformunabhängig sind. Schließlich ist es uns jederzeit möglich, einzelne Komponenten, die zur Verarbeitung notwendig sind, durch andere zu ersetzen (z. B. einen anderen Editor zu verwenden).

Die nachfolgend genannten Arbeitsschritte, welche wir im HyTex-Projekt von der Erstellung bis zur Nutzung des terminologischen Netzes durchführen, haben daher nur beispielhaften Charakter – der texttechnologische Ansatz ermöglicht es, die Vorgehensweise an jeder beliebigen Stelle der Verarbeitungskette an andere Erfordernisse anzupassen:

1. Eingabe und Wartung des Wissensnetzes mit dem Werkzeug K-Infinity und Export in eine K-Infinity-eigene XML-basierte Repräsentation
2. Konvertierung des K-Infinity-Exportformats nach XML Topic Maps (XTM)
3. Durchführung von Inferenzen und Überprüfungen auf dem Wissensnetz mittels XSLT

4. Überführung in ein grafisch aufbereitetes, strukturiertes Glossar in einen Hypertext (repräsentiert in HTML und SVG) mittels XSLT.

An verschiedenen Stellen kommt in dieser Kette die Programmiersprache XSLT zum Einsatz, eine funktionale Programmiersprache, die darauf optimiert ist, XML-Dokumente einzulesen und zu erzeugen. Nur der erste dieser Schritte geschieht manuell, alle anderen können automatisch durchgeführt werden. Die Schritte werden in den folgenden Abschnitten genauer beschrieben.

4 Erstellung mit K-Infinity

Komplexe Wissensnetze aufzubauen und zu pflegen ist eine aufwändige Aufgabe, die man am besten mit einer spezialisierten Software erledigt, die über eine Visualisierungs- und Verwaltungskomponente für Netzstrukturen und Mechanismen zur Konsistenzprüfung verfügt. Solche Mechanismen sollten z.B. sicherstellen, dass beim Löschen eines Konzeptes im Wortnetz auch alle Lexeme gelöscht werden, die dieses Konzept lexikalisieren. Wird ein Konzept gelöscht, das in einer Hyperonym-Beziehung zu anderen Konzepten steht, muss geprüft werden, wie mit den Hyponymen verfahren werden soll, ob sie gelöscht oder mit dem nächsthöheren Konzept in der Hierarchie (als Hyponyme) verbunden werden sollen. Auch die Vergabe von IDs und Verweisen auf IDs ist eine aufwändige und fehleranfällige Aufgabe, für die man am besten spezialisierte Editoren nutzen sollte.

Für die Forschungen im HyTex-Projekt, speziell für den Aufbau und die Pflege des TermNet, hat uns die Firma intelligent views das komfortable Werkzeug *K-Infinity* zur Verfügung gestellt (vgl. <http://www.i-views.de/>). *K-Infinity* unterstützt den Aufbau und das Editieren von Wissensnetzen mit einer grafischen Oberfläche (siehe Abb. 1), in der die Entitäten des Wissensnet-

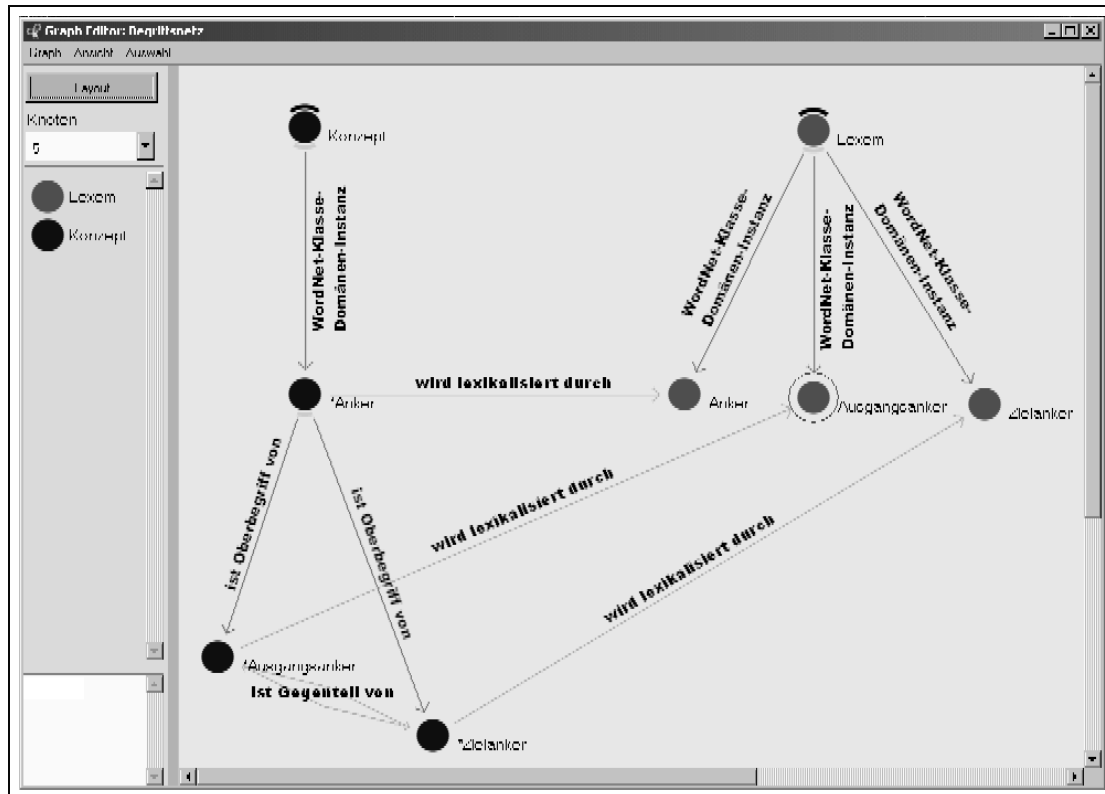


Abbildung 1: Das Werkzeug zur Verwaltung von Wissensnetzen K-Infinity ermöglicht u. a. eine grafische Eingabe von Begriffen und Relationen.

zes direkt manipuliert werden können. Es automatisiert die Verwaltung von IDs und das Umsetzen von Verweisen beim Löschen von Konzepten (in K-Infinity als „Begriffe“ bezeichnet). Weil auch nachträgliche Umbenennungen von Konzepten und Relationen ohne größeren Aufwand möglich sind, kann der Aufbau des Netzes flexibel an die Anforderungen der jeweiligen Anwendung angepasst werden, was gerade in einem Forschungsprojekt von großem Vorteil ist.

Allerdings gibt es gerade bei der Modellierung von Wortnetzen (im Stil des WordNet) eine Reihe von spezifischen Anforderungen (z. B. bestimmte Inferenzen), die mit K-Infinity nicht durchgeführt werden können. Der in HyTex verfolgte texttechnologische Ansatz ermöglicht es uns aber, sie an einer späteren Stelle in der Verarbeitungskette (siehe Abschnitt 6) zu lösen. K-Infinity bietet nämlich die Möglichkeit, ein eingegebenes Wissensnetz in ein K-Infinity-eigenes XML-basiertes Format zu exportieren. Die Fir-

ma intelligent views hat uns darüber hinaus ein XSLT-Stylesheet zur Verfügung gestellt, das dieses Format automatisch in das XML Topic Map-Format überführt. Die dann vorliegende Repräsentation wird im Folgenden beschrieben.

5 Repräsentation des terminologischen Netzes als XML Topic Map

Topic Maps stellen eine standardisierte Notation zur Repräsentation von Netzwerken aus Informationseinheiten dar. Es gibt zwei syntaktische Varianten des Topic Map Standards: 1999 erschien er als ISO-Standard (ISO, 2000), der auf einer SGML-Syntax basiert. Im Jahr 2001 wurde eine XML-Syntax zur Nutzung im WWW für den ISO-Standard entwickelt: XML Topic Maps (XTM, PEPPER & MOORE 2001), ein Industriestandard, der inzwischen in den ISO-Standard integriert wurde. Eine sehr gute Einführung in Topic Maps findet sich bei RATH 2002. Wir beschreiben hier nur diejenigen Eigenschaften

Wortnetze als XML Topic Maps

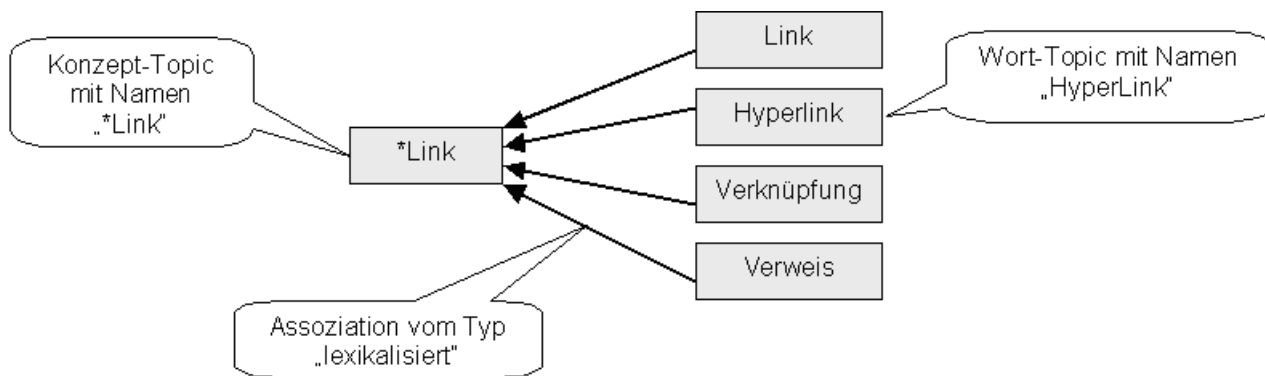


Abbildung 2: Abbildung von Konzepten und Lexemen auf Topics in der Topic Map.

von Topic Maps, die für die Repräsentation von Wortnetzen interessant sind.

Grundbausteine von Topic Maps – die Knoten des Netzwerks – sind sogenannte *Topics*. Die semantischen Beziehungen zwischen Topics – die Kanten – heißen *Assoziationen* (associations). Ein Topic kann über *Topic-Anker* (occurrences) zudem mit beliebigen adressierbaren Ressourcen, z. B. HTML-Dokumenten, verknüpft werden. In diesem Fall lassen sich Topics und Assoziationen als Metadaten zu den Dokumenten betrachten, und es ergeben sich vielfältige Anwendungsmöglichkeiten für Navigation und Suche, die wir hier aber nicht diskutieren.

Topics, Assoziationen und Topic-Anker lassen sich typisieren. Die Typen sind in Topic Maps selbst wiederum Topics, so dass sich über Typen innerhalb desselben Formalismus Aussagen machen lassen, z. B. können Typen selbst wieder Typen haben oder durch Assoziationen mit anderen Topics verbunden werden. Dies macht Topic Maps zu einem sehr mächtigen Standard. Die Typisierung ist äquivalent mit einer Kante vom Typ *instance-of*, durch sie wird also ausgedrückt, dass zwischen einem Topic und seinem Typ eine Klasse-Instanz-Relation besteht.

Jedes Topic hat einen obligatorischen eindeutigen Bezeichner (ID), und (optional) einen oder mehrere Namen.

Entsprechend der in Wortnetzen vorhandenen grundlegenden Unterscheidung zwischen Konzepten und Lexemen enthält unsere Topic

Map zwei Arten von Topics: Konzept-Topics und Wort-Topics.

Konzept-Topic: Für jedes Konzept der Domäne, für das es terminologisierte Ausdrücke gibt, führen wir ein Topic ein. Es hat einen Namen, der mit einem * gekennzeichnet ist. Jedes Konzept-Topic verbinden wir über eine Assoziation vom Typ `WordNet-Klasse-Domänen-Instanz` mit einem Topic des Namens *Konzept*.

Wort-Topic: Für jeden in der Fachdomäne terminologisierten Ausdruck – d.h. jedes Lexem, das einen Terminus darstellt – deklarieren wir ebenfalls ein Topic in der Topic Map. So erhalten wir Wort-Topics mit den Namen *Link*, *Hyperlink*, *Verknüpfung* und *Verweis*. Jedes Wort-Topic verbinden wir ebenfalls über eine Assoziation vom Typ `WordNet-Klasse-Domänen-Instanz` mit einem Topic des Namens *Lexem*.

Um nun die Zugehörigkeit der Wort-Topics zu ihrem jeweiligen Konzept-Topic auszudrücken, werden alle Wort-Topics eines Konzepts mit dem Konzept-Topic durch eine Assoziation vom Typ `lexikalisiert` verbunden (vgl. Abbildungen 1 und 2).

Lexikalische und konzeptuelle Relationen lassen sich nun auf getypte Assoziationen zwischen den Wort-Topics bzw. den Konzept-Topics abbilden. Ein Assoziationstyp für eine lexikalische Re-

lation ist ist Abkürzung für. Diese Relation besteht z. B. zwischen den Lexemen (bzw. Wort-Topics) *Link* und *Hyperlink*. Ein Beispiel für eine konzeptuelle Relation ist die Hyperonymie, sie besteht z. B. zwischen den Konzepten (bzw. Konzept-Topics) **Link* und **1:n-Link*. Die Assoziationstypen werden selbst wiederum als Topics repräsentiert.

Im HyTex-Projekt wird aus der Topic Map später vollautomatisch ein Glossar erzeugt. Für die hypertextuelle Verwendung verbinden wir die Wort-Topics durch Topic-Anker (occurrences) mit verschiedenen Textstellen des Korpus, z. B. mit Definitionen der Termini und mit Termverwendungsinstanzen. Diese Topic-Anker haben ebenfalls verschiedene Typen. Daraus werden später automatisch (typisierte) Hyperlinks von den Korpus-Dokumenten zum Glossar und umgekehrt erzeugt.

Einige Konzepte werden in unserer Modellierung des terminologischen Netzes durch ein Attribut gekennzeichnet, um verschiedene Arten von Ko-Hyponymie in Fachsprachen ökonomisch ausdrücken zu können, siehe dazu RUNTE ET AL. 2003, in diesem Band. Ein Attribut wird ebenfalls durch einen Topic-Anker repräsentiert, welcher auf eine in das Konzept-Topic eingebettete Ressource verweist, die den Attributwert enthält.

Ein weiteres Konstrukt von Topic Maps – neben Topics, Topic-Typen, Assoziationen, Assoziationstypen, Topic-Ankern und Topic-Anker-Typen – ist das Konstrukt des Skopus (scope). Ein Skopus ist ein Gültigkeitsbereich oder Kontext, in dem eine Aussage gültig ist. Im Semantic Web ist diese Möglichkeit von großer Bedeutung, da unterschiedliche Organisationen, Gruppen, und Menschen verschiedene Sichtweisen auf die Welt haben, die in Topic Maps nebeneinander existieren können. Zum Beispiel kann derselbe Gegenstand unterschiedlich benannt werden, oder in verschiedenen Klassifikationssystemen unterschiedlich eingeordnet werden. Auch ein Sko-

pus ist selbst wieder ein Topic. Bei Bedarf benutzen wir die Skopen, um die Topics der Domäne von denjenigen Topics zu unterscheiden, die der WordNet-Modellierung dienen: Alle Topics, die innerhalb der Topic Map als Typen dienen (z. B. das Topic mit dem Namen *ist Abkürzung für*) bekommen den speziellen Skopus *TermNet* zugewiesen. Es wurde vorgeschlagen, eine solche Trennung zwischen „regulären“ und „deklarativen“ Topics auf andere Weise vorzunehmen (Topic Map Templates, RATH 2000). Eine Standardisierung dieses Verfahrens in Form einer Schemasprache für Topic Maps (Topic Map Constraint Language, TMCL, <http://www.isotopicmaps.org/tmcl/>) ist jedoch noch nicht abgeschlossen.

6 Inferenzen und Überprüfungen auf dem Wissensnetz

Die Datenhaltung in der Topic Map ist redundanzfrei: Redundante Information, wie z. B. die explizite Modellierung der Synonymie-Relation, ist aus Gründen der schwierigeren Lesbarkeit, Änderbarkeit und weniger effizienten Speicherung und Verarbeitung nicht erwünscht. Für einen Nutzer, dem das Wissensnetzes auszugsweise präsentiert wird, kann solche redundante Information jedoch sehr wertvoll sein. Aus diesem Grund führen wir zwei Arten von Inferenzen auf dem Wissensnetz aus, die explizite Relationen erzeugen, welche vorher nur implizit vorhanden waren:

1. Wir inferieren die Synonymie-Relation zwischen verschiedenen Lexemen aus ihrer Verbindung mit dem Konzept (über die *lexikalisiert*-Relation). Das Ergebnis wird dem Nutzer grafisch präsentiert (Abbildung 3).
2. Die Relation der Disjunktivität wird aus den in Abschnitt 5 eingeführten Attributen und Attributwerten abgeleitet: Ko-Hyponyme, die gleiche Attributwerte aufweisen, sind disjunkt.

Außerdem führen wir begrenzt Konsistenzprüfungen durch, die für unser Wissensnetz spezifisch sind und daher nicht schon von K-Infinity vorgenommen werden können. Wir überprüfen, ob jedes Konzept mit mindestens einem Lexem verbunden ist und ob jedes Lexem mindestens einem Konzept zugeordnet ist. Wenn dies nicht der Fall ist, werden Warnmeldungen ausgegeben. Von Fischer wurden eine Reihe weiterer Konsistenzprüfungen für Wortnetze vorgeschlagen und durchgeführt; für WordNet: FISCHER 1997, für GermaNet: GUPTA 2002. Einige dieser Überprüfungen müssen wir nicht vornehmen, da sie bereits durch den Wissensnetz-Editor K-Infinity abgefangen werden (z.B. zyklische oder „abgekürzte“ Hyperonymie-Relationen) oder sich auf Relationen beziehen, die im TermNet nicht verwendet werden. Die Durchführung anderer Überprüfungen ist wünschenswert.

Die Inferenzen und Überprüfungen geschehen derzeit mit einem XSLT-Stylesheet. Prinzipiell sind natürlich beliebige andere Inferenzen auf dem Wissensnetz möglich. Das Ergebnis der Inferenzen ist wiederum eine in XTM repräsentierte Topic Map, die gegenüber der ursprünglichen Topic Map um einige Assoziationen erweitert ist. Die Programmierung in XSLT ist relativ aufwändig, da die Netzstruktur in XTM nicht direkt auf das (intrinsisch hierarchische) XML-Modell abgebildet werden kann und XTM daher stark mit Verweisen arbeitet. Von der ISO wird jedoch derzeit ein Standard geplant, der eine Anfragesprache für Topic Maps spezifiziert, vergleichbar mit SQL für relationale Datenbanken. Diese „Topic Map Query Language“ (TMQL) wird es später erlauben, Inferenzen auf einer Topic Map auf einfachere Weise zu formulieren, als es derzeit mit XSLT möglich ist.

7 Nutzung als Hypertext und Visualisierung

Aus der Topic Map wird automatisch – ebenfalls durch ein XSLT-Stylesheet – ein Glossar er-

zeugt, aus dem der Benutzer Informationen über Fachtermini gewinnen kann, siehe auch RUNTE ET AL. 2003, in diesem Band, und LENZ ET AL. 2002. Aus jedem Wort-Topic (d. h. zu jedem Lexem) wird ein Glossareintrag erzeugt, der verschiedene Links in das Korpus beinhaltet, z. B. Links zu Definitionen des Terminus. Diese werden aus den Topic-Ankern erzeugt. Umgekehrt wird von Termverwendungsinstanzen im Korpus auf Glossareinträge verlinkt.

Obwohl der Ausdruck „Topic Map“ eine grafische und räumliche Dimension impliziert und zur Veranschaulichung des Konzepts und der Vorteile von Topic Maps oft Beispiel-Visualisierungen herangezogen werden, beinhaltet der Standard selbst keine grafischen Komponenten. Da XTM XML-basiert ist, bietet es sich an, XSLT zu nutzen, um XTM-Daten in eine hochwertige grafische Präsentation zu transformieren, die tatsächlich das Kartenhafte an Topic Maps erkennen lässt.

HTML bietet für eine angemessene Umsetzung komplexer XTM-Strukturen nicht genügend Mittel. Abhilfe schafft der seit 2001 als W₃C-Recommendation vorliegende, ebenfalls XML-basierte Standard Scalable Vector Graphics (SVG). SVG-Dokumente lassen sich über ein Plugin in (X)HTML-Dokumente einbetten und bieten eine Javascript-Schnittstelle, so dass auch anspruchsvolle interaktive Karten mit verhältnismäßig wenig Aufwand automatisch aus XTM-Daten erstellt werden können, die den Anspruch von Topic Maps, Landkarten für das semantische Netz zu sein, erfüllen.

Eine relativ einfache Visualisierung jedes Lexems und seiner lexikalischen und konzeptuellen Relationen, die den jeweiligen Glossareintrag ergänzt und in diesen eingebettet wird, erzeugen wir auf diese Weise automatisch aus der Topic Map (Abbildung 3).



Abbildung 3: SVG-Visualisierung eines Lexems, wie sie einem Hypertextnutzer präsentiert wird. Lexeme und Konzepte werden verschiedenfarbig dargestellt, zu jedem Konzept werden dessen Lexeme angegeben. Durch Anklicken eines anderen Lexems kann der Nutzer zu dessen grafischer Darstellung navigieren.

8 Wortnetze als Topics Maps?

Nachdem wir am Beispiel unseres terminologischen Netzes beschrieben haben, wie sich Wortnetze als Topic Maps modellieren lassen, möchten wir an dieser Stelle einen Ausblick geben, welche Vorteile sich für die WordNet-Community ergeben könnten, wenn diese Modellierung für verschiedene Wortnetze verwendet würde.

Eine interessante Eigenschaft von Topic Maps liegt in der Möglichkeit, festzulegen, dass ein Topic mit einem anderen Topic identisch sein soll. Dies geschieht über einen URI-Verweis. Die beiden Topics müssen nicht zwangsläufig in der derselben Topic Map vorliegen, sondern es können auch „öffentliche“ Topics deklariert werden, die sogenannten Published Subject Indicators (PSIs), auf die allgemein Bezug genommen werden kann und soll.

Solche öffentlichen Topics werden von Standardisierungsorganisationen, Firmen und anderen Organisationen herausgegeben. Die Organisation OASIS (Organization for the Advance-

ment of Structured Information Standards) hat z. B. PSIs für Länder und Sprachen der Welt herausgegeben. Wenn nun zwei verschiedene Topic Maps z. B. auf das öffentliche Topic für die Sprache Deutsch referieren, dann ist klar, dass in jedem Fall dasselbe gemeint ist, auch wenn es in der einen Topic Map *allemand* und in der anderen *german* heißt.

Für das Vereinigen von Topic Maps (*Merging*) gibt der XTM-Standard explizite Regeln an, die Topic-Map-konforme Software implementieren muss. Eine dieser Regeln bewirkt, dass zwei Topics, die auf dasselbe öffentliche Topic referieren, miteinander vereinigt werden. Das so entstandene neue Topic enthält alle Eigenschaften der beiden ursprünglichen Topics. Im Beispiel würde das vereinigte Topic für die deutsche Sprache also beide Namen tragen und die Assoziationen zu anderen Topics aus beiden Topic Maps übernehmen.

Von diesem Mechanismus könnte die WordNet-Community profitieren. Man könnte PSIs

Wortnetze als XML Topic Maps

für verschiedene WordNet-Relationen veröffentlichten, ebenso für *Lexem* und *Synset*. Dann können dieselben Relationen in konkreten Wortnetzen unterschiedlich benannt werden, durch Verweis auf den PSI wird ihre Identität sichergestellt, und gleiche Relationen würden in einem möglichen Merging-Prozess miteinander vereinigt. Durch Wahl entsprechender Skopen kann sichergestellt werden, dass die Herkunftsinformationen und ursprünglichen Namen erhalten bleiben. Wenn sogar für jedes Konzept ein PSI eingeführt wird – entsprechend dem Interlingual Index, wie er in EuroWordNet (VOSSEN 1998) verwendet wird – dann würden auch die Konzepte beim Merging vereinigt. Der Merging-Prozess an sich kann durch standardkonforme Topic-Map-Software durchgeführt werden.

Für die Hyperonymie definiert der XTM-Standard bereits ein öffentliches Topic. Es trägt den Namen *superclass-subclass relationship* im Skopus der englischen Sprache, aber wir sind frei, ein neues öffentliches Topic mit dem Namen *Hyperonymie* festzulegen und seine Identität mit dem superclass-subclass-Topic durch einen URI-Verweis auszudrücken.

9 Ausblick

Wir haben anhand des im HyTex-Projekts verwendeten terminologischen Netzes exemplarisch aufgezeigt, wie ein Wortnetz mit XML Topic Maps repräsentiert und mit texttechnologischen Methoden weiterverarbeitet werden kann. Im letzten Abschnitt haben wir gezeigt, dass eine Topic-Map-Repräsentation sehr gut dazu geeignet ist, verschiedene Wortnetze oder Teile davon miteinander zu vereinigen. Ebenso ist es möglich, eine Topic Map in ein Datenbankformat zu transformieren, das einen schnellen Zugriff ermöglicht.

Auf der anderen Seite findet RDF(S) im Zuge des voranschreitenden „Semantic Web“ derzeit weite Verbreitung. Zur Repräsentation von Ontologien ist der auf RDFS aufsetzende Standard OWL weitaus ausgereifter als ein entsprechender

Standard für Topic Maps, der derzeit entworfen wird.

Um diesem Dilemma zu entkommen und die Vorteile beider Repräsentationsformalisten nutzen zu können, könnte man Wortnetze zunächst in Topic Maps repräsentieren, um sie anschließend (automatisch) nach RDFS zu transformieren. Die so bereitgestellte lexikalische Ressource könnte dann im „Semantic Web“ als Web-Service zur Verfügung gestellt werden (LEMNITZER & KUNZE 2003).

Anmerkungen

- ¹ HyTex ist ein DFG-gefördertes Projekt, das seit 2002 an der Universität Dortmund unter der Leitung von Prof. Dr. Angelika Storrer durchgeführt wird. Informationen zu den Projektarbeiten finden sich unter <http://www.hytext.info/> [Zugriff April 2004].

Literatur

- BECKWITH, R.; MILLER, G. A.; TENGI, R. (1993). „Design and implementation of the WordNet lexical database and searching software.“ In: MILLER, G. A. ET AL. (1990). Five Papers on WordNet. Technical Report 43, Princeton University, Cognitive Science Laboratory, <ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf> [Zugriff April 2004, Erstveröffentlichung in: Journal of Lexicography 3(4) (1990), 235-312].
- CHRISTODOULAKIS, D.N.; KUNZE, C.; LEMNITZER, L. (HRSG.) (2002). Proceedings of the Workshop on Wordnet Structures and Standardizations, and how these Affect Wordnet Applications and Evaluation. 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, 28th May 2002.
- DEROSE, S.; MALER, E.; ORCHARD, D. (2001). XML Linking Language (XLink) Version 1.0. World Wide Web Consortium (W3C) Recommendation, June 2001, <http://www.w3.org/TR/2000/xlink/> [Zugriff April 2004].

- FISCHER, D. H. (1997). "Formal Redundancy and Consistency Checking Rules for the Lexical Database WordNet." VOSSEN, P. ET AL. (Hrsg.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Building of Lexical-Semantic Resources for NLP Applications, 22-31.
- GUPTA, P. (2002). "Approaches to Checking Subsumption in GermaNet." In: CHRISTODOULAKIS, KUNZE & LEMNITZER (2002), 8-13.
- INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO) (2000). ISO/IEC 13250:2000 Document Description and Processing Languages – Topic Maps. Geneva, Switzerland: ISO, <http://www.yr2.doe.gov/sgml/sc34/document/0129.pdf> [Zugriff April 2004].
- KUNZE, C.; LEMNITZER, L. (2002). "Standardizing WordNet in a web-compliant format: The case of GermaNet." In: CHRISTODOULAKIS, KUNZE & LEMNITZER (2002), 24-29.
- LASSILA, O.; SWICK, R. R. (1999). Resource Description Framework (RDF) Model and Syntax Specification. World Wide Web Consortium (W3C) Recommendation, February 1999, <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/> [Zugriff April 2004].
- LEMNITZER, L.; KUNZE, C. (2003). Integrating Wordnets into the Resource Description Framework, 2003, Universität Tübingen, Seminar für Sprachwissenschaft, http://www.sfs.uni-tuebingen.de/~lothar/publ/GermaNET_RDF.pdf [Zugriff April 2004].
- LENZ, E. A.; BEISSWENGER, M.; STORRER, A. (2002). „Hypertextualisierung mit Topic Maps – ein Ansatz zur Unterstützung des Textverständnisses bei der selektiven Rezeption von Fachtexten." In: TOLKSDORF, R.; ECKSTEIN, R. (Hrsg.) (2002). Proceedings Workshop XML-Technologien für das SemanticWeb (XSW 2002), Berlin, Juni 2003. Bonn: Köllen Verlag [= GI-Edition - Lecture Notes in Informatics (LNI), P-14], 151-159.
- MILES-BOARD, T.; KAMPA, S.; CARR, L.; HALL, W. (2001). "Hypertext in the Semantic Web." In: Proceedings 12th ACM Conference on Hypertext and Hypermedia (HT '01). Aarhus, Denmark, 237-238.
- PAVELEK, T.; PALA, K. (2002). "WordNet Standardization from a Practical Point of View." In: CHRISTODOULAKIS, KUNZE & LEMNITZER (2002), 30-34.
- PEPPER, S.; MOORE, G. (2001). XML Topic Maps (XTM) 1.0. TopicMaps.Org Consortium Specification, March 2001, <http://www.topicmaps.org/xtm/1.0/> [Zugriff April 2004].
- RATH, H. H. (200). "Making Topic Maps more Colourful." In: Proceedings of XML Europe 2000 Conference, Berlin, Mai 2001, <http://www.gca.org/papers/xml/europe2000/pdfs/s29-01.pdf> [Zugriff April 2004].
- RATH, H. H. (2002). "GPS des Web. XML Topic Maps: Themenkarten im Web." In: iX, Magazin für Professionelle Informationstechnik, Heft 6 (2002), 115-122.
- RUNTE, M.; BEISSWENGER, M.; STORRER, A. (2003). „Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet." In diesem Band, 113-125.
- VOSSEN, P. (1998). "Introduction to EuroWordNet." In: VOSSEN, P. (Hrsg.) (1999). EuroWordNet: A Multilingual Database with Lexical-semantic Networks. Dordrecht: Kluwer Academic Publishers, 73-89.