

Mehrsprachige Thesauri

Simon Knoll

simon.knoll@student.uibk.ac.at

Zusammenfassung Die vorliegende Seminararbeit behandelt verschiedene Methoden zur Erstellung eines mehrsprachigen Thesaurus. Dabei werden die traditionellen Verfahren, welche auf die Analyse und Verarbeitung von mehrsprachigen Textkörpern und deren Verfügbarkeit aufbauen, mittels eines neuen, auf Wikipedia basierenden Ansatzes, verglichen. Bei dem neuartigen Ansatz steht Wikipedias Linkstruktur im Vordergrund. Aufgrund dieser werden, für ein mehrsprachiges Thesaurus nützliche Informationen gewonnen. Der Vorteil dieses Ansatzes besteht darin, dass nicht Textpassagen verarbeitet werden, sondern die verschiedenen Wikipedia-Links die nötigen Informationen enthalten.

1 Einleitung

Anfang 2001 startete Wikipedia als englischsprachiges Projekt und schon Ende 2001 war es in 18 verschiedenen Sprachen erreichbar. Seitdem haben rund 285.000 registrierte und eine unbekannte Anzahl an nicht registrierter Benutzer ihren Beitrag an Wikipedia geleistet. Alleine für die deutschsprachige Wikipedia sind regelmäßig mehr als 7.000 Autoren tätig. Solch ein Pool an Informationen ist wie geschaffen für die Erprobung verschiedenster Mining-Algorithmen. Unter Mining, was zu Deutsch soviel wie schürfen bedeutet, ist die Informationsgewinnung durch Analyse von Daten zu verstehen. Unter Betrachtung der Erstellung eines mehrsprachigen Thesaurus (ein Vokabular dessen Begriffe durch Relationen verbunden sind), wird ersichtlich, dass dabei auch Daten analysiert und Informationen daraus gewonnen werden. Die zu Grunde liegenden Daten sind die Sprachen aus denen ein mehrsprachiges Thesaurus erstellt werden soll und die gewonnenen Informationen sind die Relationen der Wörter untereinander.

Laut der offiziellen Statistik von 2006 [wika], enthalten alle verschiedenen Sprachversionen der Wikipedia insgesamt 5 Mio. Artikel, 74.4 Mio. interne Links und 21.3 Mio. Links zu anderssprachigen Wikipedias. Aufgrund dieser Tatsache, liegt es nahe Mining-Algorithmen auf Wikipedia anzuwenden, um mit den gewonnenen Informationen ein mehrsprachiges Thesaurus zu erstellen. Diese Seminararbeit wird sich vordergründig mit den Wikipedia-gestützten Techniken zur Thesaurus-Generierung beschäftigen.

In Kapitel 2 werden zunächst einige grundlegende Begriffe geklärt. Kapitel 3 beschäftigt sich mit den traditionellen Verfahren der multilingualen Thesauri-Generierung. In Kapitel 4 werden die Grundgedanken, der Verwendung von

Wikipedia als Quelle für multilinguale Thesauri erklärt und die Ansätze von Erdmann [ENHN08a] besprochen. Im darauffolgenden 6ten Kapitel werden die verschiedenen Möglichkeiten zur Verbesserung behandelt, welche von Kinzler in [Kin08] vorgeschlagen werden. Abgeschlossen wird die Seminararbeit mit dem letzten Kapitel welches ein Fazit bildet und einen Ausblick auf das zukünftige Potenzial dieser Verfahren gibt.

2 Begriffserklärung

Zu Beginn werden einige Begriffe erklärt, welche im weiteren Verlauf der Seminararbeit noch öfters erwähnt und gebraucht werden.

Thesaurus Ein Thesaurus ist ähnlich wie ein Wörterbuch, eine Sammlung von Begriffen zu einem bestimmten Themenbereich. Ein Thesaurus enthält jedoch mehr als nur die korrekte Schreibweise, Informationen darüber, ob sich ein Begriff verallgemeinern (Oberbegriffe), spezialisieren (Unterbegriffe) oder anders ausdrücken (Synonyme) lässt.

Textkorpus Ein Textkorpus ist eine Sammlung von ausgewählten und organisierten Texten, welche Gegenstand sprachlinguistischer Analysen ist. Im Kontext dieser Seminararbeit wird ein Korpus durch eine Sammlung aus Wiki-Seiten repräsentiert.

Paralleler Korpus Ein paralleler Korpus unterscheidet sich von einem normalen Korpus darin, dass die Sammlung von Texten oder Äußerungen in zwei oder mehreren Sprachen vorhanden ist. Beispiele dafür sind:

- Geschäftsberichte international agierender Betriebe
- EU-Dokumente
- Äquivalenzwörterbücher

Nachdem einige grundlegenden Begriffe erklärt wurden, kann nun mit dem nächsten Kapitel fortgefahren werden, welches sich mit den traditionellen und grundlegenden Techniken der Thesaurus-Erstellung befasst.

3 Traditionelle Thesaurus-Erstellung

Um eine Vorstellung davon zu bekommen welchen Aufwand die Erstellung eines multilingualen Thesaurus beinhaltet wird in diesem Kapitel auf die traditionellen Verfahren der Thesaurus-Erstellung eingegangen. Grundsätzlich wird, ob es sich nun um ein manuelles oder automatisiertes Verfahren handelt, zwischen drei Herangehensweisen zur Erstellung eines Thesaurus unterschieden:

1. Erstellung eines neuen Thesaurus

Es wird ein neuer mehrsprachiger Thesaurus erstellt, wobei keine bereits existierenden Thesauri einer einzelnen Sprache verwendet werden. Dabei gibt es zwei Ansätze. Entweder wird mit einer Sprache gestartet und andere Sprachen werden sukzessive hinzugefügt oder es wird simultan mit mehreren Sprachen gestartet.

2. **Kombination existierender Thesauren** Die Idee ist jene, bereits bestehende Thesauren verschiedener Sprachen miteinander zu kombinieren. Dabei können die bestehenden Thesauri vereinigt oder verlinkt werden. Werden mehrere Thesauri zu einem neuen vereint, entsteht ein neuer mehrsprachiger Thesaurus. Werden mehrere Thesauri miteinander verlinkt, so bleibt jeder einzelne für sich bestehen und die enthaltenen Begriffe verlinken jeweils auf die Übersetzungen.
3. **Übersetzung eines bereits existierenden Thesaurus in eine andere Sprache**
Hierbei werden die verschiedenen Sprachen nicht gleichwertig behandelt. Die Ausgangssprache wird somit zu der dominanten Sprache im erstellten Thesaurus.

Die im Rahmen dieser Seminararbeit behandelten Ansätze (siehe Kapitel 4), erstellen jeweils einen eigenständigen und neuen Thesaurus. In den folgenden Unterkapiteln wird näher auf die traditionellen Methoden der Thesaurus-Generierung eingegangen.

3.1 Manuell

Die traditionelle Generierung mehrsprachiger Thesauri baut auf menschliche Handarbeit auf. Ein Beispiel dafür ist das EDICT [lin] Projekt, welches 1991 gestartet wurde. Es umfasst zur Zeit über 128.000 Einträge, jedoch brauchte es 18 Jahre und die Arbeit vieler Community Mitglieder um einen so umfangreichen mehrsprachigen Thesaurus zu erstellen. Zudem ist die Richtigkeit der Einträge nicht garantiert, da auch Menschen, die gerade eine fremde Sprache lernen, ihren Beitrag leisten und nicht nur Sprachlinguisten. Wird ein solches Projekt jedoch nur von Sprachlinguisten und Sprachprofis durchgeführt, dauert die Herstellung aufgrund der verminderten Teilnehmerzahl noch viel länger.

3.2 Automatisiert

Die ersten Ansätze der automatisierten Thesaurus-Generierung bauten auf die Methoden der maschinellen Übersetzung auf, welches einfach manuell eingegebene sprachlinguistische Übersetzungsregeln sind. Später wurden diese Regeln durch sprachunabhängige statistische Methoden der maschinellen Übersetzung ersetzt. Diese basieren auf die Verfügbarkeit großer mehrsprachiger Textkorpora. Solche Texte werden tagtäglich erstellt und oft auch veröffentlicht wie z.B. EU-Dokumente oder Geschäftsberichte international agierender Betriebe. Die Vielzahl solcher Dokumente und die öffentliche Zugänglichkeit führte schlussendlich zur Idee sie als Textkorpora für maschinenbasierte Übersetzungen und Extraktion von multilingualen Terminologien zu verwenden. Grundsätzlich werden diese in parallele und vergleichbare Textkorpora unterschieden. Beide Verfahren sind sehr interessant für die automatisierte Thesaurus Konstruktion, jedoch weisen beide immer noch Probleme innerhalb der Abdeckung und Genauigkeit der Terminologie auf. Beispielsweise wird für weit verbreitete Begriffe in der Regel ein

gutes Ergebnis erzielt. Die Qualität des Ergebnisses sinkt jedoch mit der fallenden Verbreitung eines Begriffes. Doch dazu mehr in den einzelnen Unterpunkten.

Erstellung aus einem parallelen Textkorpora Die Extraktion multilingualer Inhalte aus einem parallelen Textkorpora baut auf Dokumenten in einer Sprache und deren Übersetzungen in andere Sprachen auf. Dies wird grundsätzlich in drei Schritten vollzogen:

1. Korpus-Vorbereitung

Im ersten Schritt muss der Korpus vorbereitet werden. Dieser Prozess beinhaltet die Kennzeichnung der Satzgrenzen, Wortverkettungen (Trennung von Wörtern aus anderen Wörtern oder Satzzeichen) und die Kennzeichnung von Worttypen wie Substantive, Verben und Adjektive.

2. Satz-Angleichung

Der vorbereitete Korpus muss nun Satz-angeglichen werden, d.h. jeder Satz in der Ausgangssprache wird mit einem oder mehreren Sätzen in der Zielsprache gekoppelt. Dies geschieht aufgrund von Informationen über Satzlängen oder sogenannten Schlüsselwörtern, deren Übersetzungen bereits bekannt sind.

3. Wort-Angleichung

Im Wortangleichungsprozess werden Übersetzungskandidaten für jedes Wort identifiziert und deren Übersetzungswahrscheinlichkeit berechnet. Die Wortangleichung basiert in der Regel auf die Kookkurrenz¹ der Wörter wie z.B. der Dice Koeffizient² oder für sehr ähnliche Sprachenpaare die Ähnlichkeit der Zeichenketten.

Eine der wichtigsten Fragen der mehrsprachigen Wörterbuch-Extraktion aus parallelen Korpora ist, dass während für hochfrequente Wörter in der Regel gute Ergebnisse erzielt werden, die Genauigkeit drastisch sinkt, wenn ein zu übersetzender Begriff nicht in einer großen Menge im Textkorpora vorhanden ist. Dies trifft vor allem für gebietsspezifische Begriffe zu. Außerdem ist die Genauigkeit für Sprachenpaare aus sehr unterschiedlichen Sprachfamilien relativ gering. Japanisch und Englisch haben z.B. grammatikalisch eine vollkommen andere Struktur. Die englische Grammatik ist vergleichsweise schwierig, da sie viele Wortbeugungen beinhaltet, während die japanische Grammatik sehr einfach ist und es erlaubt Wörter wegzulassen wenn diese aus dem Kontext erkenntlich sind.

Zusätzlich beinhalten parallele Korpora oft keine exakten Übersetzungen. Aus grammatikalischen Gründen oder um zusätzliche Informationen hinzuzufügen, kann daher Text hinzugefügt werden. Umgekehrt ist es auch möglich den Text zu verändern oder gewisse Textteile zu streichen.

¹ Wörter, welche gleichzeitig auftreten

² Algorithmus zur Erkennung von Wortähnlichkeit

Ein weiteres Problem der Informationsgewinnung aus mehrsprachigen Korpora ist, dass nicht für jedes Themengebiet und jede Sprache eine ausreichende Menge an parallelen Korpora gegeben ist. Oft verhindern Copyright-Bestimmungen die Analyse potentieller Korpora. Die derzeit größten parallelen Korpora sind zum einen der „Canadian Hansards corpus“ (Dokumente aus dem kanadischen Parlament in Englisch und Französisch) und der „Europarl corpus“ (Dokumente aus dem europäischen Parlament in allen offiziellen Sprachen der EU). Mit diesen beiden Korpora konnten schon zufriedenstellende Ergebnisse erzielt werden. Für englisch-japanische Wörterbuchextraktion wurde mit Abstracts von Publikationen und Softwaredokumentationen experimentiert. Jedoch im Vergleich zu anderen Sprachen ist die Anzahl der englisch-japanischen parallelen Korpora verschwindend gering.

Somit unterliegt die Thesaurus-Erstellung aus einem parallelen Textkorpus der Verfügbarkeit von Dokumenten und derer wortwörtlichen Übersetzungen. Daher muss die Verwendung vergleichbarer Texte in verschiedenen Sprachen in Betracht gezogen werden, um Themenbereiche abzudecken welche nicht so verbreitet sind wie andere.

Erstellung aus vergleichbaren Textkorpora Da nicht für alle Sprachen und Themengebieten ausreichend parallele Korpora vorhanden sind, ist die Verwendung von vergleichbaren Korpora sehr interessant. Ein vergleichbarer Korpus enthält keine exakten Übersetzungen, jedoch Texte aus demselben Themenbereich. Daraus wird auf eine gemeinsame Terminologie geschlossen. So werden z.B. für die Übersetzung von Englisch nach Japanisch Abstracts von japanischen Patenten mit deren nicht wörtlichen englischen Übersetzungen oder Zeitungsartikel verwendet. Die Sammlung solcher Korpora erweist sich daher verhältnismäßig einfach. Jedoch die Extraktion brauchbarer Informationen umso schwerer.

Somit haben beide Arten von Textkorpora ihre Probleme, wobei diese von grundlegender Natur sind. Vergleichbare Textkorpora werden immer komplizierter zu verarbeiten sein als parallele und die Verfügbarkeit paralleler Textkorpora wird immer geringer sein als die von vergleichbaren. Einen völlig anderen Weg schlägt der Wikipedia-Ansatz ein, welcher im folgenden besprochen wird.

4 Der Wikipedia Ansatz

Wikipedia wurde und wird von der Wikipedia Community, welche eine große Anzahl an aktiven Teilnehmern besitzt, manuell erstellt und erweitert. Somit bietet sich Wikipedia vorzüglich als Plattform für einen multilingualen Thesaurus an, da menschlicher Aufwand mehrfach genutzt wird. Dadurch wird kein menschlicher Mehraufwand benötigt, um passende Textkorpora zu suchen oder zu erstellen. Weiters ist Wikipedia durch die Arbeit und der Erweiterung der Community sprachlich immer auf dem neusten Stand. Dies ist besonders für einen Thesaurus wichtig, da sich Sprache und Sprachgebrauch im Laufe der Zeit

verändert. Vergleichbare parallele Textkorpora ³, welche sich meist auf einen Themenbereich beschränken (Gesetzestexte, Patenttexte, ...), erreichen nicht den Umfang wie Wikipedia, welches viele verschiedene Themenbereiche darunter auch Fachspezifische beinhaltet.

Die traditionellen Verfahren zur Erstellung mehrsprachiger Thesauri bauen, wie im vorigen Kapitel erläutert, auf statistische und computerunterstützte linguistische Methoden auf, was einen hohen Rechenaufwand und teilweise ungenaue Lösungen bedeutet. Jene Ansätze, welche Wikipedia als Datenquelle verwenden, beziehen sich auf die dichte Linkstruktur der Wikipedia anstatt die Texte selbst zu analysieren und vergleichen. Hierbei werden Wortverwandtschaften und Übersetzungen aufgrund der Verlinkungen festgestellt. Somit wird davon ausgegangen, dass wenn zwei Artikel in zwei verschiedenen Sprachen über Sprachenlinks miteinander verlinkt sind, sie die selbe Thematik beinhalten. Die Schlussfolgerung daraus ist, dass die Titel der beiden Artikel die gegenseitigen Übersetzungen sind (siehe Kapitel 4.1).

Datengewinnung durch Wikipedia als Plattform, sog. *Wikipedia mining*⁴ ist ein noch sehr junges Forschungsgebiet. Wie auch die Idee, multilinguale Informationen aus Wikipedia zu extrahieren. Nichtsdestotrotz existieren bereits verschiedene Ansätze, z.B. haben Adafre und Rijke [AdR06] ein bilinguales Wörterbuch aus der Wikipedia Linkstruktur erstellt, um dann einen parallelen Korpus von Wikipedia Artikeln zu erzeugen. Declerck [DPV⁺06] verwendete bilinguale Terminologien um Labels von Ontologien zu übersetzen. Dabei wurde jedoch die Verwendung von Linktexten und Weiterleitungsseiten (siehe Kapitel 4.1) außen vor gelassen. Im Lösungsvorschlag von Erdmann [ENHN08a] werden Linktexte und Weiterleitungsseiten im Prozess der Thesaurus-Generierung miteinbezogen. Das folgende Unterkapitel behandelt die verschiedenen Linkstrukturen welche im obgenannten Lösungsvorschlag verwendet werden.

4.1 Wikipedias Linkstruktur

Wikipedia enthält eine sehr dichte Link-Struktur. Es beinhaltet Links zwischen Artikeln innerhalb einer Sprache und zwischen verschiedenen Sprachen. Im weiteren Verlauf werden die einzelnen Arten der Links genannt und beschrieben, welche benötigt werden, um multilingualen Inhalt aus Wikipedia zu extrahieren.

³ Zwei oder Mehrsprachiger Textkorpus

⁴ http://wikipedia-lab.org/en/index.php/Wikipedia_Mining, besucht: April 2009

Sprachen Links

Sprachenlinks sind Links, welche zwei verschiedensprachige Artikel, die inhaltlich die selbe Thematik behandeln, verbinden. Der Englische Artikel „Fruit preserves“ ist z.B. mit dem deutschen „Marmelade“ Artikel verlinkt (siehe Abb. 1). Standardmäßig befindet sich die Sprachenlinks links vom Artikel in einer sogenannten Sidebar (siehe Abb. 2). In Wikipedia werden Sprachenlinks mit folgendem Syntax erzeugt `[[language code:article title]]`. Wobei der „language code“ die Sprache des Zielartikels definiert und der „article title“ ist der Titel des Zielartikels. Da alle Artikeltitle einer Sprache unique sind, reicht der Artikeltitle aus, um einen Artikel zu identifizieren.

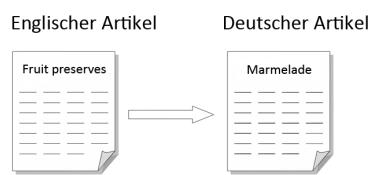


Abbildung 1. Sprachenlinks zwischen Artikeln, in Anlehnung an [ENHN08a]

Das Bild zeigt einen Ausschnitt einer Wikipedia-Seite. Oben links sind zwei Buttons für 'Permanent link' und 'Cite this page' zu sehen. Darunter befindet sich ein Menü 'languages' mit einer Liste von Sprachoptionen: العربية, Český, Deutsch, Español, Esperanto, Français, Hrvatski, Íslenska, Italiano, עברית, Nederlands, 日本語, Norsk (nynorsk), Polski, Português, Română, Русский, Simple English, Slovenščina. Rechts daneben ist ein Bereich '8 External links' und ein Abschnitt 'Variations' mit den Überschriften 'Confit' und 'Conserve'. Unter 'Confit' steht 'Main article: Confit' und eine Beschreibung: 'Confit, which is the past participle for been seasoned and cooked with hone'. Unter 'Conserve' steht eine Beschreibung: 'A Conserve is a jam made of fruit ste'. Darunter folgt ein Absatz: 'Although under EU law, a fruit consen'. Ein weiterer Absatz beginnt mit 'Often the making of conserves can ca long that fruit will break down and liqu fruits are not particularly suitable for n'. Ein abschließender Absatz beginnt mit 'Due to this shorter cooking period, no'. Ganz unten ist der Text 'An alternate definition holds that cons' zu sehen. Am unteren Rand des Screenshot-Bereichs steht 'Fruit butter'.

Abbildung 2. Ausschnitt eines Wikipediaartikels mit Sprachenlinks

Weiterleitung

Weiterleitungsseiten sind Wikipedia-Artikel, welche keine Informationen enthalten. Sie enthalten lediglich einen Verweis/Weiterleitung auf den Hauptartikel. Dies führt zu einem komfortableren Zugang an die Inhalte von Wikipedia. Öffnet ein Benutzer eine solche Weiterleitungsseite, wird er automatisch auf den Zielartikel weitergeleitet. Sucht ein Benutzer im englischen Wikipedia nach „Jam“, wird er automatisch auf den Artikel „Fruit preserves“ weitergeleitet. Wenn die automatische Weiterleitung deaktiviert ist, verbleibt der Benutzer auf einer Weiterleitungsseite (siehe Abb. 4). Eine Weiterleitungsseite wird durch den Eintrag `#REDIRECT [[article title]]` gekennzeichnet, wobei „article title“ auf den Hauptartikel verweist.

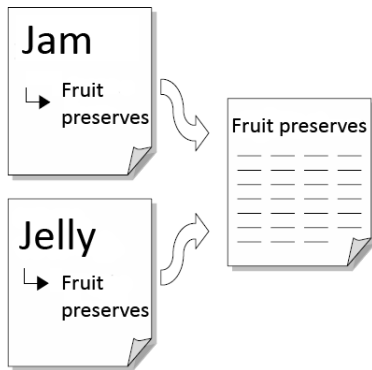


Abbildung 3. Weiterleitung zwischen Artikeln, in Anlehnung an [ENHN08a]



Abbildung 4. Ausschnitt einer Weiterleitungsseite

Linktext

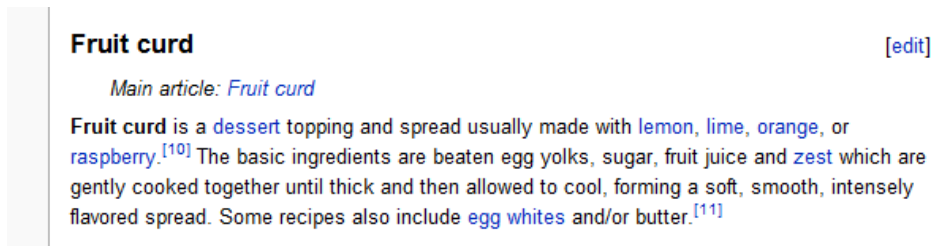


Abbildung 5. Bsp. für Linktexte

Ein Linktext ist der Textteil eines Links. Also der Teil vom Link der im Browser angezeigt wird. Einige Beispiele für Linktexte sind in Abb. 5 ersichtlich *z.B. lemon, lime, orange, raspberry*. Standardmäßig wird im Wikipedia Quellcode `[[article title]]` verwendet, um auf andere Artikel zu verweisen. Hierbei wird der Titel des Zielartikels für den Linktext verwendet. Jedoch ist es möglich, den Linktext mit `[[article title | link text]]` selbst zu bestimmen. Oftmals unterscheiden sich der effektive Artikeltitel und der Linktext nur in Groß-/Kleinschreibung. Manchmal werden die Linktexte aber abgeändert damit sie sich besser in den Kontext fügen. Somit kann dieses Wissen dafür verwendet werden um computerlinguistische Aufgaben zu bewältigen.

Ein-/Ausgehende Links

Für alle vorangegangenen Link-Typen kann die Richtung festgestellt werden. Wie in Abb. 6 ersichtlich, handelt es sich bei einem Vorwärts-Link um einen ausgehenden Link und bei einem Rückwärts-Link um einen eingehenden Link. Forschungen über die Web-Mining-Strukturen, wie z.B. über Google's PageRank [PBMW99] oder Kleinberg's HITS [Kle99], haben bereits bewiesen, dass die Anzahl der eingehenden Links einer Seite von unschätzbarem Wert sind, um objektive und verlässliche Daten daraus zu extrahieren. Jedoch sind diese Algorithmen nicht untäuschbar. Eine bekannte Möglichkeit PageRank von Google auszunutzen ist z.B. das sogenannte GoogleBombing [wikb]. Um aus Wikipedia multilinguale, terminologische Informationen zu gewinnen, sind ein- und ausgehende Links nützlich, z.B. ist die Anzahl der eingehenden Links dafür verwendbar, um die Qualität eines Übersetzungskandidaten zu bestimmen.

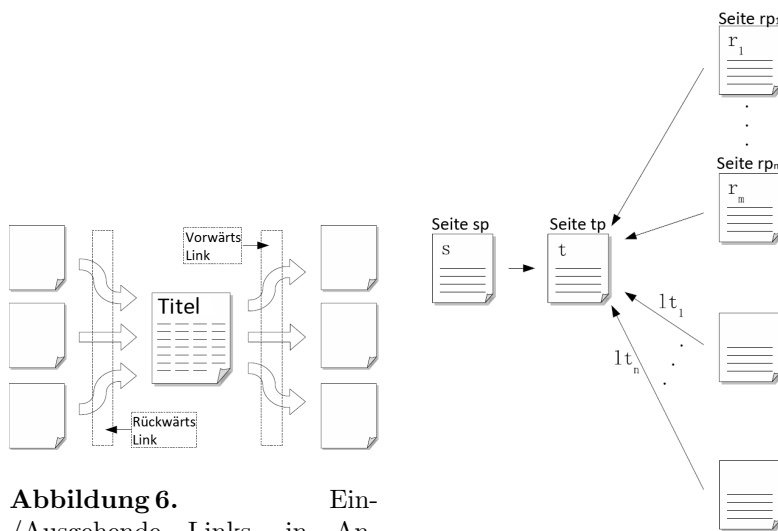


Abbildung 6. Ein-/Ausgehende Links, in Anlehnung an [ENHN08a]

Abbildung 7. Wikipedias Linkstruktur, in Anlehnung an [ENHN08a]

In diesem Unterkapitel wurden die verschiedenen Arten von Links und ihre Funktion in Wikipedia besprochen. Mit dem Wissen um ihre Funktion wird ersichtlich, welche wichtige Informationen sie beinhalten. So weist der Sprachenlink eines Artikels auf einen potentiellen Übersetzungskandidaten hin. Der Titel einer Weiterleitungsseite oder die Linktexte sind mögliche Synonyme des Hauptartikels bzw. des verlinkten Artikels. Zusätzlich werden diese Übersetzungs- und Synonymkandidaten anhand ihrer ein- und ausgehenden Links evaluiert, um auf

ihre Güte rückzuschließen. Im folgenden Unterkapitel werden die Techniken zur Auswahl und Bewertung der Übersetzungskandidaten und Synonyme behandelt.

4.2 Auswahl/Bewertung der Übersetzungskandidaten und Synonyme

Bevor mit der Erstellung eines Thesaurus begonnen werden kann, wird ein Wörterbuch benötigt, welches über die Sprachenlinks erstellt wird (siehe Abb. 1). Laut Erdmann [ENHN08a] gibt es drei Methoden, um die Abdeckung eines Thesaurus zu verbessern. Die „redirect page method (RP method)“, die „link text method (LT method)“ und die Kombination der beiden Methoden „(RP \cup LT Method)“. Die von Erdmann vorgeschlagenen Methoden werden nun im folgenden besprochen.

Die Grundlage Um ein Wort s zu übersetzen, wird zunächst der dazu passende Wikipedia Artikel extrahiert. Diesen Artikel wird fortan sp genannt, was für *source page* steht. Wird ein Artikel gefunden, welcher s als Titel beinhaltet, so übernimmt dieser die Rolle der *source page* sp . Ist s äquivalent zum Titel einer Weiterleitungsseite, so wird die Zielseite dieser Weiterleitungsseite als sp verwendet. Gibt es nun einen Sprachenlink in der gewünschten Sprache im Artikel sp zu einer Seite tp , dann wird der Titel t des Artikels tp als Übersetzungskandidat ausgewählt und der Menge der Übersetzungskandidaten hinzugefügt somit:

$$TC(s) = \{t\}.$$

Diese Menge bildet nun die Grundlage der Übersetzungskandidaten.

Verbesserung durch Weiterleitungsseiten (RP method) Um die zuvor erstellte Basis mit anderen Übersetzungskandidaten (potentielle *Synonyme*) zu erweitern werden die Titel der Weiterleitungsseiten der Menge hinzugefügt. Seien nun R die Titel der Weiterleitungsseiten rp wird die Menge der Übersetzungskandidaten TC als

$$TC(s) = \{t\} \cup R(rp)$$

definiert, also als Vereinigung des Titels des direkt verlinkten Artikels tp und der Titel der Weiterleitungsseiten rp , welche auf tp verweisen. Nicht alle Weiterleitungsseiten sind als Übersetzungskandidat geeignet, da in Wikipedia z.B. für einen weit verbreiteten Rechtschreibfehler eine Weiterleitungsseite erstellt wird, die dann auf den Hauptartikel weiterleitet. Diese müssen im Laufe der Auswahl berücksichtigt und ausgefiltert werden. Ähnlich wie im Google PageRank Verfahren [PBMW99] lässt sich aufgrund der eingehenden Links, den sog. „backward links“ ein Urteil über die „Qualität“ des Übersetzungskandidaten fällen. Begriffe, welche semantisch nicht mit dem Ausgangstitel verbunden oder nur Weiterleitungen von Rechtschreibfehlern sind, weisen in der Regel eine niedrige Rückwärtsverlinkung auf. Um dies nun in die Praxis umzusetzen, gibt es für jede Weiterleitungsseite rp eine Bewertung s_{rp} , welche mit folgender Formel definiert ist

$$s_{rp} = \frac{|e\text{ingehende Links von } rp|}{|e\text{ingehende Links von } tp \text{ und von allen Weiterleitungsseiten von } tp|}$$

Die Tabelle 1 enthält zwei Beispiele für diese Bewertungsformel. Normalerweise

Hauptartikel/Weiterleitungsseite	eingehende Links	Bewertung
Regenschirm	46	0,938
Paraplui	1	0,02
Parapluie	2	0,04
Mobiltelefon	363	0,509
Mobiltelephon	4	0,005
Handyverbot	2	0,002
Händi	4	0,005
Mobiltelefone	15	0,021
Cellphone	2	0,002
Handy	322	0,452

Tabelle 1. Beispiel für die Bewertung für die Zielseite und deren Weiterleitungsseiten

haben die Zielseiten eine höhere Bewertung als die mit ihnen verbundenen Weiterleitungsseiten. Wenn der umgekehrte Fall eintritt ist dies ein Indiz dafür, dass eine Weiterleitungsseite der bessere Übersetzungskandidat ist.

Verbesserung durch Linktexte (LT method) Wie schon in Kapitel 4.1 erwähnt, sind Linktexte der lesbare Textteil von Wikipedia Links. Sei tp ein Übersetzungskandidat, der über einen Sprachenlink gefunden wurde, wird das Ergebnis verfeinert, indem die Linktexte aller eingehenden Links der Seite tp in die Menge der Übersetzungskandidaten TC miteinbezogen werden, also:

$$TC(s) = \{t\} \cup LT(tp).$$

Ein Beispiel für Linktexte auf einen Übersetzungskandidaten tp ist in Abb. 7 ($lt_1 \dots lt_n$) ersichtlich. Wie bei der Verwendung der Weiterleitungsseiten, wird auch bei den Linktexten ein Verfahren benötigt, welches nicht themenbezogene Links ausfiltert. Dafür wird jedem Linktext eine Bewertung s_{lt} zugeordnet. Dabei wird die Anzahl der eingehenden Links eines Übersetzungskandidaten tp mit dem Linktext lt in Relation mit allen eingehenden Links von tp gestellt. Formell bedeutet dies

$$s_{lt} = \frac{|e\text{ingehende Links von } tp \text{ mit Linktext } lt|}{|e\text{ingehende Links von } tp|}$$

In Tabelle 2 ist ein Beispiel für die Bewertungsfunktion der Linktexte enthalten.

Linktext	eingehende Links	Bewertung
mobile phone	2059	0,9846
mobile	30	0,0143
mobile web	2	0,0009

Tabelle 2. Beispiel der Linktext Bewertung [ENHN08b]

Verbesserung durch Linktexte und Weiterleitungsseiten (RP \cup LT Method) Die Letzte von Erdmann in [ENHN08a] erwähnte Methode um die Menge der Übersetzungskandidaten ist die Kombination der zwei vorhergehenden Methoden (LT und RP Methode) also formal

$$TC(s) = \{t\} \cup R(tp) \cup LT(tp)$$

Werden beide Methoden kombiniert, so gestaltet sich die Bewertung komplexer, da mehrere Fälle unterschieden werden müssen. Sei c ein Übersetzungskandidat aus der Menge TC und

$(c \in (\{t\} \cup R) \wedge \in LT)$ c ist der Titel der Zielseite t oder der einer Weiterleitungsseite aus R und zur selben Zeit ein Linktext LT . Somit ist die Bewertung die gewichtete Summe von s_{rp} und s_{lt} , also der Bewertungen von Weiterleitungsseite und von Linktext.

$$s = (w_{rp} \cdot s_{rp}) + (w_{lt} \cdot s_{lt}).$$

$(c \in (\{t\} \cup R) \wedge \notin LT)$ c ist der Titel der Zielseite t oder der einer Weiterleitungsseite aus R aber kein Linktext LT . Hierbei zählt nur die gewichtete Bewertung der Weiterleitungsseite

$$s = (w_{rp} \cdot s_{rp}).$$

$(c \in LT \wedge c \notin (\{t\} \cup R))$ c ist ein Linktext aber weder der Titel einer Zielseite t noch der einer Weiterleitungsseite aus R . Dabei wird die Bewertung nur durch ein gewichtetes s_{lt} berechnet.

$$s = (w_{lt} \cdot s_{lt}).$$

Die Gewichtungen w_{rp} und w_{lt} wurden zur Normalisierung der Bewertungen verwendet, dabei wurde jeweils 0,5 ausgewählt also $w_{rp} = w_{lt} = 0,5$.

Erdmanns Lösungen bilden eine gute Basis, jedoch sind sie noch ausbaufähig und schöpfen nicht alle Möglichkeiten, welche Wikipedia bietet, aus. Es besteht unter anderem die Möglichkeit, Kategorienseiten und Artikelabschnitte zu verwenden, um Subsumtionsrelation (Verallgemeinerungen oder Spezifizierungen eines Begriffes) zu erkennen. Verallgemeinerungen und Spezifizierungen werden in Erdmanns Lösungen nicht erkannt und berücksichtigt. Weiters fehlt Erdmanns Lösungsvorschlag eine Glosse⁵ zu den Begriffen, wie es in Thesauren üblich ist. Im folgenden Kapitel werden Erweiterungen behandelt, welche Kinzler im Kontext seiner Masterarbeit [Kin08] entwickelt hat.

⁵ Bedeutungserklärung eines Begriffes

5 Verbesserungsmöglichkeiten

Erdmanns Ansatz [ENHN08a] bietet zwar die Möglichkeit Synonyme zu finden, jedoch wird dies beim Prototyp ⁶ ihrer Arbeit nicht oder nur beschränkt miteinbezogen, obwohl die Bewertung der Übersetzungskandidaten eine gute Möglichkeit bilden würde um Synonyme auszuwählen. Eine Möglichkeit ist, anhand eines Grenzwertes der Bewertung die zu verwendenden Synonyme auszuwählen, wobei der Begriff mit der höchsten Bewertung als direkte Übersetzung gehandelt wird. Ein ausgereifteres Verfahren bietet Daniel Kinzler in seiner Diplomarbeit WikiWord [Kin08]. Die wichtigsten Unterschiede werden im Folgenden besprochen.

5.1 Glosse und Verwendung der Begriffserklärungsseiten

Glossen werden in Thesauren verwendet, um die Bedeutung eines Begriffes zu erläutern. Das WikiWord Projekt nimmt jeweils den ersten Satz eines Artikels in der Zielsprache her, um eine Glosse zu erstellen. Dies funktioniert, da es in Wikipedia eine Konvention ist, mit dem ersten Satz, den betreffenden Artikel kurz und knapp zu beschreiben. Beispiele für diese Konvention sind in Tabelle 3 ersichtlich.

Artikeltitel	Erster Satz
Cola	Cola, auch Kola, ist ein koffein- und kohlenstoffhaltiges Erfrischungsgetränk.
Lana	Lana a. d. Etsch ist eine Marktgemeinde in Südtirol, Italien südlich von Meran.
Salsa (Tanz)	Salsa ist ein moderner Gesellschaftstanz aus den USA und Lateinamerika, der paarweise oder in der Gruppe getanzt wird.

Tabelle 3. Beispiel für den ersten Satz eines Wikipedia-Artikels

Begriffsklärungsseiten (Disambiguierungen) dienen in Wikipedia dazu um zwischen den möglichen Bedeutungen eines Terms zu unterscheiden. Um daraus Informationen zu gewinnen, werden aus dem Text der Begriffserklärungsseite alle Wiki-Links extrahiert, die auf einen Artikel verweisen, der eine mögliche Bedeutung beschreibt.

5.2 Extraktion der Vorlagen (*Templates*)

WikiWord analysiert die Verwendung von Vorlagen. Vorlagen oder sog. *Templates* geben Strukturen vor, wie ein Artikel eines bestimmten Themengebiet (z.B.

⁶ <http://wikipedia-lab.org:8080/WikipediaBilingualDictionary>, besucht: April 2009

Ortschaft, Chemikalien, Namen, Schauspieler,...) geformt werden muss. Somit ergibt sich aus der Extraktion der verwendeten Vorlagen die Möglichkeit, Wörter zu klassifizieren.

5.3 Klassifikation

Die aus den Wiki-Links extrahierten Konzepte bzw. Wörter werden sofort klassifiziert und einer der folgenden Klassen zugeordnet:

Place Orte, Regionen, geografische Einheiten. Einträge dieser Klasse sind für den Aufbau bzw. die Erweiterung eines Ortslexikon sinnvoll. Beim Aufbau eines Wörterbuches dagegen können Orte gegebenenfalls ausgelassen werden.

Person Natürliche Personen sind für den Aufbau bzw. die Erweiterung eines Personenregisters wichtig. Beim Aufbau eines Wörterbuches dagegen werden Personen in der Regel nicht aufgenommen.

Organisation Organisationen wie Firmen, Regierungs- und Nichtregierungsorganisationen (NGOs), etc. Beim Aufbau eines Wörterbuches werden Organisationen meist nicht aufgenommen.

Name Vor- oder Nachnamen an sich (nicht die betreffenden Personen). Solche Einträge sind für den Aufbau bzw. die Erweiterung eines Namenslexikons sinnvoll. Beim Aufbau eines Wörterbuches dagegen spielen Namen in der Regel keine Rolle.

Time Zeitperioden (wie z. B. 15. Jahrhundert) oder auch ein wiederkehrendes Datum (wie z. B. 6. April). Solche Einträge eignen sich für den Aufbau bzw. die Erweiterung eines Almanachs⁷. Beim Aufbau eines Wörterbuches hingegen werden sie nicht verwendet.

Number Zahlen an sich. Diese können beim Aufbau eines Wörterbuches übergangen werden.

Lifeform Lebensformen, also biologische Taxa, Klassen, Familien, Rassen usw. von Tieren und Pflanzen. Diese können beim Aufbau eines Wörterbuches unter Umständen ausgelassen werden.

Other Sonstige Konzepte. Alle, denen keiner der oben angegebenen Typen zugeordnet werden konnte.

Durch diese Klassifizierung können nicht erwünschte Konzepte gleich zu Beginn aus dem Prozess der Thesaurus-Generierung ausgeschlossen werden.

5.4 Berücksichtigung von Verwandtschaft der Artikel/Begriffe

Die Verwandtschaft von Artikeln/Begriffen wird aufgrund von Querverweisen (Wiki-Links) zwischen den Artikeln bestimmt. Wenn zwei Artikel sich gegenseitig über Wiki-Links referenzieren, kann davon ausgegangen werden, dass die Konzepte, die diese Artikel beschreiben, miteinander verwandt sind.

⁷ Ein Almanach oder auch Jahrbuch genannt, ist ein periodisch erscheinendes Nachschlagewerk

5.5 Berücksichtigung von Ähnlichkeit der Artikel/Begriffe

Die Ähnlichkeit von Konzepten wird über die Sprachenlinks bestimmt: wenn zwei Artikel über Sprachenlinks denselben Artikel in einer anderen Sprache referenzieren, so werden die Konzepte, die diese beiden Artikel beschreiben, als ähnlich angesehen. Der Grund ist, dass Sprachenlinks immer auf ähnliche (oder, idealerweise, äquivalente) Artikel verweisen.

5.6 Berücksichtigung der Subsumierung

Anhand von Wiki-Kategorienseiten Subsumtionsrelation⁸ werden direkt aus den Kategorie-Seiten⁹ der Wikipedia abgeleitet, denen eine Seite zugeordnet ist. Kategorie-Seiten in Wikipedia, sind spezielle Seiten, welche mehrere Artikel einem Überbegriff zuordnen. Wie eine Kategorie Seite aussieht wird in Abb. 8 ersichtlich. Dabei wird nicht zwischen Kategorienseiten und Artikeln unterschieden, das heißt, Kategorien sind ebenfalls einfach Konzepte/Begriffe.

Anhand von Artikelabschnitten In Wikipedia erzeugt jede Überschrift einen Anker, auf den verlinkt werden kann. Dies wird vor allem dann verwendet um einen Begriff zu spezifizieren, welcher keinen eigenen Artikel besitzt. Somit steht der Linktext des Links, welcher auf einen Anker verweist in einer Relation zum dazugehörigen Gesamtartikel. Durch dieses Wissen, werden weitere Informationen zur Subsumierung extrahiert.

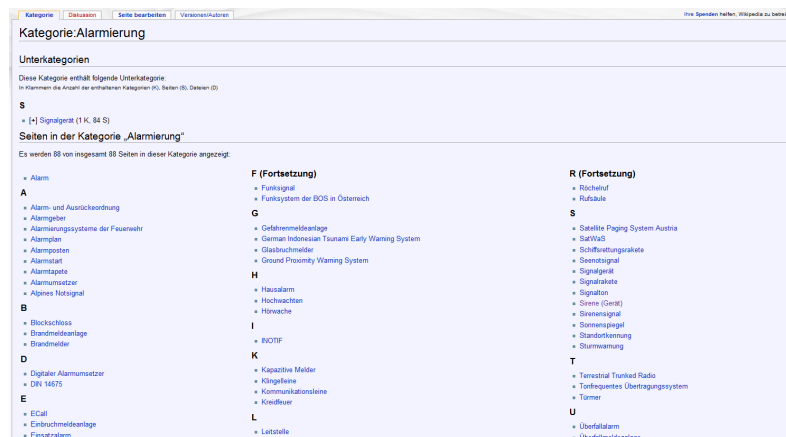


Abbildung 8. Eine Kategorienseite der Wikipedia

⁸ Verallgemeinerungen oder Spezifizierungen eines Begriffes

⁹ <http://de.wikipedia.org/wiki/Wikipedia:Kategorien>, besucht: April 2009

Kinzler konstruierte unter Verwendung der genannten Aspekte einen soliden, webbasierten und multilingualen Thesaurus, welches unter seiner Web-Seite ¹⁰ erreichbar und frei zugänglich ist. Dieses Thesaurus betrachtet, anders als die Lösung von Erdmann, Verwandtheit und Ähnlichkeit verschiedener Konzepte. Zudem wird die Spezifizierung und Generalisierung von Konzepten berücksichtigt.

6 Fazit

Diese Seminararbeit zeigte die verschiedenen Methoden, welche nötig sind, um anhand von Wikipedia ein mehrsprachiges Thesaurus aufzubauen. Der klare Vorteil des Wikipedia-Ansatzes gegenüber traditioneller Verfahren ist, dass die Informationen über Wortverwandtheit und Übersetzungskandidaten aufgrund der Artikelverlinkungen festgestellt werden. Traditionelle Verfahren bauen auf den Vergleich verschiedensprachiger Textkorpora auf. Der Vergleich von Textkorpora ist, wie in Kapitel 3.2 geschildert, ein aufwendiger Vorgang, da die Texte selbst analysiert werden müssen. Ein Beispiel für ein funktionierendes mehrsprachigen Theaurus liefert das Wikiword Projekt¹⁰. Es beachtet die Spezialisierung und Abstrahierung der gesuchten Begriffe und umfasst zur Zeit 3 Sprachen (Deutsch, Französisch und Englisch). Wikiword kann auch als normaler einsprachiger Thesaurus für jede der genannten Sprachen verwendet werden.

Übersetzungen ausgehend von Englisch in andere Sprachen funktionieren gut, da die englische Wikipedia die meisten Artikel enthält und von der größten Anzahl an Benutzern gewartet wird. Jedoch gestaltet sich die Übersetzung von einer wenig verbreiteten Sprache in eine andere nicht so populäre Sprache schwieriger. Dem kann Abhilfe geschafft werden, indem Englisch als Pivotsprache verwendet wird. Angenommen zwei verschiedensprachige Wikipedias in Sprachen \mathcal{A} und \mathcal{B} besitzen untereinander keine oder nur wenige Verlinkungen. Beide sind aber mit der englischen Wikipedia \mathcal{C} gut verlinkt. Wird nun ein Begriff von Sprache \mathcal{A} in Sprache \mathcal{B} übersetzen, wird dies über die Pivotsprache \mathcal{C} bewerkstelligt. Somit verläuft der Übersetzungsvorgang wie folgt: $Sprache_{\mathcal{A}} \rightarrow Sprache_{\mathcal{C}} \rightarrow Sprache_{\mathcal{B}}$.

Größter Kritikpunkt ist, wie bei allen auf Wikipedia basierenden Projekten, dass jeder Benutzer die Möglichkeit hat Artikel zu verändern oder neue Artikel/Links einzufügen. Dies birgt eine große Fehlerquelle in sich. So ergaben Evaluierungen aus [Ham07], dass allein unter den Sprachenlinks verschiedener Paarungen (deutsch-englisch, deutsch-französisch, deutsch-italienisch) bis zu 5 Prozent Inkonsistenzen, wie z.B.:

- Ein Hauptartikel in einer Sprache verlinkt auf eine Disambiguierungsseite.
- Ein Hauptartikel in einer Sprache verlinkt auf eine falsche Übersetzung.

¹⁰ <http://toolserver.org/~daniel/wikiword/wikiword.php>, besucht: April 2009

- In einer Sprache existieren mehrere Artikel, deren Inhalt redundant oder überlappend ist, welche alle auf einen Artikel in einer anderen Sprache verlinken. Dieser ist jedoch nur mit einem Artikel rückverlinkt.

aufwiesen.

Der Wikipedia-Ansatz, welcher aufgrund der Analyse von Wiki-Links arbeitet, birgt noch ein anderes Potential, außer der Thesaurus-Erstellung in sich. In Kombination mit einem traditionellen Verfahren zur Thesaurus-Generierung, wie z.B. die Generierung aus vergleichbaren Korpora (siehe Kapitel 3.2), besteht die Möglichkeit inkonsistente Wiki-Links zu erkennen. Dafür werden zwei über Sprachenlinks verknüpfte Artikel als vergleichbare Korpora verwendet, um ihre Zusammengehörigkeit zu bestätigen oder zu widerlegen.

Zukünftig können traditionelle Verfahren der Thesaurus-Generierung mit den Wikipedia-Ansätzen kombiniert werden, um zusammen bessere Ergebnisse zu erzielen. Bestehende Thesauren würden durch die Ergebnisse der Wikipedia-Ansätze angereichert werden. Dies bedeutet dass Thesauren, welche über Wikipedia generiert werden, in bestehenden multilingualen Thesauren integriert werden. So ist die zukünftige Rolle der Generierung multilingualer Thesauren über Wikipedia eine additive und soll nicht die traditionellen Verfahren verdrängen, da eine gewisse Kontrolle vonnöten ist, um ein Maß an Korrektheit zu garantieren.

Literatur

- AdR06. S. F. Adafre and M. de Rijke: *Finding Similar Sentences across Multiple Languages in Wikipedia*, Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, (2006), pages 62–69.
- DPV⁺06. T. Declerck, A. Pérez, O. Vela, Z. Gantner and D. Manzano-Macho: *Multilingual lexical semantic resources for ontology translation*, (2006), pages 1492–1495.
- ENHN08a. M. Erdmann, K. Nakayama, T. Hara and S. Nishio: *An Approach for Extracting Bilingual Terminology from Wikipedia.*, J. R. Haritsa, K. Ramamohanarao and V. Pudi (editors), *DASFAA*, volume 4947 of *Lecture Notes in Computer Science*, Springer, 2008, pages 380–392.
- ENHN08b. M. Erdmann, K. Nakayama, T. Hara and S. Nishio: *Extraction of Bilingual Terminology from a Multilingual Web-based Encyclopedia*, *Journal of Information Processing*, volume 16, (2008), pages 68–79.
- Erd08. M. Erdmann: *Extraction of Bilingual Terminology from the Link Structure of Wikipedia*, Master's thesis, Osaka University-Graduate School of Information Science and Technology, 2008.
- Ham07. R. Hammwöhner: *Interlingual Aspects of Wikipedia's Quality*, M. A. Robert, M. L. Markus and B. Klein (editors), *12th International Conference on Information Quality (ICIQ-2007)*, M.I.T., 2007, universität Regensburg.

- IFL09. *Guidelines for Multilingual Thesauri*, Technical Report 115, International Federation of Library Associations and Institutions, 2009.
- INHN08. M. Ito, K. Nakayama, T. Hara and S. Nishio: *Association thesaurus construction methods based on link co-occurrence analysis for wikipedia*, *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, ACM, New York, NY, USA, 2008, pages 817–826.
- Kin08. D. Kinzler: *Automatischer Aufbau eines multilingualen Thesaurus durch Extraktion semantischer und lexikalischer Relationen aus der Wikipedia*, Master's thesis, Universität Leipzig, 2008, also available at <http://lips.informatik.uni-leipzig.de/pub/2008-4>.
- Kle99. J. M. Kleinberg: *Authoritative sources in a hyperlinked environment*, *Journal of the ACM (JACM)*, volume 46(5), (1999), pages 604–632.
- lin. The EDICT Dictionary File, http://www.csse.monash.edu.au/~jwb/j_edict.html, besucht am 24.04.2009.
- MEN08. T. H. Maike Erdmann, Kotaro Nakayama and S. Nishio: *A Bilingual Dictionary Extracted from the Wikipedia Link Structure*, *Database Systems for Advanced Applications*, Springer, 2008, pages 686–689.
- MMW06. D. Milne, O. Medelyan and I. H. Witten: *Mining Domain-Specific Thesauri from Wikipedia: A Case Study*, *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE Computer Society, Washington, DC, USA, 2006, pages 442–448.
- NHN08. K. Nakayama, T. Hara and S. Nishio: *A Search Engine for Browsing the Wikipedia Thesaurus*, *Database Systems for Advanced Applications*, (2008), pages 690–693.
- PBMW99. L. Page, S. Brin, R. Motwani and T. Winograd: *The PageRank Citation Ranking: Bringing Order to the Web.*, Technical Report 1999-66, Stanford InfoLab, 1999, previous number = SIDL-WP-1999-0120.
- wika. Wikipedia-Statistik-Tables, Links zu anderen Wikipedias, <http://stats.wikimedia.org/DE/TablesDatabaseWikiLinks.htm>, besucht am 25.05.2009.
- wikb. Google-Bombe – Wikipedia, Die freie Enzyklopädie, <http://de.wikipedia.org/w/index.php?title=Google-Bombe&oldid=58470596>, besucht am 24.04.2009.