Hypertextualisierung mit Topic Maps – ein Ansatz zur Unterstützung des Textverständnisses bei der selektiven Rezeption von Fachtexten

Eva Anna Lenz, Michael Beißwenger, Angelika Storrer

{lenz,beisswenger,storrer}@hytex.info

Abstract: Der Kurzbeitrag berichtet über ein Projekt „Hypertextualisierung auf textgrammatischer Grundlage“ (HyTex), in dem erforscht wird, wie sich linear organisierte Dokumente mit semiautomatischen Methoden auf der Grundlage von textgrammatischem Markup und der linguistisch motivierten Modellierung terminologischen Wissens in delinearisierte Hyperdokumente überführen lassen. Ziel ist es, eine Sammlung von Fachtexten so in einen Hypertext zu überführen, dass terminologiebedingte Verständnisschwierigkeiten beim Lesen durch entsprechende Linkangebote aufgelöst werden, so dass die Fachtexte auch von Semi-Experten der Domäne selektiv gelesen werden können. Der Schwerpunkt des Beitrags liegt auf der Modellierung terminologischen Wissens mit XML Topic Maps und dessen Stellenwert für die automatische Erzeugung von Hyperlinks.

1 Projektrahmen

Als Hypertextualisierung bezeichnet man die Aufgabe, sequenziell organisierte Dokumente für die Publikation in einem Hypertextsystem, z.B. dem World Wide Web, aufzubereiten. Für diese Aufgabe braucht man Konversionstools auf der technischen Seite. Auf der konzeptionellen Seite benötigt man Prinzipien und Strategien (a) zur Zerlegung von Dokumenten in Hypertextknoten (Segmentierung), und (b) zur Verknüpfung der Hypertextknoten durch Hyperlinks (Re-Konnexion).

Das im folgenden skizzierte Projekt „Hypertextualisierung auf textgrammatischer Grundlage“ (HyTex, vgl. http://www.hytex.info/), das als Teilprojekt der DFG-finanzierten Forschergruppe „Texttechnologische Informationsmodellierung“ an der Universität Dortmund durchgeführt wird (bis 31.3.2002 am Institut für deutsche Sprache in Mannheim), ist interdisziplinär angelegt, mit einem sprachwissenschaftlichen Schwerpunkt. Es beschäftigt sich insbesondere mit der konzeptionellen Seite der Hypertextualisierung. Parallel und ergänzend zu statistisch basierten und KI-orientierten wissensbasierten Ansätzen möchten wir in unserem textgrammatisch basierten Ansatz Kohärenzstrukturen in Dokumenten (semi-)automatisch durch Markup annotieren.[1] Dieses Markup nutzen wir für die Gene-

[1] Kohärenzstrukturen lassen sich grob umreißen als semantische Zusammenhänge zwischen Textteilen, die von einem Rezipienten auf der Grundlage seines Sprach- und Weltwissens rekonstruiert werden müssen.

rierung von Hypertextsichten, die den Nutzern beim Browsen eines Hypernetzes genau diejenigen Wissensvoraussetzungen anbieten, die zum Verständnis des aktuell rezipierten Inhalts benötigt werden (*Linking nach Wissensvoraussetzungen).*

Anwendungsgebiete dieser Strategie sind Szenarien, in denen ein Leser Semi-Experte für eine Fachtextdomäne ist und Fachtexte selektiv liest. Ein Semi-Experte kennt nicht alle (Fach-)Termini der Domäne und erkennt sie teilweise nicht als solche. Selbst wenn er einen Ausdruck als Terminus identifiziert, stellt sich für ihn die Frage, ob eine Definition in dem aktuell rezipierten Fachtext zu finden ist, und, falls ja, an welcher Stelle. Solche Szenarien finden sich z.B. im Kontext interdisziplinärer Forschung, Studium und Ausbildung, Journalismus, und Fachlexikographie. Die Strategie „Linking nach Wissensvoraussetzungen“ erlaubt die Generierung von Hypertext-Sichten, die den Rezipienten beim selektiven Lesen eines Fachtextes so unterstützen, dass terminologiebedingte Verständnisprobleme durch entsprechende Linkangebote gelöst werden können.

Die in HyTex zu entwickelnden automatischen Strategien zur Segmentierung und zum Linking werden erprobt an einer Sammlung von Dokumenten zur Fachdomäne Texttechnologie, d.h. zunächst an einem abgeschlossenen Textkorpus. Die Übertragbarkeit auf offene Dokumentenmengen soll in der zweiten Projektphase erprobt werden. Das zu diesem Zweck aufgebaute Fachtextkorpus enthält in seinem Kernbestand wissenschaftliche Abhandlungen zum Thema, sowie wichtige Dokumente wie XML-Spezifikationen, Glossare, FAQs etc. Weiterhin werden diskursive Dokumente wie z.B. Beiträge aus Foren, Mailinglisten und Chats zum Thema berücksichtigt.

Das Projektziel besteht darin, zu untersuchen, wie das Linking nach Wissensvoraussetzungen auf der Basis eines linguistisch motivierten Ansatzes geschehen kann. Auf der einen Seite erleichtern Techniken und Standards des „Semantic Web“ unsere Forschungsarbeit. Auf der anderen Seite zielen wir darauf ab, mit konzeptionellen Ansätzen das Semantic Web für die menschliche Informationsverarbeitung (im Gegensatz zur Informationsverarbeitung durch maschinelle Agenten) in der genannten speziellen Nutzungssituation durch das Linking nach Wissensvoraussetzungen besser erschließbar zu machen.

2 Skizze der Projektarchitektur

Die Strategien zur Hypertextualisierung des Korpus operieren über Repräsentationen von Wissen auf dreierlei Ebenen, die in der Veranschaulichung der Projektarchitektur in Abbildung 1 (von unten nach oben) dargestellt sind:

Ebene des Fachtextkorpus: Das in den Dokumenten sprachlich manifeste Wissen über die Vernetztheit der Inhalte ist repräsentiert als textgrammatisches und linguistisches Markup. Annotiert werden insbesondere sprachliche Mittel der Textverknüpfung, Definitionen von Fachtermini und Termverwendungsinstanzen.

Ebene des Modells der Fachtextdomäne: Terminologisches Wissen über die Domäne repräsentieren wir mit XML Topic Maps (XTM, [PM01]) unter Verwendung der von der lexikalischen Datenbank WordNet benutzen Strukturierung [Fel98b].

Ebene der Benutzermodellierung: Annahmen über das Vorwissen bestimmter Nutzertypen. Dieses repräsentieren wir zunächst als (statische) Nutzerprofile, in einer späteren Projektphase auch in Form von Nutzungsprotokollen, aus denen sich dynamisch die Wissensvoraussetzungen erschließen lassen, die ein Nutzer auf seinem individuell gewählten Leseweg bereits erworben hat.

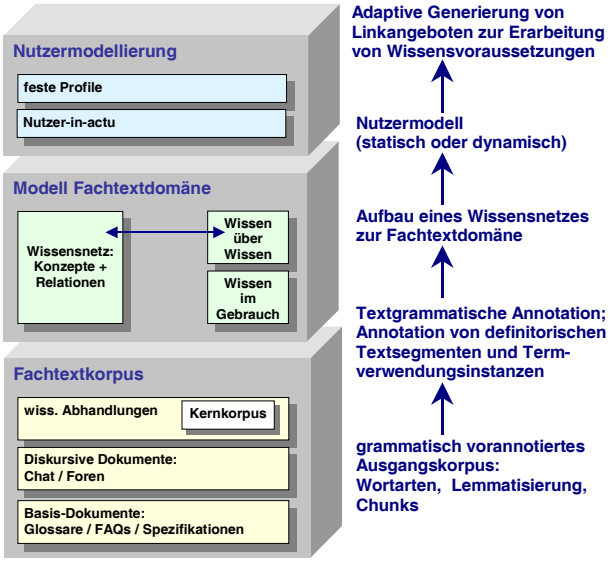

Abbildung 1: Die HyTex-Projektarchitektur.

Im Rahmen der Strategie „Linking nach Wissensvoraussetzungen“ verfolgen wir zwei Ziele: (a) Hypertext-Sichten auf die Korpustexte zu generieren, deren einzelne Hypertext-Knoten kohäsiv geschlossen sind[2], und (b) Wissensvoraussetzungen, die der Verwendung von Termini in den Dokumenten zu Grunde liegen, über Linkangebote zu den entsprechenden Definitionen für einen selektiv zugreifenden Benutzer rekonstruierbar zu machen.

Auf (a) gehen wir in diesem Kurzbeitrag nicht näher ein. Für (b) unterscheiden wir zwischen drei Arten von Wissensvoraussetzungen:

- Eine *intratextuelle Wissensvoraussetzung* liegt vor, wenn ein Ausdruck als Terminus im Sinne einer Definition verwendet wird, die der Autor im Vortext explizit eingeführt hat.
- Eine *extratextuelle Wissensvoraussetzung* liegt vor, wenn der Autor einen Terminus im Sinne der Definition eines anderen Autors (in einem anderen Dokument) verwendet.

[2]Ein Hypertext-Knoten ist kohäsiv geschlossen, wenn es keine aus dem Textteil hinausweisenden sprachlich manifesten Bezüge – wie z.B. „siehe Kapitel 5“ – enthält, es sei denn, diese sind (z.B. über Linkangebote) rekonstruierbar.

- Eine *domänenspezifische Wissensvoraussetzung* liegt vor, wenn ein Terminus ohne genaue Angabe einer Definitionsstelle im Sinne einer in der betreffenden Fachsprache eingespielten Festlegung verwendet wird.

Die Generierung von Linkangeboten zur Rekonstruktion solcher terminologiebedingten Wissensvoraussetzungen erfolgt auf der Grundlage einer teilautomatischen Annotation sowohl von Textsegmenten, in deren Rahmen Termini definitorisch eingeführt werden (*definitorische Textsegmente*) als auch von Verwendungen der Termini (*Termverwendungsinstanzen*). Sowohl die (teilautomatischen) Verfahren zur Auffindung und Annotation als auch das Linking für intra- und extratextuelle Wissensvoraussetzungen werden in [BLS] beschrieben; wir konzentrieren uns in diesem Beitrag auf die Modellierung domänenspezifischer Wissensvoraussetzungen mit XTM.

Für die Implementierung unserer Architektur nutzen wir XML als Austauschformat für alle Komponenten: für das Markup des Textkorpus, für die Topic Map (XTM-Syntax) und für die Benutzermodellierung. Dies erlaubt es uns, Dokumente im Web-Kontext zu erstellen und wiederzuverwenden und vorhandene Werkzeuge zu nutzen, z.B. zur Validierung. Zur Erzeugung des späteren Hypertextes aus der textgrammatischen Annotation und der Topic Map benutzen wir XSLT. Der XSLT-Code wird automatisch aus deklarativen Regeln erzeugt, die leichter änderbar sind als Programmcode und die ebenfalls in XML repräsentiert sind; zu Einzelheiten siehe [LS02].

Im HyTex-Projekt kombinieren wir bereits bewährte Ansätze der Hypertextualisierung auf der Basis von Textauszeichnungen, einer Ontologie und einem Benutzermodell mit linguistischen Methoden (textgrammatisches Markup, Auszeichnung von definitorischen Textsegmenten und Termverwendungsinstanzen, WordNet) und Techniken des Semantic Web (XML, Topic Maps, XSLT). Mit dieser spezifischen Kombination verfolgen wir das Ziel der Herstellung eines Hypertextes für das oben beschriebene Anwendungsszenario des Semi-Experten mit selektiver Lesestrategie.

3 Modellierung terminologischen Wissens mit Topic Maps und dessen Nutzung zum Linking

Für den Fall, dass ein Terminus relativ zu seinem Gebrauch in der Fachsprache verwendet wird (domänenspezifische Wissensvoraussetzung), ist es für den selektiven Leser eines Fachtextes sehr hilfreich, wenn ihm einerseits die Information gegeben wird, dass es sich überhaupt um einen Terminus handelt, und er andererseits weitere Informationen über den Terminus erhalten kann. Solche Informationen möchten wir durch Generierung entsprechender Hypertext-Sichten bereitstellen (Abbildung 2).

Im folgenden skizzieren wir, wie terminologisches Wissen über die Fachtextdomäne als lexikalisches Netz modelliert werden kann (3.1) und wie diese Modellierung mit XTM implementiert wird (3.2). Abschließend zeigen wir exemplarisch, wie die Topic Map zur Generierung domänenspezifischer Linkangebote genutzt und aus Teilen der Topic Map ein erweitertes Glossar erzeugt werden kann (3.3).

3.1 Strukturierung des Wissensnetzes auf der Basis von WordNet

Die grundlegenden Konzepte und Relationen, die zwischen den Termini der Fachtextdomäne bestehen, modellieren wir unter Rückgriff auf den psycholinguistisch motivierten Ansatz der lexikalischen Datenbank WordNet [Fel98b], der ursprünglich für den englischen Wortschatz entwickelt, inzwischen aber im Projekt EuroWordNet für andere europäische Sprachen erweitert wurde. Im Folgenden skizzieren wir nur die Aspekte der Modellierung, die für unseren Anwendungsbereich relevant sind; eine detaillierte Beschreibung des Ansatzes und deren Anwendungsmöglichkeiten in Sprachtechnologie und Information Retrieval finden sich im Sammelband von [Fel98b].

Grundeinheiten von WordNet sind Konzepte (Begriffe), die durch sog. „synsets“ repräsentiert sind. Ein Synset besteht aus den (graphisch repräsentierten) Wörtern, in unserem Fall terminologischen Ausdrücken, die das enstprechende Konzept in einer Einzelsprache lexikalisieren. Die Wortformen „Hyperlink“ und „Link“ bilden also z.B. in der deutschen Hypertext-Terminologie ein als „synset“ repräsentiertes Konzept. Wortformen und Konzepte werden durch Relationen zu einem lexikalischen Netz, dem „WordNet“, verknüpft. Dabei wird zwischen semantisch-konzeptionellen und lexikalischen Relationen unterschieden (vgl. [Fel98a], S.8f).

- *Semantisch-konzeptionelle Relationen* sind Relationen zwischen den Synsets, also den Konzepten. Wichtige konzeptionelle Relationen für unsere Zwecke sind die Hyponymie (`ist_Unterbegriff_von`) mit der Umkehrrelation Hyperonymie, und die Meronymie (`ist_ein_Teil_von`) mit der Umkehrrelation Holonymie.
- *Lexikalische Relationen* bestehen zwischen Wortformen. Dazu zählen Synonymie und Antonymie (semantisches Gegenteil, z.B. warm – kalt), wobei die Synonymie-Relation insofern speziell ist, als dass sie durch das Synset ausgedrückt wird.

Beide Typen von Relationen sind binär und zum Teil gerichtet. Wir orientieren uns an den im Zuge des für EuroWordNet entwickelten und in [Vos98] beschriebenen Relationeninventars und erweitern dieses um weitere Relationen, die für unsere Zwecke sinnvoll sind, wie beispielsweise domänenspezifische Relationen (z.B. `implementiert_als`), oder die Relation `ist_Wissensvoraussetzung_für`, die über den bestehenden semantisch-konzeptionellen Relationen definiert werden kann.

Das WordNet-Modell ermöglicht aufgrund der Trennung zwischen lexikalischen und semantischen Relationen eine feinere sprachliche Modellierung als Ontologien, die sich primär auf die semantischen ISA- und Teil/Ganzes-Relationen stützen. Diese Unterscheidung hat sich in verschiedenen Projekten der Sprachverarbeitung und im Information Retrieval bewährt [Fel98b]. Die Datenhaltung ist ökonomisch, da bei den semantisch-konzeptionellen Relationen die Synsets (als atomare Konstituenten des semantischen Netzes) miteinander verbunden werden und gerade nicht die einzelnen, den jeweiligen Synsets angehörenden Wörter. Die Relationierung einzelner Wörter geschieht vielmehr auf der Ebene der lexikalischen Relationen. Beispielsweise ist es möglich, eine Abkürzung für ein Wort anders zu behandeln als dessen Synonyme.

Die in WordNet angelegte Unterscheidung zwischen Synsets und lexikalischen Einheiten (Wörtern) erleichtert es auch, Terminologien verschiedener natürlicher Sprachen aufeinander zu beziehen (EuroWordNet, [Vos98]).

3.2 Repräsentation von WordNet als Topic Map

Entsprechend der in WordNet vorhandenen grundlegenden Unterscheidung zwischen Wörtern und Konzepten (Synsets) enthält unsere Topic-Map Repräsentation zwei Arten von *Topics*:

Wort-Topic: Für jeden in einer Fachdomäne terminologisierten Ausdruck – d.h. jede lexikalische Einheit, die einen Terminus darstellt – deklarieren wir ein Topic in der Topic Map. Dies bedeutet, dass es z.B. in der Domäne der Hypertexttheorie für „Link“ und „Hyperlink“ je ein Topic gibt. Würden wir auch die Domäne der Künstlichen Intelligenz hinzunehmen, so gäbe es ein weiteres Topic für „Link“ in der Theorie der semantischen Netze. Die Wortform wird durch den *Basisnamen* (Base Name) des Topics repräsentiert.
Die Wort-Topics werden anhand der im Gesamtkorpus vorhandenen Annotationen von Definitionen automatisch erzeugt.

Konzept-Topic: Für jedes Konzept (Synset) führen wir ebenfalls ein Topic ein. Die Wort-Topics aller darin enthaltenen terminologisierten Ausdrücke (Wortformen) werden mit dem jeweiligen Konzept durch eine *Assoziation* (association) vom Typ `lexikalisiert` manuell verbunden.

Lexikalische und semantisch-konzeptionelle Relationen lassen sich nun auf getypte Assoziationen zwischen den Wort-Topics bzw. den Konzept-Topics abbilden. *Assoziationstypen* (association types) für lexikalische Relationen sind z.B. `ist_antonym_zu` und `ist_Abkürzung_für`. Da diese in Topic Maps selbst wieder Topics sind, können wir alle lexikalischen Relationen als Instanzen des *Topic-Typ* (topic type) `lexikalische_Relation` deklarieren, alle semantisch-konzeptionellen Relationen entsprechend als Instanzen des Topic-Typs `semantische_Relation`. Wie üblich drücken wir die Richtung einer Relation durch *semantische Rollen* (association role types) aus. Auch die Relationen werden explizit modelliert.

Jedes Vorkommen eines Terminus in einem Dokument wird mit dem entsprechenden Wort-Topic durch einen *Topic-Anker* (occurrence) mit der *Ankerrolle* (occurrence role type) `Termverwendung` verbunden. Anhand der Liste der Wort-Topics werden die Vorkommen im Korpus automatisch aufgefunden und mit ihnen durch Topic-Anker verbunden. Textstellen, die einen Terminus definieren oder auf andere Weise zu seinem Verständnis beitragen (z.B. in einem FAQ), werden manuell durch Topic-Anker verschiedener Ankerrollen mit Wort-Topics verbunden.

In XTM ist es möglich, dass eine Ressource direkt in einen Topic-Anker eingebettet wird. Wir verwenden diese Möglichkeit, um eine allgemeine Definition für einen Terminus direkt in die Topic Map mit aufzunehmen.

Wir benutzen das Topic-Map Konzept des *Skopus* (scope) zur thematischen Filterung. Wir weisen allen Charakteristika eines Topics (d.h. seinen Namen, Topic-Ankern und semantischen Rollen) manuell einen thematischen Skopus zu, z.B. `Hypertext` oder `texttechnologische_Standards`. Die Charakteristika aller Topics, die innerhalb der Topic Map als (WordNet-)Typen dienen (Topic-Typen, Assoziationstypen, semantische Rollen, Anker-Rollen, und Skopi), erhalten den speziellen Skopus `WordNet`. So können sie von den Topics der Domäne unterschieden werden.

3.3 Nutzung der Topic Map zum Linking

Das nach den Prinzipien von WordNet strukturierte und als Topic Map repräsentierte terminologische Wissen über das Fachgebiet „Texttechnologie“ kann nun genutzt werden, um einem Leser zur Rekonstruktion domänenspezifischer Wissensvoraussetzungen – neben den intratextuellen und den extratextuellen – entsprechende Linkangebote zu machen. Diese Links führen zu einem erweiterten Glossar, das in Abbildung 2 exemplarisch dargestellt ist. Es enthält zwei Arten von Informationen:

- *Informationen über den Terminus selbst.* Dazu gehören eine allgemeine Definition des Terminus, die auch Verweise zu anderen Teilen des Glossars enthalten kann, sowie Linkangebote, die zu Textstellen des Korpus führen, die zum Verständnis des Terminus beitragen. Sie können z.B. zu Definitionen des Terminus bei anderen Autoren führen, ggf. gewichtet nach der Einschlägigkeit des Autors und/oder dem Typ der Definition. Zudem wird auf Vorkommen des Terminus in bestimmten Textsorten verwiesen, z.B. auf FAQs.
- *Informationen über die Beziehungen des Terminus zu anderen Termini.* Zu jedem Terminus werden seine Relationen zu anderen Termini angegeben, u.a. Ober- und Unterbegriffe und Synonyme. Dabei ist es auch möglich, zu den jeweiligen „verwandten“ Termini im erweiterten Glossar zu gelangen und von dort aus Informationen über jene Termini zu erhalten. Allein diese Informationen können sehr hilfreich sein, um einen Terminus einordnen zu können.

Das erweiterte Glossar wird automatisch aus der Topic Map gewonnen. Dazu wird zu jedem Wort-Topic – d.h. zu jedem terminologisierten Ausdruck – ein Hypertext-Knoten für einen Glossareintrag erzeugt. Anhand der Topic-Anker mit der Anker-Rolle `Termverwendung` werden die domänenspezifischen Links von den Korpus-Dokumenten zu den jeweiligen Glossareinträgen erstellt. Der Text für die allgemeine Definition des Terminus wird aus der eingebetteten Ressource für das Wort-Topic übernommen. Die Links zu Textstellen anderer Dokumente entsprechen den Topic-Ankern des Topics mit verschiedenen Anker-Rollen (z.B. `bekannte_Definition` für Links zu in der Fach-Community einschlägigen Definitionen, oder `FAQ`). Synonyme werden durch die Verbindung der Wort-Topics mit ihrem Konzept über die `lexikalisiert`-Assoziation aufgefunden. Die Links zu „verwandten Begriffen“ werden aus den getypten Assoziationen (zwischen Wort-Topics bzw. zugehörigen Konzept-Topics) erzeugt.

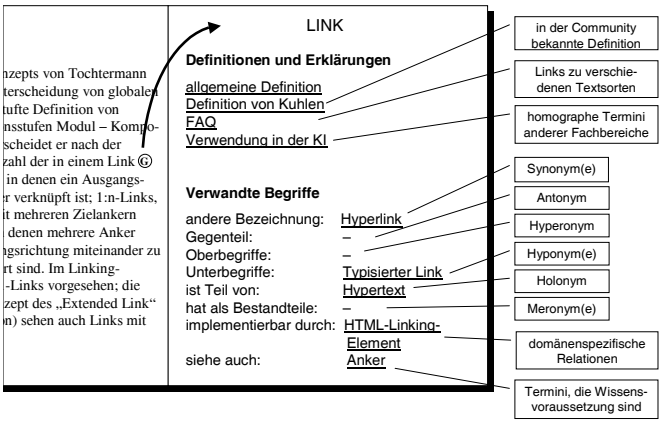

Abbildung 2: Exemplarische Darstellung eines Glossareintrags für den Terminus „Link“.

Wir gehen davon aus, dass sich verschiedene Benutzer in unterschiedlichen Nutzungssituationen für unterschiedliche Themen interessieren. Durch die Modellierung thematischer Skopi wird es ermöglicht, Themen, die den Benutzer nicht interessieren, auszublenden. Damit können auch alle Arten von Links (von Text zu Text, innerhalb des Glossars, von Text zu Glossar, und von Glossar zu Text) gefiltert und damit ggf. ausgeblendet werden.

Literaturverzeichnis

[BLS] Michael Beißwenger, Eva Anna Lenz, and Angelika Storrer. Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen. In Vorbereitung.

[Fel98a] Christiane Fellbaum. Introduction. In Christiane Fellbaum, editor, *WORDNET: an electronic lexical database*, pages 1 – 19. MIT Press Cambridge, Massachusetts, London, England, 1998.

[Fel98b] Christiane Fellbaum, editor. *WORDNET: an electronic lexical database*. MIT Press Cambridge, Massachusetts, London, England, 1998.

[LS02] Eva Anna Lenz and Angelika Storrer. Converting a corpus into a hypertext: An approach using XML topic maps and XSLT. In *LREC 2002: Third international conference on language resources and evaluation*, 2002. Im Druck.

[PM01] Steve Pepper and Graham Moore, editors. *XML Topic Maps (XTM) 1.0. TopicMaps.Org Specification*, 2001. http://www.topicmaps.org/xtm/1.0/.

[Vos98] Piek Vossen. Introduction to EuroWordNet. In Piek Vossen, editor, *EuroWordNet: a multilingual database with lexical semantic networks*, pages 73 – 89. Kluwer Academic Publishers, Dordrecht / Boston / London, 1998.