Modellierung des Sprachraums von Unternehmen

Was man nicht beschreiben kann, das kann man auch nicht finden.

Dr. Thomas Hoppe

Datenlabor Berlin, Email: thomas.hoppe@datenlabor.berlin und

Ontonym GmbH, Berlin, Email: thomas.hoppe@ontonym.de

<Kasten Kernaussagen / Empfehlungen>

1. Autoren und Nutzer von Informationen artikulieren sich mit Bezeichnungen aus unterschiedlichen Sprachräumen, die sie gegenseitig nicht unbedingt kennen.
2. Mit jeder Suche ist eine Übersetzungsaufgabe zwischen diesen Sprachräumen verbunden.
3. Zur Lösung dieser Übersetzungsaufgabe wird Hintergrundwissen über den Sprachgebrauch benötigt.
4. Dieses Hintergrundwissen kann nicht gelernt werden, es muss manuell, durch Interpretation des Sprachgebrauchs aufgebaut werden.
5. Erfahrungswerte zeigen, dass der Aufwand für eine manuelle Modellierung vertretbar gering ist und bei 20-30 Bezeichnungen pro Stunde liegt.
6. Für die Modellierung dieses Hintergrundwissens empfiehlt sich ein pragmatischer, inkrementeller Ansatz.

<Kasten Ende>

Zusammenfassung Ein zentrales Problem, welches sich beim Einsatz semantischer Technologien im Unternehmenskontext stellt, ist bedingt durch die Unterschiede im Sprachgebrauch von Unternehmen, seinen Mitarbeitern und externen Kunden oder Lieferanten und durch die Variabilität unserer Sprache. Unternehmen, Mitarbeiter und Kunden leben und artikulieren in unterschiedlichen Sprach-

räumen[1]. Sie verwenden Bezeichnungen, die der jeweils andere nicht unbedingt kennt, oder erfinden neue Bezeichnungen. Anhand von Praxisbeispielen zeigen wir, woher dieser Unterschied stammt, wie dieser Unterschied durch den Einsatz semantischer Technologien aufgehoben werden kann, was hierzu notwendig ist und warum es sich lohnt, diesen Unterschied aufzuheben.

Hintergrund

Seit den Anfängen des Internets nehmen Suchmaschinen eine wichtige Rolle bei der Suche nach Informationen ein. Ob im Internet, auf Websites oder in den Intranets von Unternehmen, Suche und Suchfunktionen bilden die zentralen Mechanismen, um Informationen zu finden. Suchmaschinen sind gekommen und gegangen, die auf Volltextindexierung basierende Funktionalität dieser Suchmaschinen hat sich jedoch kaum verändert. Erst mit Google kam es zu einer disruptiven Veränderung in der Funktionalität von Suchmaschinen. Durch die Berücksichtigung des „Verlinkungsgrades" von Webseiten, der als Indikator für die Wichtigkeit der Seite resp. der in ihr enthaltenen Informationen betrachtet werden kann, und durch die Analyse der Ankertexte (der auf Webseiten sichtbare Textteil eines Links), die als kurze Inhaltsbeschreibungen der dahinterliegenden Webseite betrachtet werden können, konnte Google die Qualität der Suchergebnisse stark verbessern.

Google hat sich im vergangenen Jahrzehnt zum „goldenen Standard" entwickelt und dominiert mittlerweile die Suche im Internet. Das Verb „googeln" ist mittlerweile im allgemeinen Sprachgebrauch als Synonym für „suche im Internet" etabliert und hat Einzug in Rechtschreiblexika gefunden. Teilweise sind die Ergebnisse, die Google liefert, sehr beeindruckend und die Vorschläge für naheliegende Suchanfragen sehr genau. Dennoch erlebt man als Benutzer, der eine spezifische Information benötigt, dass man seine Suchanfrage häufig mehrfach umformulieren muss, bis man das Gesuchte findet. Als die führende Suchmaschine, die den Großteil des sichtbaren Webs indexiert, besitzt Google einen Vorsprung bei der Analyse von Suchanfragen, Webseiten und dem Suchverhalten von Benutzern: Google kann auf statistischen Analysen großer Datenmengen aufsetzen, um Beziehungen zwischen Bezeichnungen oder Begriffen[2] zu ermitteln und damit – zu mindestens aus statistischer Sicht – einen Teil der Begriffsbedeutung – sprich die Semantik der Begriffe – zu erlernen.

---

[1] Im Gegensatz zur herkömmlichen Bedeutung des Begriffes „Sprachraum" als *geografisch abgegrenztes Verbreitungsgebiet* einer Sprache, verwende ich diesen Begriff in einer allgemeineren Interpretation als *abgegrenzte Menge von Individuen die dieselben Begriffe benutzen*.

[2] Unter „Begriff" oder „Konzept" verstehen wir in diesem Artikel analog zur Definition in Wikipedia (http://de.wikipedia.org/wiki/Begriff) den Bedeutungsinhalt von Objekten, Klassen oder Beziehungen zwischen diesen, der durch eine „Bezeichnung" textuell, bildlich oder wie auch immer geartet repräsentiert wird.

Statistische Verfahren zum Erlernen von Begriffsbedeutungen, die in der Größenordnung einer Suchmaschine im Internet funktionieren, skalieren aber nicht unbedingt herunter auf die Größenordnung von einigen Hunderttausenden oder Millionen von Dokumenten eines Unternehmens, seinen Hunderten bis Tausenden von Produkten oder Mitarbeitern, dem relativ geringen „Verlinkungsgrad“ [1] und der Diversität der Informationsquellen in den Intranets. Es ist daher nicht verwunderlich, dass Suchfunktionen auf Unternehmenswebseiten und in Intranets selber bisher kaum an die Qualität Googles heranreichen.

Ein zentrales Problem, welches sich beim Einsatz semantischer Technologien im Unternehmenskontext stellt, ist bedingt durch die Unterschiede im Sprachgebrauch von Unternehmen, seinen Mitarbeitern und externen Kunden oder Lieferanten und durch die Variabilität unserer Sprache. Unternehmen, Mitarbeiter und Kunden leben und artikulieren in unterschiedlichen Sprachräumen. Sie verwenden Bezeichnungen, die der jeweils andere nicht unbedingt kennt, oder erfinden neue Bezeichnungen.

Eine Frage der Bedeutung

Das Erkennen der Bedeutung von Bezeichnungen oder deren Semantik stellt eine wichtige Voraussetzung für die richtige Beantwortung einer Frage dar bzw. für die Lieferung der richtigen Ergebnisse. Diese Bedeutung ist den Bezeichnungen und Begriffen nicht inhärent. Das heißt, wir können nicht allein anhand der Zeichenfolge einer Bezeichnung (seiner Syntax) auf die Bedeutung schließen. Für uns Menschen stellt dies in der Regel kein Problem dar, da wir im Laufe unserer Entwicklung und unseres Lebens so viel Wissen erworben haben, dass uns in der Regel mit der Nennung einer Bezeichnung sofort der korrespondierende Begriff präsent ist oder wir eine Hypothese über ihre Bedeutung bilden können. Für Computersysteme, die eine eingegebene Zeichenfolge interpretieren müssen, ist dies jedoch ein nicht triviales Problem.

Lassen Sie uns hierzu ein kleines Experiment machen, das ich regelmäßig bei meinen Vorträgen zur Veranschaulichung der Notwendigkeit von Hintergrundwissen verwende. Bevor Sie weiterlesen, bitte ich Sie, sich die folgenden sechs Abbildungen anzusehen und zu überlegen, was die dargestellten Begriffe verbindet. Bitte nehmen Sie sich die Zeit, darüber nachzudenken, bevor Sie weiterlesen.

Abb. X.1 Bedeutungsexperiment Quellen: SXC.hu, 123rf

Sind Sie zu einer Hypothese gekommen? Falls nicht, sehen Sie sich die Abbildung nochmals an. Dieses Experiment habe ich in zahlreichen Vorträgen benutzt und bisher sind lediglich zwei Zuhörer auf die Lösung gekommen. Oft kommen die Zuhörer darauf, dass die gezeigten Begriffe etwas mit „Natur“ zu tun haben oder mit „Mengen“. Offensichtlich sind dies verbindende Ideen zwischen den Bildern. Eine „Computermaus“ hat aber wenig mit „Natur“ zu tun, so dass diese Idee widerlegbar ist. Bei den zwei Mäusen von einer Menge zu sprechen ist mathematisch zwar korrekt, widerstrebt aber unserer intuitiven Vorstellung von einer Menge, zumal beide Mäuse sehr unterschiedlich sind.

Wenn Sie noch nicht auf die Lösung gekommen sind, hilft es vielleicht, wenn Sie die dargestellten Begriffe – sofern Sie alleine sind – laut vorlesen. Falls dies noch nicht reicht, fällt der Groschen vielleicht jetzt, oder wenn ich Ihnen noch mit den Worten „Pinkepinke“, „Zaster“ oder „Mammon“ helfe.

Dies ist eines der wenigen Beispiele, mit dem sich das Konzept „synonymer Begriffe“ rein bildlich veranschaulichen lässt. Das Wesentliche bei diesem Experiment ist, dass die Information, dass diese Bilder Synonyme für den Begriff „Geld“ darstellen, nicht in den Bildern selber enthalten ist, sondern allein durch die Interpretation der Bilder mithilfe seines Hintergrundwissens dem Betrachter offenbart wird.

Als weiteres Beispiel können wir die Bezeichnungen „Boulette“, „Fleischpflanzerl“, „Fleischkücherl“ oder „Frikadelle“ betrachten oder – um in den Unternehmenskontext zu wechseln – „Klempner“, „Flaschner“ und „Spengler“. Auch in diesen Zeichenketten verbirgt sich kein Hinweis, dass diese Bezeichnungen Synonyme sind. Nur durch unsere Erfahrungen und unser Hintergrundwissen betrachten wir diese Bezeichnungen als Synonyme.

Ohne dieses Hintergrundwissen können wir lediglich die Ähnlichkeit der Begriffe anhand ihrer bildlichen oder textuellen Darstellung bestimmen (was Sie vermutlich auch versucht haben). In dieser Hinsicht verhalten wir uns wie ein Computersystem, dem es am nötigen Hintergrundwissen mangelt.

Was lernen wir aus diesem Experiment? 1.) Die Bedeutung einer Objektrepräsentation steckt nicht in der Repräsentation selber, 2.) ohne zusätzliches Wissen haben wir Schwierigkeiten, die Bedeutung der Objektrepräsentation zu bestimmen, und 3.) dieses Wissen muss entweder explizit vorliegen oder ableitbar sein.

Bedeutung von Begriffen im Unternehmenskontext

Warum aber spielen solche Wortbedeutungen im Unternehmenskontext eine Rolle? An Hand von drei Beispielen zeigen wir, dass das Problem unterschiedlicher Wortbedeutungen bei der Suche auf den öffentlichen Webseiten, bei der Suche im Extranet für Geschäftspartner und bei der Suche im Intranet eines Unternehmens auftritt.

Website-Suche bei einem Industrieunternehmen

Die Suchfunktion der BMW-Webseiten lieferte uns über viele Jahre eines unserer Standardbeispiele[3], auf das wir bei einer Analyse der Top 500 Suchanfragen von BMW gestoßen sind. In anderer Form lassen sich ähnliche Beispiele auch bei anderen Suchfunktionen finden, dennoch finden wir das BMW-Beispiel am anschaulichsten, um die Problematik der Begriffsbedeutungen aufzuzeigen.

Ich lade Sie nochmals ein, mich bei einem kleinen Experiment zu begleiten. Haben Sie einen Computer, ein Tablet oder ein Smartphone zur Hand? Wenn ja, öffnen Sie doch bitte einmal die Webseite www.bmw.de.

Rechts oben auf dieser Webseite findet sich das Feld für die Suche[4]. Suchen Sie bitte einmal nach „Russfilter“, „Rußpartikelfilter“ oder – falls Sie Schnelltipper sind – nach „Dieselrussfilter“, wie auch andere dies im Rahmen der Top 500 Suchen bereits vor Ihnen taten. Und, sind Sie mit diesen Begriffen fündig geworden? Sehr wahrscheinlich nicht. Vermutlich aber haben Sie schon ein Bild davon im Kopf, was wir bei diesem Experiment gesucht haben, richtig?

Bedingt durch unser Hintergrundwissen über Anwendungsbereiche und Wortbildungsregeln fällt es uns relativ leicht, allein aus einem oder sehr wenigen Begriffen auf das gesuchte Konzept zu schließen.

Probieren Sie nun noch mal die Variante „Partikelfilter“ oder tippen Sie „Diesel“ langsam, so dass die Autocompletion aktiv wird. Jetzt sehen Sie die Ergänzungsvorschläge „Partikelfilter“ oder „Dieselpartikelfilter“. Wenn Sie nach diesen Vorschlägen suchen, werden Sie zwar fündig; beide Suchen liefern jedoch unterschiedliche Ergebnisse, obwohl sie doch denselben Begriff bezeichnen.

---

[3] Mittlerweile hat BMW die Funktionalität durch eine Autocompletion verbessert, so dass einige unserer damaligen Beispiele leider nicht mehr so anschaulich funktionieren.

[4] Bis Herbst 2014 war dies so.

Aus unserer initialen Analyse der Top 500 Suchanfragen wurde ersichtlich, dass lediglich unter der Bezeichnung „Dieselpartikelfilter“ Informationen gefunden werden konnten. „Dieselpartikelfilter“ ist quasi die normative Bezeichnung für den Begriff im Sprachraum des Unternehmens. Ohne die korrekte Bezeichnung jemals gehört zu haben, werden wir als firmenexterne Benutzer jedoch Schwierigkeiten haben, unsere Anfrage korrekt zu formulieren und die gewünschte Information zu finden.

Wir können bei der Formulierung unserer Anfragen nur auf uns bekannte Bezeichnungen zurückgreifen oder auf Benennungen, die wir selber aus dem gesuchten Begriff ableiten und kreieren. Als Anfragender drücken wir uns damit aber mit Bezeichnungen aus unserem Sprachraum aus und können nur fündig werden, wenn diese auch zum Sprachraum des Unternehmens gehören.

Wenn das Unternehmen jedoch seinen eigenen Sprachraum besitzt, müssen wir unsere Anfragen solange reformulieren, bis wir eine Bezeichnung aus dem Sprachraum des Unternehmens verwenden oder bis wir unsere Suche vorzeitig abbrechen, da wir zu dem Schluss gelangt sind: „zu diesem Thema gibt es nichts“.

Zwar entschärft eine Autocompletion – wie im obigen Beispiel „Partikelfilter“/„Dieselpartikelfilter“ – die Problematik etwas, da wir als Nutzer bereits beim Tippen ein Feedback über die Unternehmenssprache erhalten; dennoch verbleibt das Problem, dass wir mit den vom Unternehmen vorgeschlagenen Bezeichnungen jeweils lediglich einen Bruchteil der passenden Informationen finden und diese nicht mit den Informationen integriert werden, die über eine andere Bezeichnung gefunden werden können.

Extranet-Suche bei einem Marktforschungsunternehmen

Bei einem Marktforschungsunternehmen hatten wir Gelegenheit, mehr als 4.000 Marktstudien zu analysieren, die sich über die gesamte Lebenswelt von Konsumenten erstreckten, sozusagen von „Aalabsatz“ bis „Zylindervertrieb“. Diese Studien sollten über eine Suche auch externen Partnern zugänglich gemacht werden. Auch hier fand sich eine Vielzahl von Bezeichnungen, die synonym sowohl innerhalb einer als auch über mehrere Studien verwendet werden, wie beispielsweise „Fernsehgerät“, „TV-Gerät“, „Fernseher“, „Fernsehempfänger“, „Farbfernseher“ usw. Keine dieser Studien enthielt jedoch alle Bezeichnungsvarianten, sodass bei einer reinen bezeichnungsbasierten Suche lediglich ein Bruchteil der Studien findbar war. Ohne das nötige Hintergrundwissen musste der Suchende auch hier seine Anfragen mehrfach reformulieren.

Intranet-Suche bei einem Fernsehsender

Das Intranet von Unternehmen leidet unter dem gleichen Problem. Bei einem Fernsehsender identifizierten wir in den im Intranet zugreifbaren internen Richtlinien beispielsweise die Begriffe „Kostenerstattung“, „Aufwandentschädigung“, „Aufwendungserstattung“ und „Aufwandserstattung“. Auch hier müssen Mitarbeiter unnötig Zeit mit der Reformulierung von Anfragen verschwenden, um das Gesuchte zu finden, oder sie konsultieren eine Kollegin und halten diese von der Arbeit ab.

Wir lernen aus diesen Beispielen: 1) die Sprachwelt eines Unternehmens und der externen Welt können sich sehr stark unterscheiden, 2) auch in Unternehmen werden – oft trotz einer weitgehenden Normierung von Bezeichnungen – unterschiedliche Synonyme genutzt, 3) ohne eine explizite Repräsentation dieser Synonyme kann nur ein Bruchteil der Informationen gefunden werden und 4) die mit unterschiedlichen Bezeichnungen gefundenen Ergebnis können nicht integriert präsentiert werden und vermitteln damit keinen Gesamteindruck.

Variabilität unserer Sprache und unseres Sprachgebrauchs

> „Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt“,
> Ludwig Wittgenstein [2]

Oben haben wir auf die Informationen in Unternehmen aus der Sicht des Suchenden geschaut. Betrachten wir die Unternehmensinformationen einmal aus Sicht desjenigen, der eine Information, wie z.B. diesen Text, erzeugt. Als Schreibender bin ich zum einen durch meine eigene Sprachwelt und Ausdrucksweise gebunden, zum anderen durch die Bezeichnungen, die unternehmens-, branchen- oder fachgebietsintern verwendet werden.

Wir haben bereits in der Schule gelernt, unsere Ausdrucksweise zu variieren und möchten ungern Dokumente produzieren, die so spannend zu lesen sind wie ein Telefonbuch oder ein juristischer Text. Wir bemühen uns im Allgemeinen, so interessant wie möglich zu formulieren, scheuen aber auch nicht vor „Wortneuschöpfungen“ zurück und verwenden – insbesondere in Patentschriften – durchaus auch Bezeichnungen, die das Gemeinte etwas verschleiern. Wir kreieren daher schon mal für eine Stellenausschreibung anstelle des gewöhnlichen „Vertriebsleiters“, den Begriff „Sales Manager“ oder „Direktor Sales“ und bezeichnen im Ausschreibungstext diese Position als „Leiter Verkauf“, erfinden die Position eines „Fach-IBS-Leiters“ in der Abteilung „Generische Services“ oder bezeichnen einen Computer als „Rechengerät mit integrierter Speichereinheit“.

Darüber hinaus sind wir in unserem Sprachgebrauch von Natur aus nicht immer präzise. Ich lade Sie zu einem weiteren kleinen Experiment ein. Als Sie das erste

Mal das abgebildete Tier sahen, welche Bezeichnung ist Ihnen dazu spontan eingefallen? Bitte nennen Sie wirklich nur den ersten Begriff, der Ihnen in den Sinn kam.

Abb. X.2 Bedeutungsexperiment Quelle: iStockphoto

Mit hoher Wahrscheinlichkeit wird Ihnen vermutlich der Begriff „Kamel“ in den Sinn gekommen sein, so wie bei ca. 60-75% der Personen, mit denen ich dieses Experiment durchgeführt habe. Lediglich ein geringerer Teil der Befragten, dachte spontan an die korrekte Bezeichnung „Dromedar“ oder „Camelus dromedaries“.

In unserer alltäglichen Kommunikation spielen solche Ungenauigkeiten oft eine Rolle – wir bezeichnen da schon mal ein „Auto“ mit seinem Oberbegriff „Kraftfahrzeug“ oder verwenden anstelle von „Bezeichnung“ den verwandten Begriff „Begriff“. Diese Ungenauigkeiten erzeugen aber in der Regel kaum Probleme. Entweder wir können den gemeinten Begriff anhand des Umgebungs- oder Gesprächskontextes ermitteln, oder wir leben schlichtweg mit der falschen Bezeichnung und kommen auch so klar.

Bei der Kommunikation mit einem Computer jedoch sind diese Unterschiede gravierend. Solange ein System nicht über das notwendige Hintergrundwissen über Begriffsbeziehungen verfügt, aus dem abgeleitet werden kann, dass es sich bei allen „Dromedaren“ auch immer um „Kamele“ handelt bzw. dass „Auto“ synonym zu „Automobil“ ist, und es sich dabei auch immer um „Kraftfahrzeuge“ handelt, die im Schweizerischen als „Motorwagen“ bezeichnet werden[5], kann das System lediglich rein syntaktisch auf die Eingaben des Benutzers reagieren. Es liefert ihm dann nur Ergebnisse, die die explizit angefragte Bezeichnung oder seine syntaktischen Varianten enthalten.

[5] http://de.wikipedia.org/wiki/Kraftfahrzeug (Letzter Zugriff: 12.2.2014)

Konsequenzen des Sprachgebrauchs

Wenn wir über Suche und Suchmaschinen reden, haben wir in erster Linie im Sinn, passende Dokumente zu einer Suchanfrage zu finden. Dokumente und Suchanfragen existieren aber nicht losgelöst von einem Zweck. Einerseits werden Dokumente geschrieben, um damit etwas zu bewirken, andererseits werden Anfragen an Suchmaschinen gestellt, um eine übergeordnete Frage zu beantworten. Sowohl hinter den Suchanfragen als auch hinter den Dokumenten stehen Menschen, die mit den von ihnen gewählten Bezeichnungen ihre Gedanken oder Fragen ausdrücken. Suchmaschinen jedoch können weder die Absicht der Menschen noch deren Gedanken allein aus den eingegebenen Bezeichnungen herauslesen. Um diese „Erkenntnislücke" zu überbrücken, benötigen Suchmaschinen Hintergrundwissen, um die von Autoren und Benutzern benutzten Bezeichnungen interpretieren zu können.

Die obigen Beispiele haben gezeigt, dass:

1. Objekte oder Begriffe mit unterschiedlichen Bezeichnungen benannt werden können,
2. die Bedeutung der Bezeichnungen nicht unbedingt aus den Bezeichnungen selber abgeleitet werden kann,
3. Hintergrundwissen benötigt wird, um die in Suchanfragen und Texten verwendeten Bezeichnungen interpretieren zu können,
4. Hintergrundwissen sowohl synonyme Bezeichnungen als auch Ober-/Unterbegriffsbeziehungen umfassen sollte[6],
5. sich der Sprachraum von Unternehmen vom Sprachraum ihrer Umwelt unterscheiden kann,
6. selbst in einem Unternehmen unterschiedliche Bezeichnungen für ein und dasselbe Objekt oder ein und denselben Begriff verwendet werden, und
7. Hintergrundwissen zur Abbildung der Bezeichnungen eines Sprachraums auf die entsprechenden Bezeichnungen eines anderen Sprachraums benötigt wird.

Fassen wir zusammen: Autoren und Informationssuchende formulieren mit den Bezeichnungen ihres eigenen Sprachraums. Diese Sprachräume werden durch allgemeine, umgangssprachliche Bezeichnungen, durch Fachbegriffe und durch selbst kreierte Benennungen, die Menschen erzeugen, um Konzepte auszudrücken, deren „korrekte" oder „offizielle" Bezeichnung sie nicht kennen, aufgespannt.

Solange beide Seiten, Autoren und Informationssuchende, die gleichen Bezeichnungen verwenden, um das Gemeinte auszudrücken, existiert kein Problem und Informationsangebot und Informationsbedarf der beiden Seiten finden zu einander. Wie die obigen Beispiele jedoch zeigen, existiert häufig ein Unterschied im Sprachgebrauch beider Seiten.

---

[6] Daneben erweisen sich auch funktionale Synonyme – Bezeichnungen, die synonym verwendet werden, ohne Synonyme im engeren Sinn zu sein, z.B. Projektleiter und Projektmanager – und Begriffsassoziationen – Begriffe, die zueinander in inhaltlicher Beziehung stehen, als wichtig.

Das eigentliche Problem des Findens passender Dokumente wird daher immer durch eine Übersetzungsaufgabe zwischen den unterschiedlichen Sprachräumen und den Begriffsinterpretationen der Autoren und Informationssuchenden begleitet. Wird diese Übersetzungsaufgabe nicht oder nur unvollständig berücksichtigt – wie dies viele Suchfunktionen lediglich durch syntaktische Abgleiche realisieren – kann ein Informationsbedarf in vielen Fällen nicht ausreichend befriedigt werden.

Um diese Übersetzungsaufgabe jedoch durchführen zu können, wird es notwendig, die Grenzen der gelebten Sprache beider Seiten zu überbrücken, z.B. indem Hintergrundwissen über den Gebrauch von Bezeichnungen beiderseits der Sprachgrenzen genutzt wird.

Terminologiemanagement und Unternehmensthesaurus

Ziel des Terminologiemanagements in Unternehmen ist es, den Sprachgebrauch innerhalb von Unternehmen durch die Schaffung und Nutzung eines normativen Vokabulars zu vereinheitlichen. Der Nutzen des Aufbaus und der Verwaltung einer unternehmenseinheitlichen Terminologie liegt insbesondere in der konsistenten, verbalen Präsentation des Unternehmens nach außen, in präziseren und eindeutigen Dokumentationen und Gebrauchsanleitungen insbesondere technischer Produkte und in einer Vereinfachung automatischer Übersetzungen von Dokumenten und Benutzeroberflächen [3].

Sprache ist jedoch kein starres einmalig fixierbares Gebilde. Sie passt sich immer wieder den Gegebenheiten und der Zeit an (siehe hierzu auch Kapitel 10). Auch in Unternehmen werden neue Bezeichnungen kreiert. Das Marketing erfindet neue Begriffe wie z.B. „unkaputtbar“ für bekannte Eigenschaften oder benennt Produkte um, wie z.B. „Twix“ in „Raider“. Bereichs- oder Abteilungsleitungen entwickeln neue Stellenbezeichnungen, wie z.B. „Senior Prozessmanager P & IT“ oder „Fach-IBS-Leiter“, oder im Zuge von Unternehmenszusammenlegungen neue Abteilungsbezeichnungen, wie z.B. „Services on Demand“ oder „Generische Services“. Selbst die Mitarbeiter weichen im täglichen Sprachgebrauch von den festgelegten Produktbezeichnungen ab und benennen mitunter ein Produkt nur noch durch seine zugrundeliegende Technologie.

Zweckmäßigerweise ist Terminologiemanagement daher als kontinuierlicher Prozess in einem Unternehmen zu betrachten, das zwar nicht täglich, jedoch aber in periodischen Abständen erfolgen sollte.

Nimmt man die Konsequenzen des Sprachgebrauchs im Unternehmen und seiner externen Kunden oder Nutzer ernst, dann muss das Terminologiemanagement erweitert werden. Neben der reinen Normung des internen Sprachgebrauchs muss auch der externe Sprachgebrauch einbezogen werden. Offensichtlich jedoch kann der Sprachgebrauch externer Nutzer nicht normiert werden. Daher muss eine Abbildung ihres Sprachgebrauchs auf die Terminologie des Unternehmens erfolgen.

Unternehmensthesaurus

Während für die einfachste Form des Terminologiemanagements ein Glossar ausreicht, wenn es nur darum geht, Begriffe und deren Definitionen unternehmensintern festzulegen, wird es für Zwecke der Übersetzung unumgänglich, zu Repräsentationsformen wie z.B. Synonymlisten überzugehen, um Synonyme auf normative Begriffeabzubilden. Werden darüber hinaus Begriffe auf unterschiedlichen Abstraktionsebenen benötigt (vgl. drittes Experiment), ist es unumgänglich von reinen Synonymlisten auf Thesauri überzugehen, um die Beziehung zwischen Ober- und Unterbegriffen zu repräsentieren, Querverweise zwischen Begriffskategorien zu berücksichtigen und umfangreichere Vokabulare stärker zu ordnen und zu strukturieren.

Abb X.3 zeigt einen exemplarischen Ausschnitt aus einem Thesaurus, der für

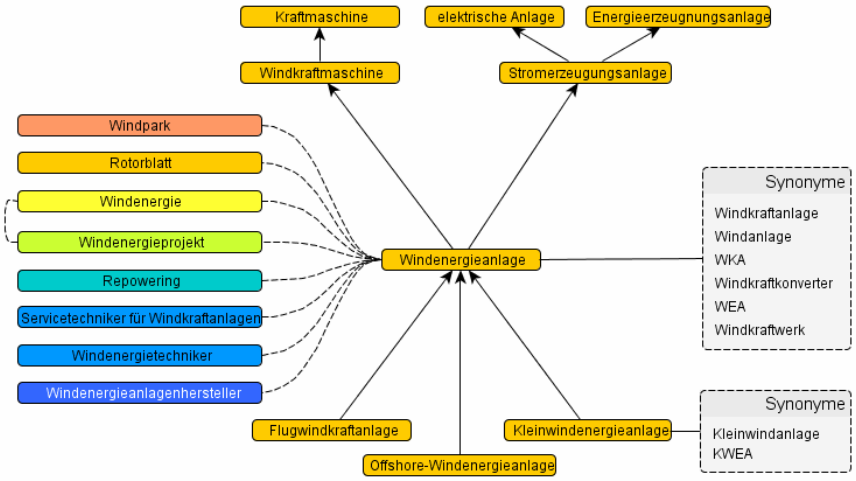

Abb. X.3 Auszug aus dem Thesaurus der Weiterbildungsdatenbank Berlin-Brandenburg zum Themenbereich „Erneuerbare Energien"

den Begriff „Windenergieanlage" einige mit ihm in Zusammenhang stehende Begriffe darstellt. Auf der rechten Seite sind synonyme Bezeichnungen in der gestrichelten Box zusammengefasst. Die Knoten oberhalb des Begriffs zeigen einige der Oberbegriffe, unter denen der Begriff „Windenergieanlage" eingeordnet werden kann; die Knoten unterhalb zeigen einige seiner begrifflichen Verfeinerungen. Auf der linken Seite stehen weitere Begriffe, die mit dem Begriff in Beziehung stehen. Die unterschiedlichen Farben der Knoten symbolisieren hierbei unterschiedliche Begriffskategorien, wie z.B. „Berufs-, Stellen- und Funktionsbezeichnungen von Personen", „Branchen- und Unternehmensbezeichnungen" oder „physikalisch/technische Begriffe".

Ein solcher Unternehmensthesaurus, der den gelebten Sprachgebrauch von Autoren und Informationssuchenden berücksichtigt, stellt das notwendige Hinter-

grundwissen dar, um zwischen den Sprachräumen beider Seiten zu vermitteln. Darüber hinaus bildet er einen ersten Ansatz, das Unternehmenswissen zu strukturieren.

Mut zur Lücke: Arbeiten mit unvollständigen Terminologien

Am Beispiel der Abb. X.3 sehen wir, dass einige Beziehungen zwischen Begriffen bereits sehr präzise definiert werden. Die Ober- und Unterbegriffe bilden eine Hyperonym/Hyponym-Beziehung ab und definieren Ober- und Unterklassen von Konzepten. Synonyme repräsentieren semantisch äquivalente Begriffe und bilden äquivalente Konzepte aufeinander ab.

Andererseits sind einige Begriffe in diesem Thesaurus noch sehr unbestimmt bzw. unvollständig. So gibt es viele andere Arten von „elektrischen Anlagen", „Stromerzeugungsanlagen" oder „Kraftmaschinen", die hier nicht abgebildet sind. Wollte man den Anwendungsbereich der „technischen Anlagen und Geräte" modellieren, wäre der obige Ausschnitt als unvollständig zu bezeichnen. Zur Beschreibung spezieller „technischer Anlagen im Bereich erneuerbarer Energien" reicht er jedoch in dieser Form vorübergehend aus.

Auch bei den Synonymen ist nicht garantiert, dass alle möglichen Synonyme erfasst sind. Es könnte durchaus passieren, dass die Bezeichnung „Windenergiekonverter" signifikant oft von bestimmten Benutzergruppen verwendet wird. Auch für Synonyme kann eine Vollständigkeit somit nicht zugesichert werden.

Neben der Verfeinerung von Begriffen kann es sich für bestimmte Anwendungszwecke als notwendig erweisen, die im Beispiel genannten Beziehungen weiter zu verfeinern und von der allgemeineren „steht in Beziehung zu"-Relation zu Relationen überzugehen, die die Art der Beziehung näher charakterisieren. Beispielsweise „umfasst" ein Windpark mehrere Windenergieanlagen, ist ein Rotorblatt „Bestandteil" einer Windenergieanlage oder werden Windenergieanlagen von Windenergieanlagenherstellern „produziert".

Offensichtlich sind in einem Anwendungsgebiet nicht alle Bezeichnungen im Voraus bekannt. Sie können einer kontinuierlichen Veränderung unterliegen und es kann zu einem späteren Zeitpunkt notwendig werden, bestimmte Begriffe weiter zu präzisieren. Für den Aufbau von Thesauri in Unternehmen heißt dies, dass ein pragmatischer Weg beschritten werden muss, um einen Unternehmensthesaurus unter den genannten Randbedingungen und begrenzten Ressourcen aufzubauen und zu pflegen. Dies bedeutet einerseits, mit den aktuell wichtigsten Begriffen und Bezeichnungen zu beginnen, sich im Voraus definierte Grenzen zu setzen und sowohl Mut zur Lücke zu zeigen und andererseits zeitweise auch bewusst Ungenauigkeiten und Vereinfachungen in Kauf zu nehmen.

Pragmatischer Aufbau von Unternehmensthesauri

Für den Aufbau von Unternehmensthesauri hat sich über mehrere Projekte eine pragmatische Vorgehensweise am geeignetsten herausgestellt, da man vermeiden möchte, dass irrelevante Bezeichnungen, Konzepte und Beziehungen erfasst und modelliert werden und um möglichst schnell zu einer nutz- und evaluierbaren Modellierung zu gelangen.

Begriffsanalyse des Anwendungsbereichs

Typische Unternehmensanwendungen benötigen oft nur wenige Begriffskategorien. Es ist daher am zweckmäßigsten anhand einer Analyse des Anwendungsbereichs diese Begriffskategorien zu identifizieren.

Beispielsweise spielen im Umfeld von Stellenausschreibungen die Begriffskategorien „Berufs-, Stellen-, Funktionsbezeichnungen“, „Skills/Kompetenzen“, „Aufgaben/Tätigkeiten“, „Branchen“ und „Arbeitsergebnisse“ eine zentrale Rolle. Namen, Produkte, Orte und Adressen, spielen in Abhängigkeit von der konkreten Anwendung hingegen nur eine nachgeordnete Rolle. Geht es hingegen darum, eine Produktsuche zu verbessern, dann spielen „Produktkategorien und -bezeichnungen“, „Eigenschaften“, „Einsatzgebiete“, „Verwendungsbeschränkungen“ die zentrale Rolle; bei einer Suche nach Marktstudien hingegen sind dies „Produktkategorien“, „Zielgruppen“ und Fachbegriffe aus dem „Marketing“ und der „Werbung“.

Bereits durch die Analyse des Anwendungsbereichs und der intendierten Nutzung des begrifflichen Hintergrundwissens können relevante Begriffskategorien identifiziert werden und es kann vermieden werden, dass die finale Modellierung in irrelevante Begriffskategorien abdriftet.[7]

[7] Es hat sich gezeigt, dass im Lauf der Entwicklung umfangreicherer Modellierungen zunehmend Begriffe an der Peripherie des Begriffsraums benötigt werden. Beispielsweise musste unser Recruitment-Thesaurus um „Produktkategorien“ und viele Begriffe und Bezeichnungen aus dem Bereich der „Technik“ erweitert werden. Insofern kann der von uns empfohlene Modellierungsansatz im Gegensatz zu den bekannten „Top-Down“- oder „Bottom-Up“-Ansätzen als „Inside-Out“ bezeichnet werden. Ausgehend von den zentralen Begriffen einer Domäne werden nach und nach Ober-, Unter- und periphere Begriffe hinzugenommen, so dass die Modellierung quasi aus der Mitte heraus wächst.

Informationsquellen

Ausgangspunkt für die Modellierung bildet – nach der Analyse des Anwendungsgebiets – die Analyse der in den Informationsquellen verwendeten Bezeichnungen. Diese Informationsquellen können beispielsweise sein:

- existierende externe Glossare, Synonymlisten, Thesauri oder Ontologien, um den fachsprachlichen Begriffsgebrauch abzubilden
- firmeninterne Datenbanken und Strukturen (Produktkataloge, Strukturierungen des Produktportfolios aus Sicht des Vertriebs, der Marketingabteilung oder des Produktmanagements) zur Erfassung und Abbildung des Unternehmenssprachgebrauchs
- Dokumente des Unternehmens und seiner Zulieferer, um den Sprachgebrauch der Autoren zu erfassen
- Analyse der Eingaben in Eingabefelder oder Logdateien von Suchfunktionen und Suchmaschinen, um den gelebten Sprachgebrauch von Mitarbeitern oder externen Benutzern zu erfassen

Durch Verfahren des Text Minings, beispielsweise der Zerlegung von Dokumenten in linguistische Entitäten wie z.B. Sätze, Komposita, Nominalphrasen, Verben, Adjektive, Adverbien, das Ignorieren von Stoppworten und die Wortstammbildung können die in diesen Quellen verwendeten Bezeichnungen bestimmt, gezählt und die Häufigkeiten der Verwendungen der extrahierten Bezeichnungen ermittelt werden.

Häufigkeitsverteilung

Die absoluten Häufigkeiten besitzen nur eine geringe Aussagekraft, ihre Verteilung charakterisiert die Begriffswelt des Unternehmens jedoch gut. Eine solche Häufigkeitsverteilung für die Anfragen an die Stellensuche von *ingenieurkarriere.de* des VDI-Verlags ist beispielhaft in Abb. X.4 dargestellt. Die rote Kurve zeigt die absolute Häufigkeit der Begriffe (linke vertikale Primärachse). Die grüne Kurve visualisiert den Logarithmus der absoluten Häufigkeit (rechte vertikale Sekundärachse), um darzustellen, dass selbst hinter der 80%-Schwelle des Paretoprinzips, die beim 1504. Begriff liegt, die Häufigkeit der Begriff noch 20 beträgt.

Wenige Begriffe werden sehr häufig verwendet (linke Seite). Sie bilden in der Regel die zentralen Begriffe der Anwendungsdomäne des Unternehmens und fungieren als Ausgangspunkt für die Modellierung. Der sogenannte „long tail“ der Häufigkeitsverteilung wird durch Synonyme und Bezeichnungsvarianten gebildet, die einzeln betrachtet zwar nur relativ selten verwendet werden, die aber zusammengenommen die Sprachvielfalt der Domäne erfassen.

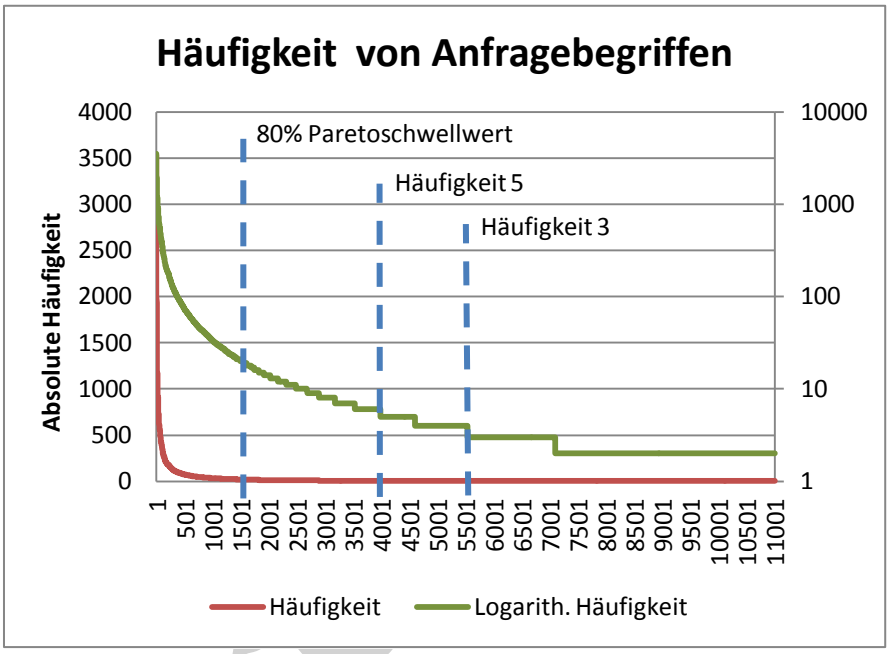

Abb. X.4 Verteilung von Anfragehäufigkeiten nach Rang der Anfrage (ingenieurkarriere.de des VDI-Verlags)

Natürlich finden sich im „long tail“ auch viele Begriffe und Bezeichnungen, die nicht unbedingt nützlich sind, von Rechtschreib- und Tippfehlern über sehr exotische Formulierungen und Begriffe, die mit dem eigentlichen Informationsangebot überhaupt nicht befriedigt werden können, bis hin zu Analyseartefakten (insbesondere beim Text Mining aus PDF Dateien), so dass Bezeichnungen, die seltener als 3-5 mal auftreten, in der Regel vernachlässigt werden können.

Die obige Abbildung zeigt darüber hinaus, dass zwischen der 80% Schwelle und einer Begriffshäufigkeit von 5 rd. anderthalb mal so viele Bezeichnungen liegen, die zwar selten verwendet werden, die jedoch noch nicht in den zu vernachlässigenden Häufigkeitsbereich fallen. Ob diese Bezeichnungen bei der Modellierung zu berücksichtigen sind, ist in der Regel eine Kosten-/Nutzenabwägung, die anhand des Anwendungsgebiets getroffen werden muss.

Generell jedoch gilt, dass es sich lohnt, die nicht berücksichtigten Bezeichnungen bei einer späteren Überarbeitung des Thesaurus mit in die Bewertung der Begriffshäufigkeiten einfließen zu lassen.

Aufwand und Nutzen

Es stellt sich natürlich die Frage nach dem Verhältnis zwischen Nutzen und Aufwand, der für den Aufbau eines Unternehmensthesaurus nötig ist. Diese Frage wird in der Regel gefolgt von der Frage, ob dieser Prozess automatisiert werden kann bzw. ob der Thesaurus nicht automatisch erlernt werden kann.

Automatischer Aufbau des Thesaurus

Unserer Erfahrung und Kenntnis nach gibt es bisher keinen Ansatz, der auf realen Daten basiert und über reine Beispielanwendungen hinausgeht, mit dem ein Unternehmensthesaurus automatisch erlernt werden kann, der gleichzeitig von Menschen verstanden und dessen Richtigkeit beurteilt werden kann. Dies liegt einerseits daran, dass die Datenmengen, die statistischen und maschinellen Lernverfahren oder Data Mining Verfahren zur Verfügung stehen – wie eingangs erwähnt – in Unternehmensanwendungen in der Regel zu gering sind. Andererseits produzieren diese Verfahren artifizielle Repräsentationen, die zwar performant sind, die aber nicht unbedingt die Realität so abbilden, dass die Ergebnisse auch noch von Menschen interpretiert werden können und als richtig beurteilbar sind.

Wesentlicher aber ist – wie unser Eingangsexperiment gezeigt hat –, dass die Bedeutung einer Bezeichnung nicht immer aus der Bezeichnung selber ermittelt werden kann und dass der Kontext, in dem sie benutzt wird, zu berücksichtigen ist. Damit aber ein Verfahren aus diesen Kontextinformationen lernen kann, müssten die Ausgangsdaten weitestgehend homogen sein. Die im Unternehmenskontext zur Verfügung stehenden Daten besitzen entweder in der Regel keine oder nur geringe Kontextinformationen, noch sind sie homogen genug, dass sie ohne umfangreichere Aufbereitung von maschinellen Lernverfahren nutzbar wären.

Aufwand manueller Modellierung

Bedingt durch diese Einschränkungen haben wir uns entschieden, um unsere ersten Demonstratoren und Anwendungen zu bauen, Thesauri von Hand zu modellieren. Dabei hat sich schnell herausgestellt, dass eine manuelle Modellierung von Thesauri doch nicht so aufwendig ist, wie sie von vielen Autoren – insbesondere Wissenschaftlern – dargestellt wird. Solange ein pragmatischer Modellierungsansatz verfolgt wird, der temporäre Unvollständigkeiten und Ungenauigkeiten in

Kauf nimmt, Mut zur Lücke zeigt und sich auf das konzentriert, was für die Aufgabenstellung zu erfassen ist, kann der Modellierungsaufwand mit 20-30 Bezeichnungen pro Stunde beziffert werden.

Dieser Erfahrungswert hat sich im Lauf der Zeit über viele Modellierungen als relativ konstant erwiesen. Die Schwankungsbreite hängt einerseits von etwaigen Hintergrundrecherchen ab, die zur Klärung der Begriffsbedeutung notwendig sind, und andererseits von der Vernetzung von Begriffen mit anderen bereits modellierten Begriffen, mit der Beurteilung und Qualitätssicherung bereits modellierter Begriffe im Lichte neuer Begriffe und der Korrektur älterer Begriffe.

Der eigentliche Aufwand zur Modellierung eines Anwendungsgebiets wird viel stärker durch die Breite des Anwendungsgebiets und damit durch die Variabilität der Begriffswelt und der Bezeichnungen bestimmt. Ein Anwendungsbereich wie z.B. Marktstudien über Consumer-Produkte, der alle möglichen Produkte behandeln kann, besitzt eine größere Begriffsvielfalt als eine Unternehmensanwendung, bei der es nur um die im Unternehmen eingesetzten Produkte und Technologien geht. Eine genauere Abschätzung des Aufwands, der zur Modellierung eines Anwendungsbereichs notwendig ist, kann in der Regel im Rahmen eines Vorprojektes erfolgen, in dem

- die anwendungsbereichsspezifischen Begriffskategorien ermittelt werden,
- die zur Verfügung stehenden Daten analysiert werden,
- anhand der Häufigkeitsverteilung eine Stichprobe der ermittelten Bezeichnungen anhand der Begriffskategorien klassifiziert wird,
- die Anzahl der insgesamt zu modellierenden Bezeichnungen aus der Stichprobenklassifikation hochgerechnet wird.

Anhand des obigen Erfahrungswerts kann beispielsweise der Aufwand zur Abschätzung einer Modellierung von 4.000 Begriffen mit durchschnittlich 160h (20 PT) beziffert werden. Legt man unternehmensinterne Personalkosten von 50 € pro Stunde zugrunde, ergäben sich Kosten von 8.000 € für die Modellierung durch einen bereits geschulten und eingearbeiteten Mitarbeiter.

Es wird ersichtlich, dass weder der Aufwand noch die Kosten ein Argument gegen eine manuelle Modellierung darstellen. Der eigentliche Nachteil manueller Modellierung liegt eher in der sich stark wiederholenden Tätigkeit, die den Charakter einer kaum enden wollenden Fließbandarbeit besitzt, für die es einer bestimmten Mitarbeitereinstellung bedarf.

Nutzen von Unternehmensthesauri

Neben dem Nutzen, den das Terminologiemanagement für ein Unternehmen besitzt, können Unternehmensthesauri für die Umsetzung von semantischen Unternehmenswendungen genutzt werden. Diese reichen von thesaurus-basierten semantischen Suchverfahren, wie sie in Kapitel 2, 23, 1 und 14 beschrieben werden, über die Realisierung von wissensbasierten Benutzeroberflächen und Portalen (Kapitel 1, 13) bis hin zur Umsetzung von semantischen Wissensmanagementan-

wendungen (Kapitel 17, 21, 24). Der Frage, wie der Nutzen eines Unternehmensthesaurus im Kontext semantischer Suche bewertet und damit auch wie der Nutzen semantischer Suche quantifiziert werden kann, gehen wir in Kapitel 9 nach.

Ein Unternehmensthesaurus kann darüber hinaus als Hilfsmittel für Analysen eingesetzt werden. Dies kann – nochmals – am Beispiel des „Dieselpartikelfilters" von BMW skizziert werden. Betrachtet man die Häufigkeit von Suchanfragen als Aussage von Suchenden über Themen, die sie interessieren, dann können durch die Analyse des Rankings der Anfragen die Themen identifiziert werden, an denen die Nutzer interessiert sind. Bei den Top 500 Suchanfragen von BMW beispielsweise belegten die Anfragen folgende Ränge: Partikelfilter (Rang 58), Russpartikelfilter (Rang 214), Dieselpartikelfilter (Rang 221), Russfilter (Rang 255), Rußpartikelfilter (Rang 260 und 309 durch unterschiedliche Kodierung des „ß") und Rußfilter (Rang 348 und 400, ebenfalls durch unterschiedliche Kodierungen).

Durch Verwendung des Hintergrundwissens eines Unternehmensthesaurus können diese Anfragen zusammengefasst werden. Das Thema „Dieselpartikelfilter" würde dann auf Rang 11 aufzeigen, dass dies ein für viele Nutzer wichtiges Thema ist. Solche wissensbasierten Anfrageanalysen bieten für Marketing, PR und Marktforschung interessante Potentiale, um proaktiv die Informationsbedürfnisse und Interessen von Nutzern zu identifizieren und sie in entsprechenden Maßnahmen aufzugreifen.

Literaturverzeichnis

1. Schillerwein, S. Der ‚Business Case' für die Nutzung von Social Tagging in Intranets und internen Informationssystemen In: Good Tags – Bad Tags Social Tagging in der Wissensorganisation, Birgit Gaiser, Thorsten Hampel, Stefanie Panke (Hrsg.) , Medien in der Wissenschaft, Band 47, Gesellschaft für Medien in der Wissenschaft e.V., Waxmann Verlag, 2008.
2. Wittgenstein, L. Tractatus Logico-Philosophicus, Project Gutenberg, http://www.gutenberg.org/files/5740/5740-pdf.pdf (Letzter Zugriff: 31.1.2014).
3. Keller, N. Terminologie-Management – ein Erfolgsfaktor für Unternehmen, 12. DTT-Symposion 2010 - Best Practices in der Terminologiearbeit, Universität Heidelberg, 15. April 2010, http://www.iim.fh-koeln.de/dtt/tutorialsundvortraege/Keller-Tutorial1.pdf (Letzter Zugriff: 20.1.2014).