OpenThesaurus: ein offenes deutsches Wortnetz

Daniel Naber (dnaber at apache.org)

März 2005

OpenThesaurus ist ein deutsches Wortnetz, das Synonymgruppen und Ober- und Unterbegriffe erfasst. Im Gegensatz zu anderen Wortnetzen wird es nicht (nur) von Linguisten weiterentwickelt, sondern steht jedem zur aktiven Mitarbeit offen. Diese Offenheit bringt Einschränkungen mit sich, was den Detailreichtum der einzelnen Einträge angeht. Andererseits erlaubt die kostenlose Veröffentlichung der Daten und der Software neue Anwendungen, die mit deutschen Wortnetzen bisher nicht möglich waren.

OpenThesaurus is a German wordnet which contains synonym sets and superordinate and subordinate relations between these synonym sets. Unlike other wordnets it is not (only) created by linguists, but everybody is free to contribute. This fact imposes some limitations to the linguistic richness of the thesaurus entries. On the other hand the free availability of both data and software allows new applications that had not been possible before with German wordnets.

Einleitung

Durch die zunehmende Verbreitung von Open-Source-Software auch im Bereich der Office-Programme besteht ein steigender Bedarf an einem frei verfügbaren deutschen Thesaurus, der sich aus den bestehenden Open-Source-Textverarbeitungen heraus nutzen lässt. Dabei bedeutet *frei verfügbar* nicht nur, dass die Daten des Thesaurus kostenlos abfragbar sein müssen, z. B. über eine Webseite. Vielmehr muss auch der Download der kompletten Daten auf den eigenen Rechner möglich sein, ebenso das Ändern der Daten und die Weiterverbreitung der ursprünglichen und der geänderten Version. Nur so ist sichergestellt, dass der Thesaurus auf die gleiche Weise verbreitet werden kann wie die Open-Source-Programme, in die er integriert ist.

Auslöser für die Entwicklung des OpenThesaurus-Projektes war die Veröffentlichung der Office-Software OpenOffice.org 1.0 im Jahre 2002 durch Sun Microsystems. OpenOffice.org (im Folgenden OpenOffice genannt) basiert auf dem bekannten StarOffice, der StarOffice-Thesaurus konnte jedoch aus lizenzrechtlichen Gründen nicht Teil von OpenOffice werden. Während der englische Thesaurus aus den Daten des Moby-Thesaurus (Institute for Language Speech

and Hearing 2000) generiert wurde[1], existierte keine entsprechende Ressource für die deutsche Sprache. In der Open-Source-Community, die OpenOffice zusammen mit Sun Microsystems weiterentwickelt, entstand die Idee einer Website, auf der jeder Benutzer die Einträge eines Thesaurus ergänzen und korrigieren kann. Ich griff diese Idee auf und implementierte das OpenThesaurus-Projekt, das seit März 2003 online ist.

Initialer Datenimport

Ausgangspunkt für den OpenThesaurus-Datenbestand war ein frei zugängliches elektronisches Deutsch-/Englisch-Wörterbuch (Richter 2004), dessen Inhalte in eine Datenbank importiert wurden. Dabei ging man von der Annahme aus, dass die Übersetzungen von englischen Termen Synonymgruppen darstellen, wie in folgenden Beispielen:

bandit: Bandit; Räuber
heraldry: Heraldik; Wappenkunde

Diese Annahme trifft nicht immer zu, wie der folgende Fall zeigt:

monitor: Bildschirm; überwachen

Wie man an den Beispielen sieht, sind im Wörterbuch alle Übersetzungen des jeweiligen Eintrags durch Semikolons getrennt, unabhängig davon, ob es sich um Synonyme oder verschiedene Bedeutungen handelt. Weiterhin werden viele englische Wörter nur mit einem einzelnen deutschen Wort übersetzt. Da solche Einträge für den Thesaurus unerheblich sind, wurden sie ignoriert. Durch den Import der vermeintlichen Synonymgruppen entstand so eine Thesaurus-Datenbasis mit ca. 25.000 Wörtern in 12.000 Synonymgruppen.

Eine derartige Nutzung eines vorhanden zweisprachigen Wörterbuchs ist weit weniger aufwändig als die halbautomatische Übertragung eines englischen Thesaurus auf eine neue Sprache (Scannell 2003) und kommt auch ohne die großen Corpora aus, die man für viele automatische Verfahren benötigt (Senellart & Blondel 2004). Nachteil des Verfahrens ist, dass die Lizenz des entstehenden Wortnetzes von der Lizenz des Wörterbuchs abhängt. Im Fall von OpenThesaurus unterliegt das Wörterbuch mit den Ausgangsdaten der GPL (GNU Project 2004), und damit gilt diese Lizenz auch für OpenThesaurus.

[1] Beginnend mit Version 2.0 von OpenOffice wird der englische Thesaurus auf den Daten von WordNet basieren.

Pflege und Weiterentwicklung

Um die Korrektur der noch reichlich vorhandenen Fehler und eine kontinuierliche Weiterentwicklung sicherzustellen, wurde der Thesaurus auf einer Website (www.openthesaurus.de) zur Verfügung gestellt. Jeder Benutzer der Website kann nach Wörtern suchen und bekommt im Falle eines Treffers alle Synonyme, Ober- und Unterbegriffe angezeigt (vgl. Abb. 1).

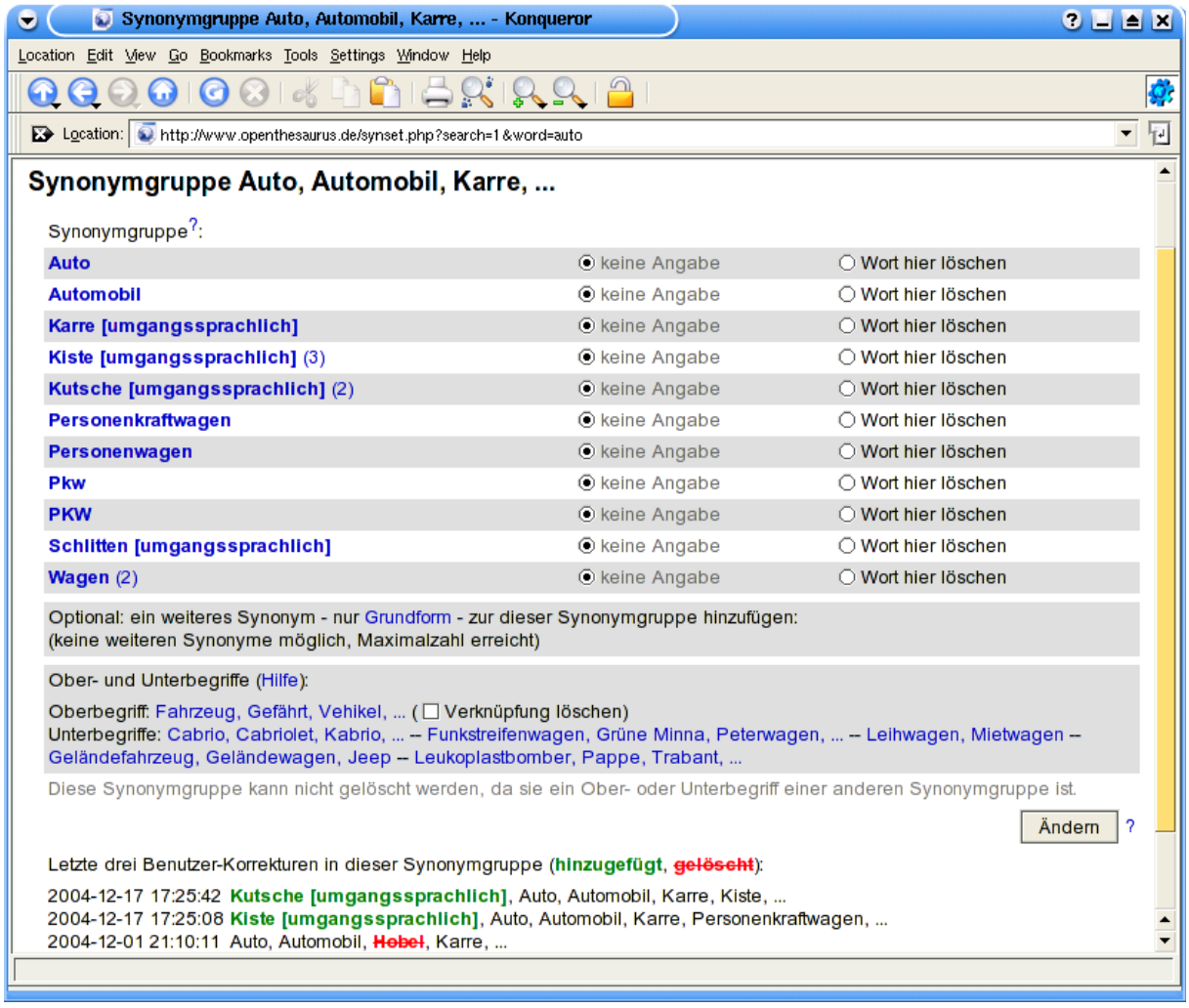

Abb. 1: Suchergebnis für die Anfrage „Auto“

Benutzer können sich auf der Website einen Benutzeraccount anlegen, der ihnen Änderungen an den Daten ermöglicht. Dazu müssen sie ihre E-Mail-Adresse angeben, an die daraufhin ein automatisch generiertes Passwort verschickt wird. Ist der Benutzer eingeloggt, kann er neue Synonymgruppen anlegen, vorhandene löschen und bei vorhandenen Synonymgruppen Wörter löschen und hinzufügen.

Jede neue Synonymgruppe kann optional in eins von 25 Fachgebieten eingeordnet werden (Chemie, Botanik, Medizin etc.). Jedem einzelnen Wort kann in Bezug auf seine Synonymgruppe optional eine der Kennzeichnungen *umgangssprachlich*, *derb*, *vulgär* oder *fachsprachlich* zugewiesen werden. In speziellen Fällen können dem Wort weitere Informationen in Klammern hinzugefügt werden, z. B. *(österr.)*. Diese werden angezeigt, müssen aber bei einer Suche nach dem Wort nicht mit angegeben werden, um es zu finden. Somit lassen sich sinnvolle Meta-Informationen speichern, ohne für jedes neue Attribut die Datenbank-Struktur erweitern zu müssen.

Zur besseren Orientierung wird bei Homonymen die Anzahl ihrer Bedeutungen angezeigt. Beispielsweise wird hinter dem Wort *Bad* in der Synonymgruppe *Bad, Schwimmbad* die Zahl *(3)* angezeigt. Durch Klick auf diese Zahl gelangt man direkt zu den anderen Bedeutungen des Worts (nämlich *Badezimmer* und *Kurbad*).

Um die Qualität der vorhanden Synonymgruppen von Nutzern der Website verbessern zu lassen, befindet sich auf der Homepage ein Link, mit dem zufällig ausgewählte Synonymgruppen angezeigt werden. Dabei erfolgt die Auswahl allerdings nicht völlig zufällig, vielmehr werden Synonymgruppen bevorzugt, die bisher selten angezeigt wurden. Mit Hilfe dieser Funktion können Nutzer an der Qualitätssicherung mitarbeiten oder durch die Daten „surfen“, ohne eine konkrete Suchanfrage einzugeben.

Dass der OpenThesaurus-Datenbestand von Nichtfachleuten geändert werden kann, hat u. a. folgende Konsequenzen:

- Die aktive und passive Nutzung soll ohne große Einarbeitungszeit möglich sein. Das ist nur mit einer einfachen und intuitiven Benutzeroberfläche möglich, die wiederum nur bei einer einfachen und verständlichen zu Grunde liegenden Datenstruktur realisierbar ist.

- Den Benutzern muss eine kurze und leicht verständliche Anleitung zur Verfügung stehen, die beschreibt, was bei Änderungen an den Daten zu beachten ist. Dies wird durch eine mit Beispielen versehene FAQ erreicht, die bewusst kurz gehalten ist und nicht alle speziellen Fälle abdeckt. Dies soll die Hemmschwelle zur aktiven Teilnahme möglichst klein halten.

- Eine gute Qualitätssicherung muss sicherstellen, dass keine absichtlich oder unabsichtlich eingefügten falschen Einträge die Datenbasis dauerhaft

verschlechtern. Dies wird im Abschnitt *Qualitätssicherung* genauer beschrieben.

Einsatzgebiete von OpenThesaurus

Da OpenThesaurus ein freies Projekt ist, steht es jedem Nutzer offen, die Daten und sogar den Sourcecode der Website für seine eigenen Zwecke zu nutzen, zu ändern und weiterzugeben. Deshalb sind dem Verfasser nicht unbedingt alle Projekte bekannt, die OpenThesaurus einsetzen. Zwei typische Einsatzgebiete sollen aber anhand von konkreten Beispielen kurz vorgestellt werden.

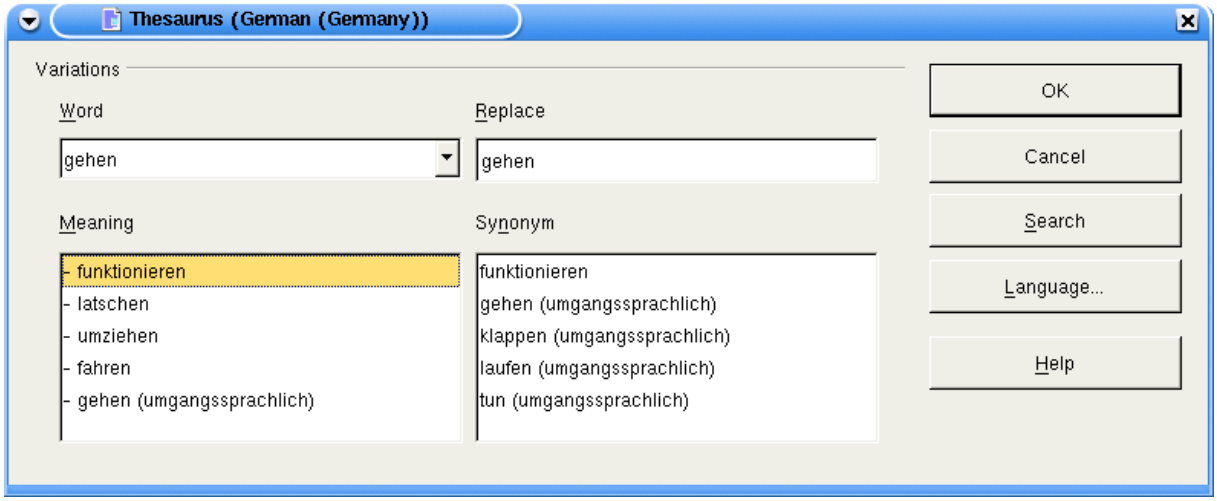

Abb. 2: Der OpenOffice-Thesaurus mit den OpenThesaurus-Daten

Einsatz als Thesaurus in einer Textverarbeitung

OpenOffice verfügt bereits über eine integrierte Thesaurus-Komponente, allerdings nicht über die nötigen Daten für einen deutschsprachigen Thesaurus. Um die OpenThesaurus-Daten mit OpenOffice nutzen zu können, ist ein Export in das von OpenOffice erwartete Format nötig. Es handelt sich dabei um ein einfaches Plain-Text-Format aus zwei Dateien[2]: Die Index-Datei beinhaltet alle bekannten Terme und einen Verweis auf eine Position in der Daten-Datei, die zu jedem Term dessen Synonyme enthält. Die so exportieren Daten werden auf einem Server des OpenOffice-Projekts abgelegt und lassen sich dann vom Benutzer komfortabel über einen OpenOffice-Wizard installieren. Nach der In-

[2] Diese Angaben beziehen sich auf OpenOffice Version 2.0, das Format in Version 1.x unterscheidet sich davon leicht.

stallation lassen sich die Synonyme zu einem gerade markierten Wort nachschlagen (vgl. Abb. 2).

Einsatz im Information Retrieval

Ein Thesaurus kann im Information Retrieval eingesetzt werden, um die Terme einer Suchanfrage mit Synonymen anzureichern und somit den Recall der Suche zu erhöhen (Mandala et al 1999). Bei Suchen auf nicht fachspezifischen Datenbeständen besteht allerdings das Problem, dass bei kurzen Suchanfragen oft nicht klar ist, wie ein mehrdeutiger Begriff mit Synonymen angereichert werden soll: *faul* kann nicht mit *verdorben* und *träge* gleichzeitig ergänzt werden, ohne die Suche weniger präzise werden zu lassen.

Abb. 3: Integration von OpenThesaurus in eine Volltextsuche

Deshalb bietet es sich an, die Suchanfrage nur nach Rückfrage beim Benutzer zu erweitern. Eine Implementierung dieser Herangehensweise ist die Volltextsuche iFinder der Firma IntraFind, die unter anderem bei der BHF-Bank genutzt wird (vgl. Abb. 3). Auf der Trefferseite werden zu jedem Suchbegriff mögliche

Synonyme angezeigt, von denen der Nutzer alle oder nur bestimmte auswählen und die Suche so ausdehnen kann. Die gewählten Synonyme werden zu den jeweiligen ursprünglichen Begriffen disjunktiv hinzugefügt, sie erhalten dabei auch die gleiche Gewichtung wie die ursprünglichen Begriffe.

Im Information Retrieval ist es sinnvoll, wenn der verwendete Thesaurus auch Schreibvarianten unterstützt, z. B. *aufwändig* und *aufwendig*. Bei OpenThesaurus ist das der Fall, allerdings wird nur die neue Rechtschreibung unterstützt. Sofern dort zwei Varianten eines Wortes zulässig sind, beinhaltet die Synonymgruppe auch beide.

Für fachspezifische Suchen ist OpenThesaurus auf Grund seines allgemeinsprachlichen Charakters weniger geeignet. Mit Hilfe der Meta-Informationen ist allerdings eine gewisse Anpassung an bestimmte Situationen möglich. Beispielsweise ist es oft sinnvoll, die als *umgangssprachlich*, *derb* oder *vulgär* gekennzeichneten Wörter herauszufiltern und nicht als Synonyme anzuzeigen. Außerdem lässt sich die thematische Kategorisierung nutzen, um ausgehend von einer OpenThesaurus-Teilmenge ein eigenes fachspezifisches Wordnet aufzubauen.

Auswertung der passiven Nutzung

Passive Nutzung bezeichnet hier die Nutzung der Website www.openthesaurus.de zur Suche, ohne das Ändern, Löschen oder Hinzufügen von Wörtern oder Synonymgruppen. Diese Art der Nutzung ist ohne Anmeldung möglich.

Um einen Eindruck vom Benutzerverhalten zu bekommen, wurden die Ende November 2004 über einen Zeitraum von sieben Tagen protokollierten Suchanfragen ausgewertet. In dieser Zeit wurden insgesamt 7.218 Suchanfragen an die Website gestellt, also durchschnittlich 1.031 Anfragen pro Tag. Eine zufällig ausgewählte Teilmenge von 200 Anfragen wurde genauer untersucht. Bezüglich der gesuchten Wortarten zeigt sich folgende Verteilung:

Nomen	*47%*
Verben	*23%*
Adjektive	*20%*
Sonstige (Adverbien, Partizipien etc.)	*10%*

Von allen Suchanfragen erzielten 75% einen oder mehrere Treffer, d. h. dem Benutzer wurde mindestens eine Synonymgruppe angezeigt, in der das gesuchte Wort vorkam. In 20% aller Fälle erhielten die Nutzer keine Treffer, weil das gesuchte Wort nicht im Datenbestand vorhanden war. Die übrigen 5% erhielten kein Ergebnis, weil die Suchanfrage ungeeignet war (z. B. englisches Wort als Eingabe).

Gibt es für eine Suchanfrage keine direkten Treffer, wird durch drei Maßnahmen versucht, doch noch ein sinnvolles Ergebnis zu liefern:

- Es wird eine Art Tippfehlerkorrektur vorgenommen, indem die Ähnlichkeit der Suchanfrage mit allen Wörtern des Datenbestands ermittelt wird und die ähnlichsten Wörter vorgeschlagen werden. Als Ähnlichkeitsmaß dient die Levenshtein-Distanz. Die Suche nach *attakieren* liefert beispielsweise den Korrekturvorschlag *attackieren*.

- Eine Tabelle der Datenbank liefert zu über 380.000 Vollformen die jeweilige Grundform. Damit erhalten Benutzer auch bei der Suche nach Vollformen einen Treffer, obwohl die Thesaurus-Daten selber nur aus Grundformen bestehen. Die Abbildung von Vollformen auf Grundformen wurde aus den Daten des Morphologie-Tools Morphy (Lezius) generiert.

- Es wird eine Suche nach Teilworten durchgeführt. Diese liefert z. B. bei der Suchanfrage *divers* u. a. die Vorschläge *Diverses* und *Diversifikation*.

Diese Maßnahmen erhöhen die durchschnittliche Trefferquote der Suche von 65% auf die oben genannten 75%. Auf der Seite mit den Korrekturvorschlägen werden zusätzlich Links angezeigt, mit denen die Suchanfrage auf Google oder Wikipedia (de.wikipedia.org) ausgeführt werden kann.

Auswertung der aktiven Nutzung

Aktive Nutzung bezeichnet im Folgenden das Anlegen oder Löschen neuer Synonymgruppen und das Löschen oder Hinzufügen von Wörtern in Synonymgruppen[3]. Seit Bestehen des Projekts sind insgesamt 19.671 solcher Änderungen

[3] Das Ändern von Wörtern ist nur über Löschen und anschließendes Einfügen möglich.

von Nutzern ausgeführt worden, die sich wie folgt verteilen[4]:

Wort zu Synonymgruppe hinzufügen	*11.794*
Wort aus Synonymgruppe löschen	*2.000*
Neue Synonymgruppe anlegen	*2.482*
Synonymgruppe löschen	*901*
Synonymgruppe mit Oberbegriff verknüpfen	*2.494*

Die Zahl der Löschungen zeigt, dass die Benutzer die Aufräumarbeit, die für die automatisch importieren Initialdaten nötig war, zumindest zum Teil durchgeführt haben.

Insgesamt haben sich auf der Website bis heute über 600 Benutzer registriert. Es zeigt sich jedoch, dass die aktive Mitarbeit äußert ungleich verteilt ist: Die aktivsten 10 Benutzer haben 80% aller Beiträge geleistet.

Positiv ist anzumerken, dass es bisher zu keinem Fall von Vandalismus, also absichtlichen Falscheinträgen, gekommen ist. Dass dies nicht selbstverständlich ist, zeigen Projekte wie Wikipedia, die regelmäßig mit Missbrauch zu kämpfen haben. Der Grund könnte in der größeren Bekanntheit von Wikipedia liegen (Wikipedia hat ca. 100-mal mehr registrierte Benutzer) oder in der Tatsache, dass OpenThesaurus zur aktiven Mitarbeit eine Anmeldung voraussetzt.

Qualitätssicherung

Die aktive Teilnahme von Nicht-Fachleuten lässt der Qualitätssicherung eine große Bedeutung zukommen. Zur regelmäßigen Kontrolle aller Änderungen an den Daten existiert ein geschützter Bereich auf der Website, der nur dem Administrator zugänglich ist. Dort befindet sich eine Liste der zuletzt von den aktiven Nutzern durchgeführten Änderungen, die fast täglich überprüft wird. Da alle Änderungen protokolliert werden, können sie vom Administrator leicht zurückgenommen werden, wenn sie nicht sinnvoll sind. Weil zu jeder Änderung Datum und Benutzername gespeichert werden, kann man Benutzer per E-Mail auf ihre Fehler hinweisen, wenn sich falsche Einträge häufen.

[4] Die Änderungen des Projektadministrators sind hier und im Folgenden jeweils nicht mitgezählt.

Der Administrationsbereich enthält auch diverse vordefinierte Datenbankanfragen, wie z. B. die Suche nach Synonymgruppen mit besonders vielen Wörtern und nach Synonymgruppen, die sich ähneln, die also viele gemeinsame Wörter enthalten und daher eventuell zu einer Synonymgruppe zusammengefasst werden sollten. Um die Entwicklung des Projekts besser verfolgen zu können werden außerdem die Anzahl der Suchanfragen in den letzten 24 Stunden, die letzten zehn Suchanfragen und ihre Trefferzahl, und die Nutzer aufgelistet, die sich zuletzt zur aktiven Teilnahme angemeldet haben.

Als wenig hilfreich erwies sich die Idee, Nutzern auf der Startseite die Änderungen anderer Nutzer anzuzeigen und sie über die Qualität dieser Änderungen abstimmen zu lassen. Es kamen zu wenige Bewertungen zustande, die inhaltlich zu widersprüchlich waren: wird eine Änderung nur dreimal bewertet, davon zweimal als *schlecht* und einmal als *gut*, reicht das nicht aus, um daraus ohne manuellen Eingriff eine Rücknahme der Änderung zu veranlassen.

Vergleich zu WordNet, GermaNet und Wikipedia

Anders als WordNet (Fellbaum 1998) und GermaNet (Hamp & Feldweg 1997) verfügt OpenThesaurus derzeit nur über eine Relation zwischen den Synonymgruppen, nämlich die IS-A-Beziehung. Der Grund dafür ist vor allem die Sorge, dass die Nutzung der Website durch viele neue Relationen schwieriger wird und dadurch entweder die Fehlerrate von Nutzerbeiträgen stark ansteigt oder die Zahl der Nutzerbeiträge zurückgeht. Des Weiteren ist die derzeit prominenteste und wichtigste Anwendung von OpenThesaurus die Integration als Thesaurus in OpenOffice, wo über die Synonymie hinausgehende Beziehungen weniger wichtig sind.

Ähnlich wie bei WordNet kann bei OpenThesaurus eine Synonymgruppe nur genau einer anderen Synonymgruppe als Unterbegriff zugeordnet werden, so dass für Nomen eine echte Hierarchie entsteht. Die sieben Synonymgruppen der obersten Hierarchie (*psychologische Eigenschaft*; *Abstraktion*; *Entität, Instanz*, ...) wurden von WordNet übernommen und unter der künstlichen Synonymgruppe *Irgendetwas* eingeordnet, so dass genau eine Baumstruktur entsteht (statt sieben).

Anders als GermaNet wird OpenThesaurus nicht gezielt mit Hilfe eines Corpus erweitert. Dadurch erfolgt eine Konzentration auf solche Wörter, die auch wirklich Synonyme haben statt auf solche, die mit großer Häufigkeit in einem

Corpus vorkommen. Die folgende Übersicht zeigt den aktuellen Umfang von OpenThesaurus:

Synonymgruppen	*13.644*
Wörter	*32.562*
Anzahl der Wörter pro Synonymgruppe	*2,94*

Die OpenThesaurus-Website wurde in der Programmiersprache PHP und mit der Datenbank MySQL implementiert (Naber 2004). Auf Ebene der Dateiformate besteht damit keine Kompatibilität zu WordNet. Die OpenThesaurus-Daten werden als so genannter Datenbank-Dump zur Verfügung gestellt; dabei handelt es sich um eine Textdatei, die direkt in eine relationale Datenbank importiert werden kann und dort dann die gleichen Inhalte und Relationen erstellt wie auf der OpenThesaurus-Website.

PHP und MySQL sind die am weitesten verbreiteten Technologien für interaktive Websites, wodurch die Installation einer OpenThesaurus-Variante auf einem anderen Server für eine weitere Sprache vergleichsweise einfach wird. Bisher existieren OpenThesaurus-Installationen für Spanisch (ca. 5.000 Synonymgruppen), Polnisch (12.000) und Slowakisch (3.200). Die einzelnen Projekte sind allerdings unabhängig voneinander, im Gegensatz zu GermaNet/EuroWordNet besteht also zwischen den Synonymgruppen der unterschiedlichen Sprachen keine Verbindung.

Das OpenThesaurus-Projekt folgt wie die freie Enzyklopädie Wikipedia der Idee, die Benutzer eines Thesaurus bzw. Lexikons aktiv an den Inhalten mitarbeiten zu lassen. OpenThesaurus unterscheidet sich allerdings in zwei wichtigen Punkten von Wikipedia: Erstens sind Änderungen an den Daten bei OpenThesaurus nur nach einer Anmeldung möglich, bei der zwar kein Name aber eine gültige E-Mail-Adresse angegeben werden muss. Zweitens ist die Dateneingabe auf einzelne Wörter und kurze Phrasen beschränkt. Alle Relationen zwischen Wörtern und Synonymgruppen entstehen implizit, die Angabe von Links in einer speziellen Syntax wie bei Wikipedia ist damit nicht nötig und auch nicht möglich.

Fazit

OpenThesaurus eignet sich sowohl, um automatisch generierte Wortnetze manuell zu verbessern und dauerhaft zu pflegen, als auch zum manuellen Aufbau

neuer, fachspezifischer Wortnetze. Die Zugriffszahlen der Website und die Tatsache, dass OpenThesaurus auch mit Linux-Distributionen wie SUSE-Linux ausgeliefert wird, zeigen den großen Bedarf an einem frei verfügbaren deutschen Wortnetz. Als Projekt, das auf der Mitarbeit Freiwilliger basiert, kann es nicht die Vielzahl an Relationen bieten, wie sie sich in WordNet und GermaNet befinden. Seine freie Verfügbarkeit und die Nutzung einer weit verbreiteten relationalen Datenbank als technische Basis ermöglichen aber die einfache Integration in eigene freie oder kommerzielle Projekte, die in dieser Form bisher nicht möglich war.

Literaturverzeichnis

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

GNU Project (2004). *GNU General Public License*. Retrieved January 3, 2005, from Free Software Foundation Web site: http://www.gnu.org/copyleft/gpl.html#TOC1.

Hamp, B. & Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. In P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo & Y. Wilcks (Eds.), *Proceedings of the ACL/EACL-97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (pp. 9-15).

Institute for Language Speech and Hearing (2000). *Grady Ward's Moby*. Retrieved December 18, 2004, from Web site: http://www.dcs.shef.ac.uk/research/ilash/Moby/.

Lezius, W. (n.d). *Morphy - Morphologie und Tagging für das Deutsche*. Retrieved December 9, 2004, from Web site: http://www.lezius.de/wolfgang/morphy/.

Mandala, R., Tokunaga, T. & Tanaka, H. (1999). Combining multiple evidence from different types of thesaurus for query expansion. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 191-197).

Naber, D. (2004). *OpenThesaurus: Building a Thesaurus with a Web Community*. Retrieved January 3, 2005, from Web site: http://www.openthesaurus.de/download/openthesaurus.pdf.

Richter, F. (2004). *German <-> English Dictionary*. Retrieved December 9, 2004, from Web site: http://dict.tu-chemnitz.de.

Scannell, K. P. (2003). Automatic thesaurus generation for minority languages: an Irish example. In *Proceedings of Workshop on Natural Language Processing of Minority Languages with Few Computational Linguistic Resources*. Batz-sur-Mer, France.

Senellart, P. P. & Blondel, V. D. (2004). Automatic discovery of similar words. In M. W. Berry (Ed.), *Survey of Text Mining*. New York: Springer-Verlag.