Semantische Analyse und automatische Gewinnung von branchenspezifischem Vokabular für E-Commerce

Daewoo Kim

München 2007

Semantische Analyse und automatische Gewinnung von branchenspezifischem Vokabular für E-Commerce

Daewoo Kim

Dissertation
an der CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG (CIS)
der Ludwig–Maximilians–Universität
München

vorgelegt von
Daewoo Kim
aus Seoul / Korea

München, den 10.10.2007

Erstgutachter: Herr Prof. Dr. Franz Guenthner
Zweitgutachter: Herr Prof. Dr. Klaus Schulz
Tag der mündlichen Prüfung: 25.01.2008

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

I. **Erkennung der domainspezifischen Terme im jeweiligen Bereich durch EGT (Elementare Generische Terme) und domainspezifische Listen**

II. **Erstellung des Terminologie-Extraktionssystems AGBV**:
Automatische **G**ewinnung von **B**ranchenspezifischem **V**okabular aus den erstellten Korpora

Die vorliegende Arbeit ist den beiden oben genannten Zielsetzungen gewidmet. Um Webseiten für E-Commerce inhaltlich zu erfassen, wird branchenspezifisches Vokabular für die jeweiligen Bereiche (z.B. Auto, Computer, Lebensmittel) automatisch gewonnen und semantisch analysiert. Durch "AGBV" werden domainspezifische Wörter in den jeweiligen Bereichen extrahiert. Die folgende Grundannahme für domainspezifische Terme wird getroffen:

> Ein Term wird als domainspezifisch betrachtet, wenn er in einem Bereich öfter als andere Terme vorkommt und seltener in anderen Bereichen.
> Ein domainspezifischer Term beinhaltet mindestens einen domainspezifischen Teil als "Elementaren Generischen Term" (EGT, z.B. Wagen, Auto).

Zur Erkennung der KGT (Komplexe Generische Terme) wird die Affix-Anwendung von EGT in dieser Arbeit genannt und gebraucht. Bei der Affixanwendung von EGT gibt es Präfix-, Infix- und Suffixanwendung ähnlich zum Derivationsprozess:

Suffix-	$W = W_1...W_n \quad W_n$ ist ein EGT (z.B. Auto).	Renault-Autos
Präfix-	$W = W_1...W_n \quad W_1$ ist ein EGT.	Autoverkauf
Infix-	$W = W_1...W_n \quad W_{1+1}...W_{n-1}$ beinhaltet ein EGT.	Gebrauchtautomarkt

Einwortterme werden durch EGT mit Hilfe der Affixanwendung erkannt. Dann können Mehrwortterme aus den erkannten bereichsspezifischen Einworttermen schrittweise richtig identifiziert werden. Dieses Verfahren wird in der Arbeit "Bootstrapping-Verfahren mit EGT" genannt. E-Commerce-relevante Webseiten können den jeweiligen Branchen mit Hilfe von EGT (Elementare generische Terme) maschinell zugeordnet werden, was "EGT-Klassifikator" genannt wird. Die Qualität der EGT spielt eine entscheidende Rolle dafür. Das automatisch erkannte branchenspezifische Vokabular durch die "AGBV" in den jeweiligen Branchen ist eine qualifizierte Basis für einen Grundwortschatz, um manuelle Arbeiten für die linguistische Analyse miteinander zu kombinieren.

Kapitel 1

Einleitung

1.1 Zielsetzung

Die Goldsucher lernen, was Gold ist und wie man Gold suchen kann. Dann suchen sie eine Fundstelle mit der besten Qualität, um Gold zu extrahieren. Sie extrahieren das Gold, indem sie es säubern und auswählen. Man braucht für jedes Mineral eine andere Fundstelle. Auf dieser Grundlage will ich domainspezifische Korpora z.B. im E-Commerce-Bereich aufbauen. Anschließend können domainspezifische Terme in den jeweiligen E-Commerce-Bereichen aus den Korpora erkannt und für "Automatische Webseitenklassifikation" angewendet werden.

Durch statistische, linguistische und hybride Verfahren für "Schlüsselwort-Extraktion" (Keyword Extraction) können wichtige Wörter aus einem Text extrahiert werden. Dabei spielt die Worthäufigkeit eine entscheidende Rolle für die "Keyword Extraction". Aber mit den statistischen Verfahren allein kann man die Worthäufigkeit in einem Text nicht exakt genug kalkulieren. Für die korrekte Berechnung der Worthäufigkeit braucht man mindestens noch die folgenden allgemeinen linguistischen Betrachtungen:

- **Eliminierung der Stoppwörter (bzw. nicht sinntragenden Wörter)**
  (der, mit, EUR, kaufen, deutsch, regelmäßig und aktuell, Preis, ...)
- **Lemmatisierung (Stemming, Grundformreduktion)**
  (Autos → Auto, Häuser → Haus, Händler → Händler, ...)
- **Kompositazerlegung**
  (Autohändler → Auto + Händler, Hausvermietung → Haus + vermietung, ...)
- **Phrasen (bzw. Mehrwortbegriffe) erkennen**
  (Information Retrieval, ALFA ROMEO, ...)
- **Linguistische bzw. orthographische Varianten**
  (Gebrauchtwagenmarkt/ gebrauchtwagenmarkt/ GEBRAUCHTWAGENMARKT/ Gebrauchtwagen-Markt/ gebraucht-wagenmarkt/ GEBRAUCHTWAGEN-MARKT/ gebrauchtwagen-markt, ...)

- **Lexika (z.B. CISLEX) und domainspezifische Listen (z.B. Automarken) in den jeweiligen Bereichen**

- **Pronomina-Analysen (Pronomina korrekt zuorden)**
  (Sie sind Luxusautos. Ich möchte gerne eins[1] haben.)

Der Hauptteil dieser Arbeit handelt von linguistischen Verfahren.
Nach der "Keyword Extraction" können wir überlegen, welche Schlüsselwörter domainspezifisch sind. Solche Wörter (z.B. Auto, Fahrzeug, Wagen, Car, BMW, VW) sind branchenspezifisch im Automobilbereich. Aber sie sind nicht domain-spezifisch in anderen Branchen (z.B. Wein, Computer, Musik, Schmuck, Kleidung). Die auf Dienstleistungen bezogenen Wörter (z.B. Verkauf, Tuning, Finanzierung, Verkauf, Verleih) sind domain-neutral.
Auf natürliche Weise versucht man, domainspezifische Schlüsselwörter in den jeweiligen Bereichen zu extrahieren und semantisch zu klassifizieren. Dafür werden domainspezifische Korpora in den jeweiligen Bereichen verwendet.
Wenn man domainspezifische Terme in den jeweiligen Bereichen erkennen kann, können sie für die folgenden signifikanten Anwendungen effizient eingesetzt werden:

- **Erstellung von Fachwörterbüchern**

- **Verbesserung von Suchmaschinen**: z.B. fokussiertes Web-Crawling

- **Verbesserung der maschinellen Übersetzung**

- **Automatische Klassifikation von Webseiten**

Dafür wird in dieser Arbeit die Affixanwendung (Suffix-, Präfix- und Infix-Anwendung) mit den **E**lementaren **G**enerischen **T**ermen (**EGT**), z.B. Auto, Fahrzeug, Wagen, Wein, Rotwein, Handschuhe, Möbel, Jeans verwendet.
Die Qualität der verwendeten EGT für die Affixanwendung ist absolut wichtig für die Erkennung der domainspezifischen Terme im jeweiligen Bereich. EGT können automatisch erweitert werden. Aber die EGT sollten mit nötigen Fachkenntnissen zur Qualitätsverbesserung schließlich manuell verbessert werden.
Die folgenden zwei Zielsetzungen sind die Hauptaufgaben dieser Arbeit:

I. **Erkennung der domainspezifischen Terme im jeweiligen Bereich durch EGT und domainspezifische Listen (z.B. Firmennamen)**

II. **Erstellung des Terminologie-Extraktionssystems AGBV**:
Automatische **G**ewinnung von **B**ranchenspezifischem **V**okabular aus den erstellten Korpora

[1] Indefintpronomen (einer, ein[e]s, ...), Personalpronomen (Sie, ich, er, ...)

1.2 Semantische Analyse

> Ein Wort in einem Text ist für eine Branche entweder ein "branchenspezifisches Wort" oder ein "Stoppwort" (bzw. branchenneutrales Wort).

Ein Wort kann in einer Branche unter der in Kapitel 9.1. genannten Annahme semantisch überprüft werden. Damit können domainspezifische Terme den jeweiligen semantischen Klassen (z.B. Autobranche, Computer) automatisch zugeordnet werden. Das ist die Hauptaufgabe der semantischen Analyse.

1.2.1 Semantische Klassen für einfache Nomina im CISLEX

Die Wortliste, die von Stefan Langer semantisch manuell kodiert wurde, enthält 41.528 Lexeme für einfache Nomina im CISLEX. Die Lexeme werden als Grundform eingetragen. Insgesamt sind sie in 429 semantischen Klassen hierarchisch gegliedert. Davon werden 236 Klassen für E-Commerce manuell ausgewählt. In diesen 236 Klassen werden 23.921 Lexeme, die schon semantisch kodiert wurden, identifiziert. Die 23.921 Lexeme können als "Elementare generische Terme (EGT)" für die Erkennung der DST verwendet werden, wenn sie den entsprechenden E-Commerce-Bereichen richtig zugeordnet sind.

Solche kommerziellen EGT (z.B. Notebook, Laptop, Bildschirm, Keybords, DVD, PC, MP3) sind im CISLEX noch nicht semantisch kodiert, weil sie dort noch nicht vorhanden sind. Neue EGT für E-Commerce müssen erweitert werden. Sie können in zwei typischen Bereichen - Produktnamen und Dienstleistungen - sowie in Sektoren und Branchen semantisch kodiert werden. Damit kann man domainspezifische Terme (DST) im jeweiligen Bereich identifizieren und semantisch analysieren, ob ein Term zu Produktnamen, Dienstleistungen oder zu einer anderen Branche gehört. Die automatisch erstellten Kandidaten für die Erweiterung von EGT in den jeweiligen E-Commerce-Branchen für CISLEX können schließlich zur Qualitätsoptimierung manuell ausgewählt werden.

1.3 Übersicht der einzelnen Kapitel

Es folgt eine kurze Beschreibung der einzelnen Kapitel:

- 1. **Einleitung**
- 2. **Grundlagen der automatischen Terminologie-Extraktion (TE)**
  Eine Übersicht der automatischen Terminologie-Extraktion wird beschrieben. Die statistischen Verfahren 'Zipfsches Gesetz', 'TF-IDF-Gewichtung' und 'N-Gramme' werden als Grundlagen für diese linguistische Arbeit vorgestellt. In den statistischen Verfahren wird die Betrachtung der linguistischen Eigenschaften fast nicht eingesetzt. Der "Porter-Stemmer-Algorithmus" (Porter 1980) für die Lemmatisierung im Englischen wird erklärt.

- 3. **TE domainspezifischen Vokabulars mittels eines Vergleichs von Korpora**
  Allgemeine Korpora (bzw. nicht-technische Korpora) werden zur Entfernung der branchenneutralen Wörter (z.B. Umsatz, Baubeginn) verglichen, um domainspezifisches Vokabular in den jeweiligen Bereichen zu erkennen. Die drei wichtigen Anwendungen von TE domainspezifischen Vokabulars, nämlich die Erstellung von Fachwörterbüchern, die Verbesserung von Suchmaschinen (fokussiertes Web-Crawling) und die Verbesserung der maschinellen Übersetzung werden dargestellt.

- 4. **Domainspezifische Terme (DST) und ihre Relationen**
  Die Definitionen und Konventionen für "Einwort- und Mehrwortterme", "Elementare generische Terme (EGT)" und "Komplexe generische Terme (KGT)" werden vorgestellt. Die folgende Grundannahme für domainspezifische Terme im E-Commerce-Bereich wird in dieser Arbeit verwendet:

  > Ein Term wird als domainspezifisch betrachtet, wenn er in einem Bereich öfter als andere Terme vorkommt und seltener in anderen Bereichen. Ein domainspezifischer Term beinhaltet mindestens einen domainspezifischen Teil als "Elementaren Generischen Term (EGT)".

  'KFIDF' ist eine Modifikation von "TFIDF (term frequency inverted document frequency)". 'KFIDF' ist für schon kategorisierte Dokumente besser geeignet als das TFIDF-Maß, um domain-relevante Einwortterme automatisch aufzufinden. In dieser Methode wird das sogenannte Ontologie-Netz "GermaNet" für lexikal-semantische Informationen angewendet, um semantische Relationen zwischen extrahierten Termen zu erkennen.

- 5. **TE domainspezifischen Vokabulars aus einer Webseite**
  In dieser Arbeit werden domainspezifische Terme aus sechs verschiedenen Quellen innerhalb einer Webseite extrahiert. Zur Erkennung der KGT wird die Affix-Anwendung von EGT in dieser Arbeit genannt und gebraucht. Bei der Affixanwendung von EGT gibt es Präfix-, Infix- und Suffixanwendung ähnlich zum Derivationsprozess:

Suffix-	$W = W_1...W_n \quad W_n$ ist ein EGT (z.B. Auto).	Renault-Autos
Präfix-	$W = W_1...W_n \quad W_1$ ist ein EGT.	Autositze
Infix-	$W = W_1...W_n \quad W_{1+1}...W_{n-1}$ beinhaltet ein EGT.	Gebrauchtautomarkt

  Die zwei CGI-Programme, nämlich "CGI-Programm 1 mit sechs verschiedenen Quellen und N-Grammen" und "CGI-Programm 2 mit Unitex, Bootstrapping-Verfahren und phpMyAdmin (MySQL)" werden von mir erstellt, um domainspezifische Terme aus einer Seite ohne Vergleich mit Korpora zu erkennen und in Datenbanken zu speichern. E-Commerce-relevante Webseiten können den jeweiligen Branchen mit Hilfe von EGT maschinell zugeordnet werden.

- 6. **Domainspezifische Korpora aus dem Web**
  Das **Ziel der Korpora aus dem Web** ist, dass domainspezifische Korpora für die deutsche Sprache aus dem Web automatisch erstellt werden, um Einwortterme und Mehrwortterme (bzw. Phrasen) zu erkennen und zu erweitern.
  Folgende zwei Methoden für die Dokumentensammlung bzw. URL-Sammlung können angewendet werden:

  a. Extraktion aus Startseiten (z.B. www.autoscout24.de, www.vodafone.de)

  b. Extraktion mit Suchmaschinen (z.B. Google, Yahoo)

  Die Schwierigkeiten beim Aufbau der Korpora (z.B. Duplikate, komprimierte Webseiten, Cookie-Seite) werden aufgrund meiner empirischen Untersuchung erwähnt. Die Erkennungsmethoden für Einwortterme werden im Automobilbereich als Experiment durchgeführt. Für die semantische Annotation der Einwortterme in der Autobranche werden EGT, domainspezifische Listen (z.B. Automarken, Abkürzungen, Automodelle) und drei Korpora (Schmuck, Wein, Kleidung) als "Background Filter" zur Entfernung der unnötigen Wörter verwendet. Die Normalisierung der Terme wird vorgestellt.

- 7. **Extraktion der Mehrwortterme in NLP**
  Dafür werden "Lokale Grammatiken mit Unitex", LEXTER (Bourigault, 1994), FASTR (C. Jacquemin), "Mustererkennung in Perl"und "N-Gramme mit Wortfolgen" in dieser Arbeit vorgestellt.

- 8. **Erkennung der Produktterme (PT) für E-Commerce**
  Wegen der allgemeinen Unterscheidung zwischen Eigennamen und Appellativa werden die auf Produkte bezogenen Terme in dieser Arbeit als Produktterme (PT) bezeichnet, z.B. Tempo, Rama, Margarine, Handschuhe, Lederhandschuhe. Es geht um die Erkennung der PT. Die Struktur und die semantischen Merkmale der PT werden vorgestellt. Zur Erkennung der PT spielt die Affixanwendung von EGT eine entscheidende Rolle. Das CGI-Programm (Hierarchieextraktor) für die Extraktion der hierarchischen Struktur von Produkttermen wird von mir erstellt. "Semantische Klassen für E-Commerce im CISLEX" und "Erknennung der auf Dienstleistungen bezogenen Terme" werden vorgestellt.

- 9. **Automatische Gewinnung von branchenspezifischem Vokabular (AGBV)**
  In diesem Kapitel werden die folgenden Zielsetzungen experimentell durchgeführt:

  i. **Automatische Erstellung der deutschen Korpora für E-Commerce-Branchen**

  ii. **Erstellung des Terminologie-Extraktionssystems AGBV**:
  Automatische **G**ewinnung von **b**ranchenspezifischem **V**okabular aus den erstellten Korpora

Die Überprüfung der erkannten Wörter in einer Branche wird vorgestellt. Zur 'AGBV' werden 20 E-Commerce-Branchen für "Test 1" und "Test 2" ausgewählt. Die zwei Masterprogramme zum Aufbau der Korpora, nämlich "Extraktion mit Suchmaschinen" und "Extraktion aus Startseiten" werden verbessert. Das automatisch erkannte branchenspezifische Vokabular durch die **AGBV** kann als Basis für einen Grundwortschatz in den jeweiligen E-Commerce-Branchen sehr effizient benutzt werden. In dem von mir erstellten CGI-Programm "**AGBV aus einer Webseite**" werden die 20 automatisch erstellten Korpora verglichen, um domainspezifische Terme aus einer Webseite zu erkennen.

- 10. **Vergleich mit allgemeinen Korpora für AGBV**
  Im Test wurden insgesamt 25 normalisierte Datenbanken aus den 20 ausgewählten E-Comerce-Branchen, den vier zusätzlichen allgemeinen Korpora (Bibel, Politik, Gedicht, Zeitung) und "www.vodafone.de" für die Entfernung der branchenneutralen Wörter verwendet. Trotzdem bleiben unnötige und branchenneutrale Wörter (z.B. Hilfe, Kontakt) übrig. Bei der manuellen Auswahl können solche branchenneutralen Wörter gesammelt und eliminiert werden.

- 11. **Zusammenfassung und Ausblick**

Kapitel 2

Grundlagen der automatischen Terminologie-Extraktion (TE)

"Data Mining" beschäftigt sich mit strukturierten Datenbanken (structured databases). Aber es gibt große Mengen von Informationen in unstrukturierter natürlichsprachlicher Form, wie z.B. aus Webseiten oder elektronischen Texten, die sehr schnell wachsen. Bei "Text Mining" handelt es sich um Bearbeitungstechniken, die aus solchen Datenmengen Informationen suchen. "Automatische Terminologie-Extraktion", die durch verschiedene statistische und musterbasierte Methoden erreicht wird, spielt eine wichtige Rolle im Bereich von Text Mining (TM), Information Retrieval (IR), Information Extraction (IE), Natural Language Processing (NLP) und Machine Learning (ML).
Heutzutage gibt es zahlreiche aktive Forschungen und Projekte für die Extraktion der signifikanten Terme und domain-spezifischen Fachbegriffe aus z.B. Fachtexten und Webseiten der jeweiligen Domäne. Wegen der hohen Kosten und Aktualisierungsprobleme der neuen Daten kann man "manuelle Terminologie-Extraktion" von Experten nicht mehr leisten. Eine ideale Kombination ist, dass zunächst automatisch erstellte domain-spezifische Terme in den jeweiligen Bereichen zur Qualitätsverbesserung zusätzlich manuell verbessert werden können.

2.1 Definition von Terminologie und Term

Die folgenden zwei Definitionen von Terminologien werden erwähnt:

- ACHMANOVA (1966) definiert Terminologien als die "Gesamtheit der Termini eines bestimmten Produktionszweiges, Tätigkeitsbereichs oder Wissenschaftsgebietes, die einen besonderen Sektor (eine besondere Schicht) der Lexik bilden, der sich am ehesten bewußt regulieren und ordnen läßt".
  [Sch92, S. 230]
- Als Terminologie[1] wird die Gesamtheit aller Begriffe und Benennungen (Fachwörter

[1] Stand: 24.07.2007 - de.wikipedia.org/wiki/Terminologie

bzw. Termini) einer Fachsprache bezeichnet.

Eine Terminologie - das System der Termini einer Wissenschafts oder Fachsprache - ist die Gesamtheit aller Begriffe und Benennungen (Termini) einer Fachsprache.
Das systematische Sammeln von Fachausdrücken und Nomenklaturen[2] in einer oder mehreren Sprachen ist die wichtigste Aufgabe.

Ein Term ist ein spezifischer Begriff (bzw. Fachbegriff) in einem Bereich. Die Gesamtheit der bereichsspezifischen Terme ist die Terminologie in einem Bereich. Innerhalb einer Terminologie wird ein Term wie folgt definiert:

> In terminology (the study of language terms), a "**term**[3]" is a word, word pair, or word group, that is used in specific contexts for a specific meaning.

Terme beinhalten wissenschaftliche und technische Informationen.
Otman(1995) unterscheidet zwei Typen von Termen im Wörterbuch [Jac01, S. 10]:

- **Technical terms**, which denote instruments, artifacts, observations, experiments, measures.
- **Scientific terms**, which denote theoretical concepts in scientific doamins.

In diesem Kapitel werden international verbreitete Methoden für automatische Terminologie-Extraktion, die für diese ganze Arbeit nötig sind, zusammengefasst.
Im dritten Kapitel wird automatische Terminologie-Extraktion zum domain-spezifischen Aspekt mittels Vergleich von Korpora dargestellt.

2.2 Automatische Verfahren

Bereichsspezifische Terme können durch die "Automatische Terminologie-Extraktion" maschinell erkannt werden. Es gibt schon bekannte bestehende Ansätze zum Thema der Terminologie-Extraktion bzw. des automatischen Indexierens (Automatic Indexing).
Die Suche nach guten Termen bzw. Indextermen ist die wichtigste Aufgabe, um unstrukturierte natürlichsprachliche Dokumente zu analysieren. Am wichtigsten ist "Automatische Indexierung" im Bereich des "Information Retrieval", das heißt für die Auffindung der relevanten Dokumente aus einer Dokumentensammlung. Die folgenden drei automatischen Verfahren gelten sowohl für "Terminologie-Extraktion" als auch für "Automatische Indexierung". Dabei geht es um die Berechnung der Worthäufigkeit von Wörtern bzw. Wortgruppen:

- **Statistische Verfahren**: Zipfs Gesetz, TF, IDF, TF-IDF, Cosinus-Maß, N-Gramme
- **Linguistische Verfahren**: Lemmatisierung, Phrasenerkennung, POS-Mustern
- **Hybride Verfahren**: GERHARD

[2] Die Systematik einer Namensgebung (Benennung) in einem bestimmten Fachgebiet.
[3] Stand: 24.07.2007 - en.wikipedia.org/wiki/Term

2.3 Statistische Verfahren

> The "significance" factor of a sentence is derived from an analysis of its words. It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance [Luh58, S. 160].

In den 50er Jahren haben Wissenschaftler wie LUHN, SALTON und SPARCK JONES mit statistischen Verfahren experimentiert.
Die Worthäufigkeit eines Wortes im Text ist ein gutes Maß für "wichtige Wörter", genannt Indexterme oder Schlüsselwörter. Dabei werden unterschiedliche Berechnungen für die Worthäufigkeiten aus den statistischen Verfahren ohne linguistische Überlegungen behandelt.

2.3.1 Zipfsches Gesetz

Von dem amerikanischen Philologen G.K. Zipf wird das bekannte sogenannte Zipfsche Gesetz formuliert. Das Gesetz von Zipf lautet wie folgt:

$$r * f = c$$

r (Rang eines Wortes in einer Frequenzliste), f (Frequenz in einem Text), c (eine konstante Beziehung zwischen "r" und "f" / constant)

Dabei wird der umgekehrte Zusammenhang zwischen Länge und Frequenz eines Wortes betrachtet. Die am häufigsten gebrauchten Wörter sind meist sehr kurze und inhaltsleere Funktionswörter (z.B. Artikel, Konjunktionen, Präpositionen, Adverbien, Personalpronomen, Hilfsverben). Diese sind sogenannte Stoppwörter (stop words), die nicht als Indexterme gebraucht werden können. Im Deutschen sind das zum Beispiel "der", "die", "und", "oder", im Englschen "the", "a", "is".
Zipf stellt eine Verteilung auf, in der die Wörter nach ihrer Häufigkeit geordnet werden und zwei Grenzen festgelegt werden. Diese nennt er "upper cut-off" und "lower cut-off". Häufige Terme, deren Rang links der "upper cut-off" liegt, werden meist als Stoppwörter bzw. nicht signifikante Terme betrachtet. Seltene Terme, deren Rang der Wörter rechts der "lower cut-off" liegt, sind als Indexterme ebenfalls nicht geeignet, weil sie in Anfragen wenig benutzt werden. Dazwischen liegen nach dem "Zipfschen Gesetz" die signifikanten Terme (significant words).

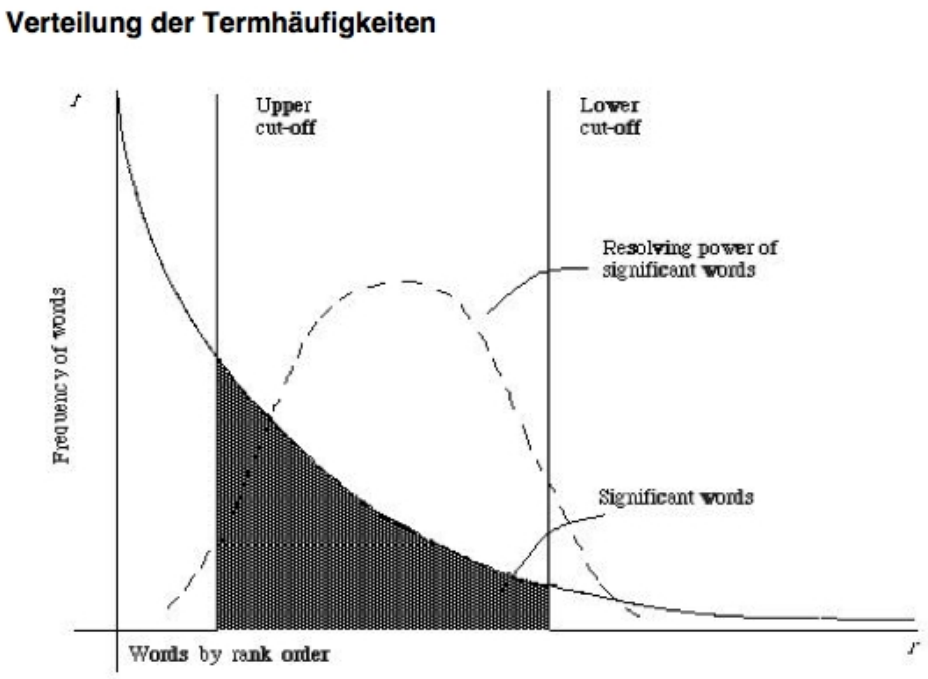


Abbildung 2.1: Verteilung der Termhäufigkeiten nach dem "Zipfschem Gesetz"

2.3.2 TF-IDF-Gewichtung

Nachdem Stoppwörter eliminiert wurden, ist die Worthäufigkeit ein wichtigr Faktor für die "Term Gewichtung" (term weighting).
Wenn das Wort T öfter in einem Dokument und selten in anderen Dokumenten vorgekommen ist, wird das Wort T als ein signifikanter Term betrachtet und bewertet. Das ist der Grundgedanke von 'TF-IDF-Gewichtung'.
Die Termfrequenz (term frequency) wird allgemein verwendet, um das Gewicht eines Wortes für ein Dokument zu ermitteln.

$$tf = \frac{n_i}{\sum_k n_k}$$

n_i (Häufigkeit eines Wortes im Dokument)
$\sum_k n_k$ (Anzahl aller Wörter des Dokuments)

Die inverse Dokumentenhäufikeit (inverse document frequence / IDF) wird gebraucht, um signifikante Terme, die in möglichst wenigen Dokumenten vorkommen, zu finden. Damit kann man die Worthäufigkeit in einer Dokumentensammlung berechnen.

$$idf = \log \frac{|D|}{|(d_i \supset t_i)|}$$

$|D|$ (total number of document in the corpus)
$|(d_i \supset t_i)|$ (number of documents where the term t_i appears - that is $n_i \neq 0$.)

Die TF-IDF-Gewichtung ($tfidf = tf * idf$) wird oft für "Information retrieval" und "Text Mining" verwendet, um die Gewichtung der signifikanten Terme zu berechnen. Im bekannten experimentellen System SMART (Salton und McGrill, 1983) wurde die TF-IDF-Gewichtung mit Cosinus-Maß als Ähnlichkeitsmaß (Skalarprodukt) [SB87, S. 3] erfolgreich eingesetzt, um relevante Dokumente in Anfragen zu ermitteln (Relevanz-Feedback[4]).

2.3.3 Vektorraummodell

Das Vektorraummodell (engl.: Vector Sprace Model (VSM)) wurde Anfang der 70er Jahre im Rahmen des SMART[5]-Projektes im Bereich von "Information Retrieval" entwickelt. Im SMART-System werden Anfragevektoren mit Dokumentvektoren mittels Ähnlichkeitsmaßen verglichen. Das einfachste Ähnlichkeitsmaß ist das folgende Cosinusmaß:

Ähnlichkeitsmaß: Cosinus eines Winkels zwischen zwei Vektoren

$$\frac{\vec{X}\cdot\vec{Y}}{|\vec{X}|\cdot|\vec{Y}|} \;\Big|\; \cos(x,y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\cdot\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Wertebereich: [-1 (Winkel:180°); 1 (Winkel:0°)]

Je kleiner der Winkel zwischen zwei Vektoren, desto größer der Ähnlichkeitswert. Die Skalarprodukte (z.B. $\vec{X} \cdot \vec{Y}$) sind die Länge der Vektoren

2.3.4 N-Gramme

Ein N-Gramm ist die Zeichenfolgen einer Länge N. Die N-Gramme z.B von "Wagen" sind die folgenden:

N-Gramm-Namen	**Wagen**
Unigramm (Monogramm)	w
Bigramm	wa
Trigramm	wag
Tetragramm	wage
Pentagramm	wagen

Wichtige N-Gramme sind das Uni-, Bi- und, Tri-Gramm. Zeichenfolgen werden in 'N-Gramme' mit einer bestimmten Anzahl von Zeichen zerlegt (z.B.

[4]Relevanz-Feedback (relevance feedback) ist eine Technik der "query expansion", die der Erweiterung einer Suchanfrage dient.

[5]SMART steht für "System for Mechanical Analysis and Retrieval of Text"

Trigramme / N = 3 : Wagen → wag, age, gen). Für die Korrektur von Tippfehlern ist dies effizient.
Nicht als Zeichenfolge sondern als Wortfolgen können diese N-Gramme in Kapitel 5.6.2.2. "N-Gramme mit Wortfolgen für Mehrwortterme" berücksichtigt und eingesetzt werden.

2.4 Linguistische Verfahren

In den statistischen Verfahren wird die Betrachtung der linguistischen Eigenschaften fast nicht eingesetzt. Vor einer statistischen Berechnung für korrekte Worthäufigkeit müssen alle Wörter auf ihre Grundform reduziert werden (Grundformreduktion). Die linguistischen Verfahren bemühen sich meist um Lemmatisierung (stemming) eines Wortes, um die Erkennung der Mehrwortterme und um die Relationen zwischen Termen mit Hilfe einer morphologischen, syntaktischen und semantischen Analyse. Sie haben die folgenden typischen Aufgaben nach Stock, W.G. [Sto58, S. 23]:

- **Eliminierung der Stoppwörter**
  (der, mit, EUR, kaufen, deutsch, regelmäßig und aktuell, Preis, ...)

- **Lemmatisierung (Stemming, Grundformreduktion)**
  (Autos → Auto, Häuser → Haus, Händler → Händler, ...)

- **Kompositazerlegung**
  (Autohändler → Auto + Händler, Hausvermietung → Haus + vermietung, ...)

- **Phrasen (bzw. Mehrwortbegriffe) erkennen**
  (Information Retrieval, ALFA ROMEO, ...)

- **Pronomina-Analysen (Pronomina korrekt zuorden)**
  (Sie sind Luxusautos. Ich möchte gerne eins[6] haben.)

[6] Indefintpronomen (einer, ein[e]s, ...), Personalpronomen (Sie, ich, er, ...)

2.4.1 Automatische Lemmatisierung

Wie ich oben erwähnt habe, ist die Worthäufigkeit ein wichtiges Maß für die Extraktion der signifikanten Terme. Diese können in verschiedenen Formen im natürlichsprachlichen Text auftauchen.
Automatische Lemmatisierung (automatic stemming) muß für die statistischen Verfahren vorher durchgeführt werden, um Worthäufigkeit exakt zu berechnen. Die Tabelle 2.2 zeigt die Notwendigkeit genauerer Berechnung der Worthäufigkeit. Aus einer Wortliste im Bereich "Autobranchen" werden die Top34-Terme, die ohne Manipulation den Ausdruck "Auto" beinhalten, nach den Worthäufigkeiten sortiert. Die 4 Terme (Auto, Autos, auto, autos) sollten zum gleichen Lemma "Auto" zusammengefasst werden. Die jeweilige Worthäufigkeit für dasselbe Lemma "Auto" muss addiert werden, um die Worthäufigkeiten korrekt zu berechnen. "Auto" und "Automobil" sollen als Synomym behandelt werden und ihre Worthäufigkeiten addiert werden.
Die Groß-/Klein-Schreibung der Wörter soll nicht als großes Problem behandelt werden. Aber in einer Sprache können Wörter in verschiedenen Formen (Einzahl/Mehrzahl, Konjugation, Deklination) geschrieben werden. Sie sollten entsprechend automatisch lemmatisiert und ermittelt werden. Außerdem gibt es auch viele Varianten der Wörter, z.B. mit Bindestrich zusammengesetzte Wörter (z.B. Auto-Reparatur für Autoreparatur) oder mit deutschen Umlauten (z.B. Autohändler für Autohaendler).
Ein Lemmatisierer, der das Lemma des jeweiligen Wortes ermittelt, sollte solche Variationen verstehen und erkennen. Bei dieser Grundformreduktion gibt es grundsätzlich zwei verschiedene Verfahren:

- **Regelbasierte Verfahren**
  Z.B. Lovins-Stemmer (Lovins 1968) und Porter-Stemmer(Porter 1980)

- **Wörterbuch-basierte Verfahren**
  Z.B. EuroWordNet: http://www.illc.uva.nl/EuroWordNet
  GermanNet: www.sfs.uni-tuebingen.de/lsd
  WordNet: wordnet.princeton.edu/

Regelbasierte Verfahren / Porter-Stemmer

Porter-Stemmer-Algorithmus (Porter 1980)

Für die Lemmatisierung der englischen Wörter werden der Lovins-Stemmer-Algorithmus (Lovins 1968) und der Porter-Stemmer-Algorithmus (Porter 1980) häufig gebraucht. Der Porter-Algorithmus beschäftigt sich nicht mit einer 100% richtigen linguistischen Grundformreduktion, sondern mit einer effizienten Berechenbarkeit der Grundformreduktion.

> Stemming algorithms perform basically two operations: suffix stripping and recoding. The suffix stripper removes from each word a word ending that is expected to be the longst suffix [Jac01, S. 17].

Die beiden oben erwähnten Stemming-Transformationen - suffix stripping and recoding - werden beim Porter-Stemmer-Algorithmus im Gegensatz zum Lovins-Stemmer-Algorithmus gleichzeitig ausgeführt.

Der Porter-Stemmer-Algorithmus besteht aus den folgenden 5 Transformationsregeln, die schrittweise durchgeführt werden.

Step	Condition	Input/output	Sample input/output
1	–	-ies → -i	ponies → poni
1	*v*	-y → -i	pony → poni
2	m > 0	-ational → -ate	relational → relate
3	m > 0	-icate → -ic	triplicate → triplic
4	m > 1	-ate → ∅	activate → activ
5	m > 1	-e → ∅	probate → probat

Tabelle 2.1: Some sample rules of Porter's algorithm [Jac01, S. 18]

Die Bedingung "*v*" bedeutet, daß der Stamm einen Vokal beinhalten muß. Die Bedingung "$m > \alpha$" ist ein Maß für die Anzahl der Konsonant-Vokal-Gruppen. Jedes Wort hat die folgende "Single-Form":

> [C](VC)m[V]
> C (consonat), V (vowel), [...] (optional), (VC)m (VC wiederholt m-mal.) [Por80, S. 132]

Der Porter-Stemmer assoziiert "deny" mit dem Stamm "deni", obwohl der Stamm "deny" ist. Er ermittelt nicht den linguistisch richtigen Stamm, sondern einen Pseudostamm. Er bemüht sich nur um effiziente Berechenbarkeit für Grundformreduktion ohne Wörterbuch.

Wörterbuch-basierte Verfahren

Regelbasierte Verfahren sind für Englisch angemessen geeignet, aber für Sprachen wie Deutsch, die stark konjugieren und deklinieren, nicht geeignet, um Grundformreduktion automatisch durchzuführen. Deswegen versuchen "Wörterbuch-basierte Verfahren" (bzw. lexikonbasierte Grundformreduktion) eine linguistisch korrekte Grundformreduktion mit Hilfe eines elektronischen Wörterbuchs zu ermitteln.
Lexikalische semantische Wortnetze - WordNet, GermaNet und EuroNet - stellen lexikalisch-semantische Informationen für Linguistische Verfahren zur Verfügung. Die Tabelle 2.2 zeigt die Notwendigkeit exakter Berechnung der Worthäufigkeit mittels "Automatischer Lemmatisierung":

Tabelle 2.2: **Notwendigkeit der automatischen Lemmatisierung**

Fre.	Terme	Lemma
3193	Autohaus	Autohaus
1654	Auto	Auto
741	Autos	Auto
684	autohaus	Autohaus
624	Autohandel	Autohandel
618	auto	Auto
564	Autohändler	Autohändler
462	Automobil	Automobil
449	Automarkt	Automarkt
386	Automobile	Automobil
386	Autohäuser	Autohaus
351	Autovermietung	Autovermietung
343	Autobörse	Autobörse
337	Autoreparatur	Autoreparatur
316	Autoservice	Autoservice
259	Autofinanzierung	Autofinanzierung
252	Autotuning	Autotuning
245	Gebrauchtautos	Gebrauchtauto
240	Autobranche	Autobranche
226	Leasingautos	Leasingauto
184	autos	Auto
181	Autozubehör	Autozubehör
177	Autowerkstatt	Autowerkstatt

Tabelle 2.2: **Notwendigkeit der automatischen Lemmatisierung**

Fre.	Terme	Lemma
126	automobile	Automobil
117	Autokauf	Autokauf
109	Autoteile	Autoteil
108	Autoreparaturen	Autoreparatur
88	autohandel	Autohandel
83	Autoverkauf	Autoverkauf
79	automobil	Automobil
75	automarkt	Automarkt
60	autozubehör	Autozubehör
57	autohändler	Autohändler
54	Autohaendler	Autohändler

2.4.2 Mehrwortgruppenerkennung / Phrasenerkennung

Die Mehrwortgruppenerkennung (Phrasenerkennung) ist die Hauptaufgabe für linguistische Verfahren. Im Artikel "Natürlichsprachige Suche" von W.G. Stock [Sto58, S. 23] wird eine Phrasenerkennung wie folgt erklärt:

> Eine Phrase ist ein Ausdruck, der aus mehreren einzelnen Wörtern besteht. Hier gilt nicht das einzelne Wort (oder dessen Wortstamm) als Schlagwort, sondern die Phrase als Ganzes.

Ein System für die Phrasenerkennung sollte z.B. "Information Retrieval" und "Alfa Romeo" als eine Einheit erkennen. Für die Suche nach Mehrwortgruppen (bzw. Phrasen) werden die folgenden vorhandenen Methoden im Bereich der Computerlinguistik oft verwendet:

- **POS-Muster** mit Hilfe von POS-Taggern (Part-of-speech)
  'NPtool' von Arppe [Arp95, S. 5]
- **NLP-Techniken (Natural language processing)**
  LEXTER, FASTR und "Lokale Grammatiken mit Unitex"

Um Mehrwortgruppen zu erkennen, werden in LEXTER (Bourigault, 1994) kategorisierte Texte in maximalen Nominalphrasen durch endliche Automaten zerlegt. Nach W.G. Stock wird ein Text in Textklumpen, die zwischen Stoppwörtern oder Satzzeichen stehen, zerlegt. Stoppwörter und Satzzeichen

werden als Begrenzer für Klumpen (Chunks) benutzt.
Die wichtigsten Mehrwortgruppen kommen aus Nominalphrasen und Nominal-Chunks. Die Suche in wichtigen Nominalphrasen und Nominal-Chunks spielt eine große Rolle dabei. In Kapitel 7 "Extraktion der Mehrwortterme in NLP" werden unterschiedliche Techniken vorgestellt.

2.5 Hybride Verfahren

Bei den statistischen Verfahren handelt es sich um die Berechnungen der Häufigkeiten von Wörtern ohne linguistische Aspekte (z.B. Lemmatisierung, Mehrwortlexeme und semantische Relationen zwischen Wörtern). Bei den hybriden Verfahren handelt es sich um die Berechnungen von Worthäufigkeiten mit linguistischen Aspekten.
Hybride Verfahren in "Automatische Terminologie-Extraktion" (TE) werden mit statistischen und linguisitischen Verfahren kombiniert, um sinnvolle Schlüsselwörter in einem Text zu ermitteln.
Ein Beispiel ist die Spezial-Suchmaschine 'GERHARD' für deutsche wissenschaftliche Webseiten im Rahmen eines DFG[7]-Projekts (1996-1998). GERHARD (GERman Harvest Automated Retrieval and Directory) entstand 1996 an der Universität Oldenburg mit dem Ziel, eine flächendeckende, roboterbasierte Suchmaschine für den deutschsprachigen Raum zu entwickeln. Die automatische Indexierung in 'GERHARD' basiert auf hybriden Verfahren. Danach versucht GERHARD eine automatische Klassifikation anhand der UDK (Universal Dezimal Klassifikation). Z.B. würde dann der Themenbereich "Umwelt und Frauen" in der UDK (mit Nummer 396,5.000.504) in dem erzeugten Lexikon folgendermaßen repräsentiert:

`Umwelt#Frauen#:396,5.000.504`
`Das # (Trunkierungssymbol)` ist dabei ein Kennzeichen für das Wortende.

Der Zahlenwert symbolisiert die Klassenzuordnung.
Je länger die Notation, desto spezifischer bzw. genauer ist die Zuordnung.

[7]DFG - Deutsche Forschungsgemeinschaft

Kapitel 3

TE domainspezifischen Vokabulars mittels eines Vergleichs von Korpora

Im vorherigen Kapitel "Automatische Terminologie-Extraktion" (TE) wurden grundlegende international bekannte Techniken erklärt. Dabei handelt es sich um die Extraktion von wichtigen Termen, genannt Indexterme oder Schlüsselwörter, aus einem Dokument oder einer Dokumentensammlung. Die automatische TE ist die Basis für Indexierung und Klassifikation.
In diesem Kapitel werden die schon vorhandenen Techniken für "TE domainspezifischen Vokabulars mittels eines Vergleichs von Korpora" vorgestellt.
Der neue Trend ist die Nutzung des Korpora-Vergleichs, um domainspezifische Terme in den jeweiligen Bereichen zu erkennen. Meine eigenen Methoden dazu werden in weiteren Kapiteln, besonders in Kapitel 9, vorgestellt.

Die Suche nach Schlüsselwörtern wird zum domain-spezifischen Aspekt hin geändert und entwickelt, um bessere Schlüsselwörter in einem Dokument bzw. Bereich zu identifizieren. Nicht alle Schlüsselwörter sind domainspezifisch. Die Erkennung der domainspezifischen Terme ist die Hauptaufgabe im Bereich "Domain Specific Terminology", um z.B. domainspezifische Lexika in den jeweiligen Bereichen zu erstellen.
Bei der automatischen Erkennung der domainspezifischen Terme in den jeweiligen Bereichen ist der Vergleich von Korpora sehr nützlich.
Durch den Vergleich der Korpora können unwichtige Wörter unter den Termkandidaten identifiziert und entfernt werden. Diese Idee liegt dieser Arbeit zugrunde. In diesem Kapitel werden die folgenden vier unterschiedlichen Ver-

gleichsmethoden von Korpora, die bisher schon verwendet wurden, vorgestellt:

a. **Technisches Korpus versus Nicht-technisches Korpus**

b. **Allgemeines Korpus als "Background Filter"**

c. **Schlüsselwortextraktion zwischen Korpora**

d. **Ähnlichkeit zwischen Korpora**

3.1 Vorherige Arbeiten ohne Vergleich von Korpora

Die folgenden bekannten vorherigen Arbeiten ohne Vergleich von Korpora beschäftigen sich mit der Berechnung der Worthäufigkeiten von Wörtern und Wortgruppen in einem Korpus bzw. einer Dokumentensammlung.

- **TF-IDF-Gewichtung**: Der Grundgedanke von 'TF-IDF-Gewichtung' basiert auf statistischen Verfahren (Term Frequency, Inverse Document Frequence).

- **Mutual Information (MI)**: I(x,y) = $\log_2 \frac{P(x,y)}{P(x)P(y)}$
  (where $P(x, y)$ denotes the joint probability and $P(x)$ and $P(y)$ denote the probability of x and y separately.)
  Beispiel: Das Korpus enthält 387267 Wörter. Der Ausdruck "Mutual Information" kommt 28 mal vor. Der Ausdruck "Mutual" kommt 134 mal vor. Der Ausdruck "Information" kommt 567 mal vor. Die "Mutual Information" im Korpus wird in Perl wie folgt berechnet:
  4.96087506310197 = log ((28/387267)/((134/387267)*(567/387267)))

- **LogLike (x, y)** $= a \log a + b \log b + c \log c + d \log d$
  $- (a+b) \log(a+b) - (a+c) \log(a+c)$
  $- (b+d) \log(b+d) - (c+d) \log(c+d)$
  $+ (a+b+c+d) \log(a+b+c+d)$

Der Begriff 'Kollokation' ist im Bereich von "Automatische Terminologie-Extraktion" allgemein wie folgt definiert.

> Kollokation ist ein statistisch assoziiertes Wortpaar - d.h. der Begriff wird synonym zu Assoziationspaar verwendet (etwa in Quasthoff/Wolff 2002)

Üblicherweise werden öfter gemeinsam auftretende Wortpaare (z.B. Information Retrieval) als Kollokationen in einem Textkorpus bezeichnet. Die oben erwähnten zwei Assoziationsmaße "Mutual Information (MI)" und "Log-Likelihood" zur Berechnung von Kollokationen werden oft bevorzugt.
Die am häufigsten untersuchten Phänomene für assoziierte Wortpaare beschäftigen sich mit seltenen Ereignissen im Korpus. Im Gegensatz zu anderen Assoziationsmaßen, z.B. MI, T-Score (T-Test), Chi-Quadrat-Test ist das Assoziationsmaß "Log-Likelihood" von Dunning (1993) für die Berechnung der Signifikanz seltener Ereignisse besonders geeignet.

Schlußfolgerung
Die statistischen Berechnungen von Worthäufigkeiten und Kollokationen werden ohne Vergleich von Korpora verwendet, um wichtige Terme in einem Textkorpus zu erkennen. Die dadurch erkannten Terme beinhalten viele unwichtige Wörter. Um solche unwichtigen Wörter zu entfernen, werden verschiedene Korpora verglichen.
Diese Arbeit stützt sich sowohl auf die grundlegenden Ansätze der statistischen Berechnungen, als auch auf die unterschiedlichen Nutzungen von Korpora für die Erkennung der domain-spezifischen Einwortterme (bzw. uniterms). Für die domainspezifischen Mehrwortterme (z.B. Information Retrieval) werden linguistische Verfahren, z.B. NLP-Techniken (Natural Lauguage Processing) oder lokale Grammatiken verwendet. Die Definitionen von Kollokation und Mehrworttermen sind nicht gleich. In dieser Arbeit wird der Begriff "Mehrwortterme" als Synonym von "complex terms" benutzt.

3.2 Technisches Korpus versus Nicht-technisches Korpus

Hier wird die Technik von Patrick Drouin mit dem Artikel "Detection of Domain Specific Terminology Using Corpora Comparision (2004)" [Dro04] vorgestellt. Die Technik arbeitet mit dem Vergleich der Korpora, um domain-

spezifische Terme, z.B. im Bereich Telekommunikation, zu identifizieren. Dafür wird ein technisches Korpus mit einem nicht-technischen Korpus verglichen, um unwichtige Wörter aus automatisch erkannten Termkandidaten zu entfernen. Dieses nicht-technische Korpus besteht aus 13736 Artikeln der englischen Zeitung 'Gazette'. Die Anzahl der verschiedenen Wörter (Word forms) ist ca. 82700. Alle verwendeten Korpora wurden zuerst tokenisiert und durch "Brill's rule-based part-of-speech tagger (Brill 1992, 1994)" getaggt. Alle Nomen als "headwords" wurden zuerst erkannt. Die Termextraktion wird mit "headwords" begonnen und im Korpus von rechts nach links analysiert. Durch den folgenden regulären Ausdruck wurden domain-spezifische Termkandidaten, die mindestens ein Nomen und maximal 6 Wörter beinhalten, automatisch erkannt:

(A|N)?(A|N)?(A|N)?(A|N)?(A|N)?N

where:
A is an adjective,
N is a noun,
(A|N) is a noun or an adjective,
? represents zero or one occurrence fo the element,
____ is an element that belongs to the SLP set. (Specialized Lexical Pivot)

Tabelle 3.1: Regulärer Ausdruck für 'TermoStat' von Patrick Drouin

Das Termextraktionssystem 'TermoStat' von Patrick Drouin ist eine neue hybride Termextraktionstechnik für technische Korpora. Dadurch werden Einwort- und Mehrwortterme, die aus Nomen als Basis und optional aus Adjektiven bestehen, erkannt. Nach der maschinellen Erkennung der Terme gibt es die zwei Bewertungsprozesse , nämlich "automatic validation" und "human validation" im 'TermoStat'. Zuerst wird "automatic validation" mit den schon gefunden Termen durchgeführt. Dann wird "human validation" mit den drei Spezialisten im Bereich Telekommunikation durchgeführt.
Zum Vergleich von technischem Korpus und nicht-technischem Korpus wurden die zwei Werte "test-value" und "probability-value" verwendet. Wenn die Worthäufigkeit des Wortes T im nicht-technischen Korpus gleich oder höher

als im technischen Korpus ist, wird sie als "probability-value" bewertet. Im umgekehrten Fall wird die Worthäufigkeit des Wortes T als "test-value" bewertet und als Termkandidat erkannt. Diese Werte sind keine booleschen[1] Werte (domainspezifisch oder "nicht domainspezifisch"). Die Werte für 'test-value' werden in die folgenden drei Gruppen eingeteilt:

- **significantly higher (SP+)**: gleich oder höher (+3.09)
- **lower (SP-)**: gleich oder niedriger als '-3.09'
- **theoretical frequency (SP0)**: zwischen '+3.09' und '-3.09'

Im Artikel wird geschrieben, dass "SP+ und SP-" für die Erkennung der domainspezifischen Terme sehr hilfreich sein kann.

Schlußfolgerung
Die unterschiedlichen Häufigkeiten der jeweiligen Wörter in zwei unterschiedlichen Korpora werden verwendet, um unwichtige Wörter aus den erkannten Termkandidaten zu entfernen.

3.3 Allgemeines Korpus als "Background Filter"

Die wesentlichen Aspekte des Artikels "Using Generic Corpora to Learn Domain-Specific Terminology (D. Vogel 2003)" [Vog03] werden hier vorgestellt. Es gibt nach Vogel beim Korpora-Vergleich keine Untersuchung für die Häufigkeiten der jeweiligen Wörter in unterschiedlichen Korpora. Ein allgemeines Korpus nur als "Background Filter" wird mit einem Target-Korpus verglichen, um unwichtige Wörter aus dem Target-Korpus zu beseitigen. Das bedeutet, dass die Überlappung zwischen einem allgemeinen Korpus und einem Target-Korpus als "nicht domain-spezifischer Teil" betrachtet werden. Dieser Teil wird aus dem Target-Korpus entfernt.
In dem Artikel wird die folgende Abbildung erklärt:

> Figure 2 represents terms in a target corpus as white oval and those in a background corpus as a gray oval. We are interested only

[1] engl. boolean

in the part of the white oval not overlapped by the gray. The gray oval filters out common terms.

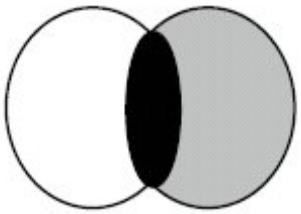


Abbildung 3.1: Figure 2. Background filters out common terms. (D. Vogel 2003)

Die drei Target-Korpora im Artikel werden zuerst wie folgt transformiert. Dann werden die zwei Background-Korpora, von denen eines 631443 Wörter und das andere 2796354 Wörter enthält, transformiert.
Die Interpunktionen im Text werden dabei ignoriert. Stoppwörter werden durch den Token '<X>' ersetzt und später für Bi- und Trigramme wieder verwendet.

Vor der Transformation:
General Motors Corp. unveiled a prototype electric car it says outpaces some gas-burning sports cars and runs twice as far between charges than previous electric models. The two-seater Impact, which tapers at the rear like a Citroen, can travel 120 miles at 55 mph before recharging and zooms from 0 to 60 mph in eight seconds, GM Chairman Roger Smith said at a news conference Wednesday.
Nach der Transformation:
generic motor corp unveiled <X> prototype electric car <X> <X> outpaces <X> gas burn sport car <X> run <X> <X> <X> <X> charge <X> <X> electric model <P> <X> two-seater impact <X> taper <X> <X> rear <X> <X> citroen <X> travel <X> mile <X> <X> mph <X> recharge <X> zoom <X> <X> <X> mph <X> <X> <X> <X> chairman roger smith <X> <X> <X> new conference <X>

Tabelle 3.2: Vor und Nach der Transformation (D. Vogel 2003)

Mit Bindestrich zusammengesetzte Wörter werden getrennt. Wörter werden durch "Porter Stemmer" lemmatisiert, der in Kapitel 2.4.1. ("Automatische Lemmatisierung") vorgestellt wurde. Die längste Wortform von demselben Stamm (z.B. (stem) announc: announcing, announcement, announced, announcements, announc) wird dabei als das Lemma (z.B. "announcement")

ausgewählt. Getrennte Wörter (z.B. auto maker) werden durch zusammengesetzte Wörter (z.B. automaker) ersetzt, wenn die Punktbewertung des Unigrammes höher als die des Bigramms ist.
Uni-, Bi-, und Trigramme werden im System erstellt. Durch das Assoziationsmaß "Log Likelihood Ration (LLR)" wird die "Term-Scoring Statistic" bewertet. Im Hintergrund werden die zwei genannten Background-Korpora als Filter benutzt, um nicht benötigte Wörter in einem Target-Korpus zu identifizieren und zu entfernen.

Schlußfolgerung
Im Artikel wurde geschrieben, dass die Nutzung eines allgemeinen Korpus als "Background Filter" für die Erkennung der wichtigen Terme in einem domainspezifischen Korpus eine sehr gute Idee ist.
Zusätzlich sollte man dabei bedenken, dass auch in der Überlappung zwischen einem allgemeinen Korpus und einem Target-Korpus domainspezifische Terme mit unterschiedlichen Worthäufigkeiten enthalten sein können. Deshalb sind die unterschiedlichen Worthäufigkeiten zwischen Korpora für die Erkennung der domainspezifischen Terme sehr nützlich.

3.4 Schlüsselwortextraktion zwischen Korpora

In "Comparing Corpora using Frequency Profiling" (P. Rayson und R. Garside) [RG00] wird gezeigt, dass Schlüsselwörter (key words) durch "Frequency Profiling" mit Hilfe der unterschiedlichen Worthäufigkeitslisten aus den gewünschten zwei Korpora erkannt werden können. Als Vergleichsbeispiele dienen hierbei die folgenden drei bekannten annotierten Korpora:

- **Brown corpus** (one million words of American English)
- **LOB corpus** (Hofland & Johansson 1982 - one million words of British English)
- **BNC (British National Corpus)**
  (Aston & Burnard 1998 - one hundred million words)

Jedes Wort in den zwei Frequenzlisten aus den gewünschten zwei Korpora wird durch das Assoziationsmaß "log-likelihood" berechnet. Dies wird durch die folgende Tabelle 3.3 ausgeführt.

	CORPUS ONE	CORPUS TWO	TOTAL
Freq of word	a	b	$a+b$
Freq of other words	$c-a$	$d-b$	$c+d-a-b$
TOTAL	c	d	$c+d$

Tabelle 3.3: Contingency table for word frequencies (P. Rayson und R. Garside)

'E' steht dabei für die Erwartungswerte. 'E1' für ein Korpus und 'E2' für ein anderes Korpus:

$$E1 = c * (a+b)/(c+d) \text{ und } E2 = d * (a+b)/(c+d)$$

Jedes Wort wird durch "log-likelihood (LL)" wie folgt berechnet:

$$LL = 2 * ((a * log(a/E1)) + (b * log(b/E2)))$$

Schlußfolgerung
Wenn es zwei (oder mehrere) Korpora in einem domainspezifischen Bereich gibt, können Schlüsselwörter durch die oben genannte Berechnung mit Hilfe der jeweiligen erstellten Worthäufigkeitslisten erkannt werden.

3.5 Ähnlichkeit zwischen Korpora

Im Artikel "Comparing Corpora (A. Kilgarriff 2001)" [Kil01] werden die verschiedenen Berechnungen (z.B. X^2-test, Mann-Whitney ranks test, Log-likelihood (G^2)) zur Ähnlichkeit zwischen Korpora (corpus similarity) vorgestellt. Zuerst werden Schlüsselwörter in einem Korpus (oder Text) durch die folgende Abbildung "Basic contingency table" erkannt:

> There are 'a' occurrences of 'w' in text X (which contains $a+c$ words) and b in Y (which has $b+d$ words).

Das beste Resultat für die Berechnung zur Ähnlichkeit zwischen Korpora (corpus similarity) lieferte der "Mann-Whitney ranks test" und das zweit-beste

	X	Y	
w	a	b	$a+b$
not w	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d=N$

Tabelle 3.4: Basic contingency table (A. Kilgarriff)

der "X^2-test".

Schlußfolgerung
Bei dem im Artikel "Comparing Corpora" vorgestellten Verfahren handelt es sich um Berechnungen zur Ähnlichkeit zwischen Korpora (corpus similarity). In dieser Arbeit geht es jedoch um die Extraktion und Erkennung von domainspezifischem Vokabular aus den jeweiligen Korpora.

3.6 Anwendungen der TE domainspezifischen Vokabulars

Ohne die zugehörige Terminologie kann man die jeweiligen Fachkenntnisse nicht verstehen und erweitern. Die wichtigsten Anwendungen der Terminologie-Extraktion (TE) domainspezifischen Vokabulars sind die folgenden:

- **Erstellung von Fachwörterbüchern**
- **Verbesserung von Suchmaschinen**: fokussiertes Web-Crawling (Focused Web Crawling), Indexierung, Klassifikation
- **Verbesserung der maschinellen Übersetzung**

Erstellung von Fachwörterbüchern
Jedes Fachwörterbuch besteht aus einem zugehörigen Wortschatz. Dieser Wortschatz muss ständig geändert und verbessert werden, um neue Fachbegriffe in dem jeweiligen Bereich zu erfassen. Dafür ist das Web als Korpus sehr hilfreich. Nach der schon genannten Annahme in Kapitel 3 wird ein Korpus als eine Quelle für das bereichsspezifische Vokabular in den jeweiligen Bereichen betrachtet. Ein Korpus kann aus dem Web leicht erstellt und aktualisiert werden, um öfter vorkommende wichtige und neue Fachbegriffe zu extrahieren.

Dies stellt eine Hauptmotivation dieser Arbeit dar.

Verbesserung von Suchmaschinen

Eine schnell wachsende Anzahl an Webseiten muss klassifiziert werden. Die Klassifikation der Webseiten ist die aktuell relevante Aufgabe von Suchmaschinen. Die manuelle Themenzuordnung für die ständig zunehmende Anzahl von Webseiten ist unmöglich. Deshalb ist die automatische Klassifikation unvermeidlich. Webseiten, die zu einem Thema gehören, können durch Ähnlichkeitsmaße automatisch erkannt und zugeordnet werden.
"Fokussiertes Web-Crawling" ist die Suche nach domainspezifischen Webseiten. Die Qualität von "fokussiertem Web-Crawling" sollte einen großen Einfluß auf die Klassifikation haben. Die automatische TE domainspezifischen Vokabulars ist die Basis für fokussiertes Web-Crawling, Indexierung und Klassifikation.

Verbesserung von maschineller Übersetzung (Machine Translation)

'SYSTRAN' ist der weltweit führende Hersteller von Übersetzungssoftware. Die folgenden acht Wörter für einen Test mit "SYSTRAN" wurden vom Englischen ins Deutsche übersetzt:

Englisch	**Deutsch**
black	Schwarzes
box	Kasten
black box	Flugschreiber
domain	Gebiet
specific	spezifisch
domain specific	Gebietsbesondere
terminology	Terminologie
domain specific terminology	Gebietsbesondereterminologie
"domain specific terminology"	"spezifische Terminologie des Gebietes"

Tabelle 3.5: Test mit 'SYSTRAN' (Stand: 03.08.2007 / www.systran.de)

Es gibt verschiedene deutsche Übersetzungen von "black box" z.B. "Black Box", Fahrdatenschreiber, Flugschreiber, Unfallschreiber. Die Demo-Version von 'SYSTRAN' im Internet liefert nur eine Übersetzung.
Die Übersetzung von "black box" ist hierbei keine Bedeutungszusammenset-

zung von 'black' und 'box'. Das bedeutet, dass das System den Term "black box" als Mehrwortlexem bzw. Mehrwortterm mit Hilfe von verwendeten Lexika erkennt. Die deutschen Übersetzungen von "domain specific" und "domain specific terminology" sind aber eine Bedeutungszusammensetzung wie 'Gebietsbesondere' und 'Gebietsbesondereterminologie' und falsch übersetzt. Die Suchmaschine Google liefert keinen Treffer für "Gebietsbesondere" und "Gebietsbesondereterminologie". Das bedeutet, dass der Term "domain specific terminology" als Mehrwortlexem mit Hilfe von verwendeten Lexika nicht erkannt werden kann.
Die alternative deutsche Übersetzung mit Anführungszeichen ist "spezifische Terminologie des Gebietes". Die Übersetzung in "SYSTRAN" muss eine richtige Entwicklung sein, obwohl die deutsche Übersetzung nicht zufriedenstellend ist.
Als deutsche Übersetzung von "domain specific" im Bereich der Computerlinguistik sollte "domänenspezifisch" oder "domain-spezifisch" (bzw. bereichsspezifisch, branchenspezifisch) erwartet werden. In diesem Fall sollte der Ausdruck "domain specific" als eine Einheit übersetzt werden.

Ein zukünftiges System für maschinelle Übersetzungen muß in der Lage sein, solche Mehrwortterme als eine Einheit bzw. Mehrwortlexeme zu erkennen und z.B. durch Maximum-Matching mit Hilfe von allgemeinen Lexika und domainspezifischen Lexika richtig zu übersetzen.
Für vorhandene und ständig neu entstehende Fachtermini müssen Übersetzungssysteme ständig aktualisiert und verbessert werden. Bei den maschinenlesbaren Lexika für die maschinelle Übersetzung gibt es jedoch einen Mangel an domainspezifischen Fachbegriffen, die als eine Einheit bzw. Mehrwortlexeme erkannt und übersetzt werden. Die automatische TE domainspezifischen Vokabulars wird dafür sehr nützlich sein.

3.7 References

- **Asmussen, Jørg**. Automatic detection of new domain-specific words using document classification and frequency profiling. In: Proceedings of the Corpus Linguistics 2005 Conference, Vol. I, Birmingham 2005. [Asm05]

- **Patrik Drouin**. Detection of Domain Specic Terminology Using Corpora Comparision. In: Proceedings of the fourth International Conference on Language Resources and Evaluation, Lissabon, 2004. [Dro04]

- **Patrik Drouin**. Term extraction using non-technical corpora as a point of leverage. In Terminology, 9(1), pages 99–115, 2003. [Dro03]

- **M. Hong, S. Fissaha, J. Haller**. Hybrid Filtering for Extraction of Term Candidates from German Technical Texts. Conference TIA-2002, Nancy, 3 et 4, mai 2003. [HFH01]

- **Adam Kilgarriff**. Comparing Corpora. International Journal of Corpus Linguistics 6:1, pages 1–37, October 2001. [Kil01]

- **S. O'Shaughnessy**. Dynamische Erkennung domänenspezifischen Vokabulars. Magisterarbeit im Studiengang Computerlinguistik, Oettingenstr.67, 80538 München, 2006. [O'S06]

- **Rayson, P. and R. Garside**. Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL 2000), pages 1–6, 2000. [RG00]

- **Riloff, Ellen and Shepherd, Jessica**. A Corpus-Based Apporach for Buildung Semantic Lexicons. In: Proceedings of the second Conference on Empirical Methods in Natural Language Processing (EMNLP-2), 1997. [RS97]

- **Riloff, Ellen and Shepherd, Jessica**. A Corpus-Based Bootstrapping Algorithm for Semi-Automated Semantic Lexicon Construction. In: Journal of Natural Language Engineering, Bd. 5, Nr. 2, S. 147–156, 1999. [RS99]

- **David Vogel**. Using Generic Corpora to Learn Domain-Specific Terminology. Workshop on Link Analysis for Detecting Complex Behavior, Washington, DC, USA, 2003. [Vog03]

- **Wu, Yi-fang B. and Bot, Razvan S. and Chen, Xin**. Domain-specific Keyphrase Extraction. In: Proceedings of the 16 International Joint Conference of Artificial Intelligence, S. 668–679, 1999.
- **Feiyu Xu and et al.**. An Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms und their Relations with Bootstrapping. Proc. Of the 3rd International Conference on Language Resources and Evaluation, 2002. [Xea02]
- **Feiyu Xu and Daniela Kurz**. Text Mining for the Extraction of Domain Relevant Terms and Term collocations. [Stand: 04.08.2007, www.coli.uni-saarland.de/publikationen/softcopies/Kurz:2002:TME.pdf] [XK]

Kapitel 4

Domainspezifische Terme (DST) und ihre Relationen

Es gibt uneinheitliche Definitionen von 'Wort' [Bußß90, S. 849]. Besonders die Zählung der Wörter ist nicht eindeutig. Nach H. Bergenholtz/J. Mugdan (2000) [Mug00] herrscht in folgenden Fällen Uneinigkeit:

- bei Komposita (Samstagnachmittag, Tränengasgranaten),
- bei Schreibungen mit Bindestrich oder Gedankenstrich (Rugby-Nationalmannschaft, Rugby-Fans, Nizza-Paris),
- bei Abkürzungen (dpa, CRS),
- bei Zahlen (18.),
- bei Präpositionen mit enklitischem Artikel (ins, zum, im),
- bei zusammengesetzten Verbformen ("hat ... gezogen"),
- bei Verben mit abgetrenntem Präfix ("stiegen aus").

Schließlich ist ungeklärt, ob Interpunktionszeichen als Wörter oder Teil von Wörtern gelten sollen. Beispielsweise ist die Tokenisierung von Abkürzungen (z.B. Prof., Dr., bzw., W. Bergenholtz) sehr problematisch.

Die Komposita sind eine wichtige Quelle für domainspezifische Terme im Deutschen. Die Tendenz bei deutschen Fachausdrücken geht dahin, zusammengehörende Wörter als Einwort zu schreiben. Mit Bindestrich zusammengesetzte Wörter (z.B. Hals-Nasen-Ohren-Klinik) kann man als Einwort oder

Mehrwort zählen. Aber das ist ein Konventionsproblem. Sie werden als Einwort bzw. Einwortterm betrachtet, weil sie kein Leerzeichen beinhalten. In dieser Arbeit ist ein Wort eine Folge von Zeichen, die nicht durch Leerzeichen getrennt sind. Durch Leerzeichen werden Einwortterme und Mehrwortterme unterschieden:

- **Einwortterm**: Hals-Nasen-Ohren-Klinik, Information-Retrieval-System, Informationsgewinnung, Informationsrückgewinnung
- **Mehrwortterm**: Hals-Nasen-Ohren Klinik, Information Retrieval System, Information Retrieval

4.1 Einwortterme

Die Hauptzielsetzung dieser Arbeit ist die Erkennung der domainspezifischen Terme bzw. Einwort- und Mehrwortterme in den jeweiligen Bereichen. Die Erkennung der Einwortterme wird zuerst durchgeführt. Dann können Mehrwortterme daraus erkannt werden.
Laut Hoffmann [Hof88] werden sie auf der linguistischen Ebene in Simplizia, Derivationen, Komposita, Abkürzungen und Kurzwörter unterschieden:

- **Simplizia** sind in der Wortbildung nicht zusammengesetzte oder abgeleitete Wörter, die als Ausgangsbasis für Neubildungen verwendet werden können, z.B. Hund zu Hundesteuer, mies zu Miesling. [Bußß90, S. 686]. Sie sind ein sehr großer Teil der Fachtermini, meist Nomina und Adjektive. "Elementare Generische Terme (EGT)" in dieser Arbeit haben fast gleiche Eigenschaften wie Simplizia.
- **Derivationen** bezeichnen sogenannte Präfix- und Suffixbildungen bzw. Affixbildungen. Z.B. Verkauf, Versicherung, Schönheit. Wörter wie z.B. Hepatitis, die auf das Suffix "-itis" enden, sind meist medizinische Fachtermini.
- **Komposita** sind zusammengesetzte Wörter. Unter Komposita verstehen wir Wörter, die ohne Ableitungsmittel aus zwei oder mehreren selbstän-

dig vorkommenden Wörtern gebildet sind [Dro84, S. 401]. Die Bestimmungswörter stehen links, sie erklären das Grundwort näher. Das letzte Wort ist das Grundwort, das die Wortart der ganzen Zusammensetzung festlegt. Die Beziehung zwischen Bestimmungswort und Grundwort hängt von der Bestimmung des Grundwortes ab, das selbst Kompositum sein kann. Z.B. (Informationsgewinnung, Informationsrückgewinnung / information retrieval), (Betriebssystem / operating system, system software), (Recyclingpapier / recycling paper)

- **Abkürzungen** (abbreviation, acronym) sind Initialwörter. Sie kommen in Fachtexten sehr oft vor und sind sehr bereichsspezifisch. Sie entstanden aus einem Wort, z.B. PS (Pferdestärke), PKW (Personenkraftwagen), LKW (Lastkraftwagen), Kfz (Kraftfahrzeug) oder einer Wortgruppe z.B. ADAC (Allgemeiner Deutscher Automobil-Club). Sie sind fast immer mehrdeutig.

- **Kurzwörter** können aus den drei folgenden Arten bestehen [Fle92, S. 218-221]:
  (a) aus Anfangssegmenten: Akku (Akkumulator), Auto (Automobil), Foto (Fotografie), Lok (Lokomotive), Taxi (Taxameter)
  (b) aus Endsegmenten: Bus (Autobus / Omnibus), Rad (Fahrrad)
  (c) aus mittleren Segmenten: Basti (Sebastian), Lisa (Elisabeth)
  Sie werden im CISLEX-Wörterbuchsystem als einfache Formen bzw. Nomen wie Simplizia behandelt.

4.2 Mehrwortterme

In dieser Arbeit wird bei **Mehrworttermen** (multi-word terms) angenommen, dass sie mindestens ein Leerzeichen zwischen den Teiltermen enthalten. Sie bestehen terminologisch aus mehreren getrennten Worten, bzw. zusammengesetzten Wortgruppen (word groups), die interne syntaktische Strukturen bzw. Phrasen haben, und als Einheit betracht werden sollten (z.B. ALFA ROMEO, MERCEDES-BENZ Sprinter 31).
Die meisten terminologischen Mehrwortterme sind Nominalphrasen. Bei der Nominalphrasenerkennung werden im allgemeinen 'Part-of-Speech-Muster' (POS-

Muster) verwendet. Die wichtigsten POS-Muster für Mehrwortterme sind Adjektiv-Nomen (A+N), Nomen-Nomen (N+N) und Nomen-Präposition-Nomen (N+P+N). In "NPtool" von Arppe [Arp95, S. 5] werden syntaktische Muster erfolgreich für die Erkennung der englischen Nominalphrasen eingesetzt, was zu der Erkenntnis führt, dass ca. 80% aller 3.137 untersuchten Terme einem der Muster "Nomen (N)" (45%), "Nomen-Nomen (N+N)" (19%) und "Adjektiv Nomen (A+N)" (17%) zuzuordnen sind. Für die POS-Muster wird ein POS-Tagger verwendet.

4.3 Grundannahme für domainspezifische Terme

Die folgenden Beispiel-Terme bzw. Stichwörter können durch die Terminologie-Extraktion automatisch erkannt werden:

Neuheiten, BMW, Vergleichstest, Audi, Sportwagen

Die domainspezifischen Terme hierbei sind BMW, Audi und Sportwagen. BMW und Audi sind international bekannte Automarken. Terme wie z.B. Sportwagen werden als "Komplexe Generische Terme (KGT)" in dieser Arbeit betrachtet. Das Kompositum 'Sportwagen' besteht aus zwei Wörtern, nämlich Sport und Wagen. Das Wort 'Sport' kann domainspezifisch im Bereich von Sport sein. Aber das Wort 'Sport' ist nicht domainspezifisch im Automobilbereich. Das Wort 'Wagen' ist domainspezifisch im Automobilbereich. Solche Terme (z.B. Wagen, Auto) sind "Elementare Generische Terme (EGT)". Die EGT sind eine Basis für die Kompositabildung. Sie haben prinzipiell die gleichen Eigenschaften wie Simplizia. Sie sind Kernwörter (bzw. Grundwörter) für "Komplexe Generische Terme (KGT)".
Die Hauptzielsetzung dieser Arbeit ist, domainspezifische Terme im E-Commerce-Bereich für die deutsche Sprache zu erkennen.
Die folgende Grundannahme für domainspezifische Terme im E-Commerce-Bereich gilt in dieser Arbeit:

> Ein Term wird als domainspezifisch betrachtet, wenn er in einem Bereich öfter als andere Terme vorkommt und seltener in anderen Bereichen. Ein domainspezifischer Term beinhaltet mindestens

einen domainspezifischen Teil als "Elementaren Generischen Term (EGT)".

4.4 Elementare Generische Terme (EGT)

"Elementare Generische Terme" können nicht sinnvoll zerlegt werden (z.B. Wein, Auto, Schmuck, Möbel). Im CISLEX werden zwei Definitionen für einfache und komplexe Formen (bzw. Nomen) verwendet [MM05, S. 31-32]. Diese Definitionen für "Elementare Generische Terme" (EGT) und "komplexe Generische Terme" (KGT) übernehme ich in dieser Arbeit. 'EGT' werden wie folgt definiert:

> Ein Wort W ist ein EGT genau dann, wenn es keine sinnvolle Zerlegung $W = W_1W_2$ gibt, so daß W_1 eine Folge von Morphemen ist und W_2 ein EGT.

'EGT' sind keine Abstrakta (z.B. Service, Tuning, Informationen) sondern Konkreta bzw. Basis-Wörter zur Erkennung produktbezogener Terme im E-Commerce (z.B.Auto, Wein, Handschuhe, Möbel, Jeans). Produktbezogene Terme (z.B. Wein, Weinöffner) können durch die EGT mit Hilfe von Affixanwendung automatisch erkannt werden. Dabei ist die Qualität der EGT sehr wichtig für die Erkennung der domainspezifischen Terme in den jeweiligen Bereichen.
Zum Beispiel können solche domainspezifischen Terme (Auto, Wagen, usw.) als EGT und semantische Klasse von 'Auto' betrachtet und kodiert werden. Bekannte domainspezifische Abkürzungen (z.B. PKW, LKW, KFZ) gehören zu EGT. Marken bzw. Firmennamen (z.B. BMW, VW, Audi) sind abstrakt wie Organisationsnamen. Sie sollten ähnlich wie EGT behandelt werden.

4.4.1 Eigenschaften der EGT

Elementare generische Terme sind meist Einwortterme. Es gibt verschiedene Übersetzungsvarianten für ein Wort (z.B. Flugschreiber, Fahrdatenschreiber, Black Box, Unfallschreiber für "Black Box"). International anerkannte Ausdrücke können in den deutschen oder internationalen Webseiten häufig vorkommen. Die Übersetzungsvarianten und jeweiligen international anerkannten

Ausdrücke (z.B. Black Box) können als Synonym behandelt werden, wenn sie auf dasselbe Lemma bei der semantischen Kodierung wie folgt referiert werden können:

Semantische Kodierung
Flugschreiber;EGT=Branche.**Black Box**;
Fahrdatenschreiber;EGT=Branche.**Black Box**;
Black Box;EGT=Branche.**Black Box**;
Unfallschreiber;EGT=Branche.**Black Box**;

Tabelle 4.1: Semantische Kodierung

Mehrwortterme (z.B. Black Box) müssen als EGT betrachtet werden, weil sie nach der schon genannten Grundannahme (in Kapitel 4.3.) mindestens einen domainspezifischen Teil beinhalten. International verbreitete Englische Terme (z.B. automobile, car, motorcar, USB, PC) gehören deshalb zu EGT für die deutsche Sprache. Sie und jeweilige Übersetzungsvarianten können semantisch verschieden annotiert und auf dasselbe Lemma referiert werden, um die automatische Terminologie-Extraktion (TE) domain-spezifischen Vokabulars in den jeweiligen Bereichen signifikant zu verbessern.
Die typischen weiteren Eigenschaften der EGT sind die folgenden:

- **elementar, generisch**: z.B. Öl, Schuh, Wein, Auto, Platte, Apfel
- **Material**: z.B. Öl, Wasser, Air, Gold, Eisen
- **käuflich**: z.B. Auto, Kleid, Tasche

Die EGT können für einen Bereich, z.B. Automobil, Wagen, Fahrzeug (Automobilbereich), oder für mehrere Bereiche, z.B. Sitz (Automobilbereich, Möbel), Reifen (Automobilbereich, Schmuck), erfasst werden. Bei der semantischen Annotation können EGT unterschiedlich gekennzeichnet werden.

4.5 Komplexe Generische Terme (KGT)

Ein komplexer generischer Term (KGT) beinhaltet mindestens einen EGT.
Es gibt ca. 70.000 - 100.000 EGT und ca. 12.000.000 KGT für die deutsche Sprache. Im Allgemeinen ist der am Ende stehende 'EGT' der Kopf der 'KGT'. Er kann auch das Grundwort der Komposita benennen, das die Wortart der ganzen Zusammensetzung festlegt. KGT werden wie folgt definiert:

> Ein morphologisch-komplexes Wort $W = W_1...W_n$ ist ein KGT genau dann, wenn $W_1, ..., W_{n-1}$ Morpheme sind und W_n ein Wort mit denselben morphologischen Eigenschaften wie W ist. W beinhaltet mindestens einen EGT.

Beispiele für KGT:

Autofahrerbrille, Autofahreratlas, Funkautos, Ein-Liter-Auto, ADAC-Autositze, Autohändler, Autovermietung, Einsteigerkletterschuh, Sportkletterschuh, Fußballschuhe

Schlußfolgerung: KGT müssen mindestens einen EGT beinhalten. Die EGT sind eine Teilmenge von KGT. Ein EGT kann einem oder mehreren Bereichen zugeordnet werden. In den jeweiligen Bereichen gibt es eine endliche Menge (z.B. 10000) von EGT, um domainspezifische Terme bzw. KGT für E-Commerce-Branchen zu erkennen und zu klassifizieren.

4.6 Wortformen im CISLEX

Das DELA-System (**D**ictionnaire **é**lectronique du **LA**DL) ist das elektronische Wörterbuch des Französischen, das nach den Kriterien der Vollständigkeit, Korrektheit und Abgeschlossenheit entwickelt wurde [GM94, S. 3].
Das System wurde am LADL (Laboratoire d'Automatique Documentaire et Linguistique, Universität Paris 7) unter Leitung von Prof. Maurice Gross entwickelt. Man nennt das System auch das LADL-Format. Zur Zeit wird das LADL-Format für fünf europäische Sprachen verwendet.

Das deutsche CISLEX-Wörterbuchsystem
Das CISLEX ist ein elektronisches Wörterbuch des Deutschen, ähnlich wie das DELA-System für Französisch. Das am CIS entwickelte elektronische CISLEX-System wurde gemäß dem LADL-Format entworfen. Im CISLEX wurden die folgenden fünf Wortformen verwendet:

- **einfache Formen (EF)**
- **komplexe Formen (KF)**
- **Eigennamen oder EN-Formen**

- **Fremd- und Fachwörter oder F-Formen**
- **Abkürzungen und Morpheme oder S-Formen**

Lexeme sind Lexikoneinträge. Einwortlexeme spielen eine wichtige Rolle in der deutschen Sprache. Im CISLEX werden die folgenden Wortarten unterschieden:

Nomen (N)	Pronomen (PRON)	Präposition (PRAEP)
Adjektiv (A)	Adverb (ADV)	Konjunktion (KONJ)
Verb (V)	Partikel (PART)	Interjektion (INTJ)
Determinatoren (DET)	Verbpartikel (VPART)	

Das deutsche Kernlexikon (CISLEX-DKL) beinhaltet einfache Formen und komplexe Formen, so wie das DELA-System. Diese Definitionen von EF und KF entsprechen den Definitionen für "Elementare Generische Terme (EGT)" und "Komplexe Generische Terme (KGT)" in dieser Arbeit.

4.7 KFIDF: TFIDF-based single-word term classifier

In der Trendanalyse werden domain-spezifische Terme und ihre semantischen Relationen aus Webseiten und Fachtexten der jeweiligen Domäne extrahiert. 'KFIDF' ist eine Modifikation von "TFIDF (term frequency inverted document frequency)". 'KFIDF' ist für schon kategorisierte Dokumente besser geeignet als das TFIDF-Maß, um domain-relevante Einwortterme automatisch aufzufinden. Das KFIDF-Maß ist definiert wie folgt [Xea02]:

KFIDF (w, cat) = docs (w, cat) * LOG($\frac{n*|cats|}{cats(word)} + 1$)
docs (w, cat) = number of documents in the category cat containing the word w, n = smoothing factor
cats (word) = the number of categories in which the word occurs

Die Grundannahme von 'KFIDF' ist wie folgt:

> Ein Term wird als relevant betrachtet, wenn er in einer bestimmten Kategorie öfter als andere Terme vorkommt und in allen anderen Domänen seltener.

Domain-relevante Einwortterme (domain relevant single-word terms) aus kategorisierten Dokumenten können durch das oben erwähnte KFIDF-Maß besser extrahiert werden. In dieser Methode werden verschiedene assoziative Maße, nämlich "Mutual Information", 'Log-Likelihood' und 'T-Test' (bzw. T-Score) für Term-Kollokationen (term collocations), z.B. "zur Verfügung stehen", angewendet, um Mehrwortterme aus extrahierten Einworttermen zu erkennen. Das Log-Likelihood-Maß lieferte das beste Resultat für "low-frequency data" und Adjektiv-Nomen-Kombinationen.
In dieser Methode wird das sogenannte Ontologie-Netz "GermaNet" für lexikal-semantische Informationen angewendet, um semantische Relationen zwischen extrahierten Termen zu erkennen.

4.8 GermaNet - Semantisches Wortnetz

Die bekanntesten Wortnetze sind WordNetz (1985), GermaNet (1996) und EuroWordNet (1996). Wortnetze bzw. Thesauri sind lexikalische Datenbanken im Web, in der die semantischen und lexikalischen Beziehungen zwischen den Wörtern erstellt werden.

> GermaNet is a lexical-semantic net that has been developed within the LSD Project at the Division of Computational Linguistics of the Linguistics Department at the University of Tübingen. It has been integrated into the EuroWordNet (EWN), a multilingual lexical-semantic database. (von der GermaNet-Homepage[1])

"GermaNet" (Hamp und Feldweg (1997)) ist ein lexikal-semantisches Wortnetz für die deutsche Sprache wie WordNet für Englisch (Miller et al.(1993)) und EuroWordNets für 8 europäische Sprachen (Vossen 1998). Es ist eine lexikalische Taxonomie für den deutschen Grundwortschatz und ein von Hand

[1] GermaNet: www.sfs.uni-tuebingen.de/lsd [13.12.2006]

erstellter Thesaurus, in dem Wörter und ihre Relationen strukturiert werden. Lexeme[2] mit gleicher Bedeutung bzw. Synonyme sind zu semantischen Konzepten (Gruppen) zusammengefasst und werden als Synsets (set of synonyms) bezeichnet. Sie sind das zentrale Konzept der lexikalischen Kodierung. Zur Zeit gibt es 53.312 Synsets für die semantische Klassifikation der Wörter in GermaNet (Stand 13.12.2006, GermaNet-Homepage). Diese werden wie bei 'WordNet' sehr fein verarbeitet. Die Konzepte der Synsets werden für die Klassifikation verwendet.
Zwischen Synsets gibt es semantische Relationen, innerhalb derer zwei verschiedene Relationstypen unterschieden werden:

- **Lexikalische Relationen (Synonymie, Antonymie (Ist-Gegenteil-Von)[3])** sind bidirektionale Beziehungen. (z.B. Synonymie (öffnen -><- aufmachen, Auto -><- Wagen), Antonymie (kalt <-> warm))

- **Konzeptuelle Relationen** sind Hyponymie ('is-a'), Hyperonymie, Meronymie (Teil-Ganzes-Relation / part-whole relation), Holonymie (Umkehrung der Meronymie), Implikation (Verbkonzepte in einem logischen Zusammenhang) und Kausation.
  Z.B. ist Frucht ein Hyperonym (Oberbegriff) zu Apfel. Apfel ist ein Hyponym (Unterbegriff) zu Frucht.
  Arm ist ein Meronym zu Körper. Körper ist ein Holonym zu Arm.
  "Schnarchen" impliziert "schlafen" (Implikation), "öffnen" verursacht "offen" (Kausation).

Die Hyponymie-Beziehung ist die zentrale Relation für die hierarchische Struktur des Wortschatzes.
Eigennamen und Abkürzungen sind in GermaNet nicht integriert. Bei GermaNet-Anfragen kann man für ein Wort alle Bedeutungen und für zwei Wörter ihre Relationen finden:
Z.B. "Bank" - Sitzmöbel, Geldinstitut und wirtschaftliche Institution

[2]Lexeme haben untrennbar miteinander verbundene Funktionen [Sch92, S. 2]. Wort als lexikalische Einheit bzw. Element des Wortschatzes [Bußß90].

[3]Synonymie (Bedeutungsgleichheit), Antonymie und Opposition (Bedeutungsgegensatz), Homonymie (Ein Homonym ist ein Lexem, das für unterschiedliche Bedeutungen haben. - gleiches Wort, aber verschiedene Bedeutungen wie Bank (Sitzmöbel) und Bank (Geldinstitut) -

“Internet-Service-Provider Firma” - “Internet-Service-Provider” ist ein Hyponym zu “Firma” [Xea02].

Kapitel 5

TE domainspezifischen Vokabulars aus einer Webseite

5.1 WWW und HTML

Der Begriff "World Wide Web" wird häufig als Web, WWW oder W3 abgekürzt. Das WWW besteht aus Webseiten (Dokumenten), die in Markup-Sprachen (z.B. SGML, HTML und XML) geschrieben werden. HTML (HyperText Marktup Language) ist eine Teilmenge von SGML (Standard Generalized Markup Language). HTML ist einfacher und deutlicher als SGML. Es ist die international anerkannte Textbeschreibungssprache, mit der Texte formatiert werden sollen. XML (Extensible Markup Language) ist auch eine Teilmenge von 'SGML' und eine erweiterbare Auszeichnungssprache, in der Tags und Dokumententypen selbst definiert werden können. Diese Arbeit behandelt hauptsächlich Webseiten, die in HTML geschrieben sind.
Eine Webseite beinhaltet einen Inhalt (bzw. Text), HTML-Tags und Entities[1]. In HTML-Tags gibt es Anfangstags (z.B. <p>, ,) und Endtags (z.B. , , </table>). Endtags sind nicht immer obligatorisch sondern manchmal optional (z.B. <p>, <td>,
, <img>). Zwischen Groß- und Kleinschreibung wird nicht unterschieden.

[1]Das sind Steuersequenzen für Sonderzeichen z.B. ö (ö), ß (ß), < (<), > (>).

5.2 Meta-Keywords und Titelangaben

Ein HTML-Dokument wird zwischen "<html>...</html>" eingeschlossen und besteht aus zwei Teilen: Kopf (head) und Körper (body). Im Kopf werden Titel-Angaben und Meta-Angaben für Web-Server, Web-Browser und Suchmaschinen erstellt. Im Körper kann man den Inhalt des Dokuments mit Hilfe von HTML-Tags darstellen.
Im allgemeinen sind die Titel-Angaben die wichtigste Stelle für Suchbegriffe. Deshalb müssen domainspezifische Terme, die aus Titel-Angaben extrahiert wurden, zur Termgewichtung besonders behandelt werden.
Beim "Google Ranking-Algorithmus" wird das Title-Tag bedeutend gewichtet, weil relevante Terme direkt im 'Title' vorkommen sollten.
Im Kopf gibt es verschiedene Meta-Angaben. In dieser Arbeit werden nur zwei Meta-Tags, nämlich Meta-Keyword-Tag und Meta-Description-Tag behandelt, weil sie eine relevante Quelle für domain-spezifische Terme sein können.
Meta-Keywords sind Schlüsselbegriffe in Meta-Tags. Sie können für höhere Suchmaschinen-Platzierungen (bzw. Suchmaschinen-Optimierung[2]) effizient gebraucht werden und gleichzeitig auf eine andere Art mißbraucht werden. Durch den 'Meta-Keyword-Tag' kann man Stichwörter (Keywords) für eine Webseite erstellen. Durch den 'Meta-Description-Tag' kann man eine zusammengefasste Beschreibung für eine Webseite in 2-3 Sätzen erstellen. Ein Beispiel für den 'Head-Teil' aus einer E-Commerce-Seite (www.autoscout24.de [Stand: 27.11.2006]) aus dem Bereich "Autobranchen" ist hier zu sehen:

<head>
<title>AutoScout24 Europas Automarkt für Gebrauchtwagen und Neuwagen</title>
...
<meta name = "description" content = "AutoScout24 Ihr grosser Automarkt mit ueber 1,4 Millionen aktuellen Angeboten die groesste Auswahl in Europa">
<meta name = "KEYWORDS" content = "Automarkt, Autoboerse, Autohaendler, Gebrauchtwagen, Alfa Romeo, anbieten, Anhaenger, Ankauf, Audi, Austin, Auto, BMW, Cabrio, Caravan, Chevrolet, Chrysler, Citroen, Daimler, EU, EU-Neuwagen, Fahrzeug, Ferrari, Fiat, Finanzierung, Ford, Frankfurt, Gebrauchtwagen, Gelaendewagen, Golf, Gutachten, Haendler, Hamburg, Harley, Honda, Jahreswagen, Kfz, Kia,

[2]Als Suchmaschinenoptimierung bezeichnet man die Anpassung einer Webseite, um bessere Positionen in den Suchergebnissen einer Suchmaschine zu erreichen und somit mehr Besucher zu erhalten. (Quelle: www.informationsarchiv.net/lexikon/850.suchmaschinenoptimierung.html [27.11.2006])

Koeln, kostenlos, Lada, Leasing, Limousine, Lkw, Mazda, Mercedes, Mitsubishi, autoscout24, autoscout, autoscout24.de, Motorrad, Muenchen, Neuwagen, Neufahrzeug, Nissan, Oldtimer, Opel, Peugeot, Pickup, Polo, Porsche, Pkw, Renault, Rover, suchen, Verkauf, Volvo, VW, Wagen, Wohnwagen, Wohnmobil">
...
</head>

Gute Termkandidaten stammen aus Titel- und Meta-Angaben. Und sie müssen bei der Gewinnung der domain-spezifischen Termen besonders beachtet werden.

5.3 Terme aus sechs verschiedenen Quellen

In dieser Arbeit werden domainspezifische Terme aus den folgenden sechs verschiedenen Quellen innerhalb einer Webseite extrahiert:

- Terme aus der URL-Adresse
- Terme aus der Titelangabe
- Terme aus den Meta-Keywords
- Terme aus der Meta-Description
- Terme aus dem Ankertext
- Terme aus dem Body-Teil (bzw. Inhalt einer Webseite)

Terme aus dem Body-Teil
Der Inhalt eines Dokuments wird im Body-Teil in HTML dargestellt. Meta- und Titel-Angaben sollten auch im Body-Teil vorkommen. Dadurch kann man domainspezifische Terme erkennen und erwerben.

Terme aus dem Ankertext
Der Begriff "Ankertext (Link-Text, anchor text)" bezeichnet den anklickbaren Text mit einem Verweis (Hyperlinks) auf eine andere Webseite. Zwischen den HTML-Tags <a> und </A> können Ankertexte geschrieben werden.

Händlerbereich[3]
<img ...>Europas Automarkt für Gebrauchtwagen

Ein Ankertext sollte meist phrasenweise wie oben für eine kurze Beschreibung einer Webseite erstellt werden. Daraus können gute Terme erworben werden. Es gibt typische unwichtige Ankertexte (z.B. home, next, zurück), die beim Termerwerb (term acquisition) genau wie Stoppwörter eliminiert werden müssen.

Terme aus der URL-Adresse

Eine URL-Adresse[4] verweist auf eine Webseite. Die meisten bekannten URL-Adressen für internationale Firmen können Firmennamen und Haupttätigkeiten für die kommerzielle Bekanntmachung beinhalten. Daraus kann man Firmennamen in Kombination mit den Titelangaben und gute Terme identifizieren und erwerben.
Eine URL wird von hinten nach vorne gelesen. Sie setzt sich zusammen aus TLD (Top-Level-Domain), SLD (Second-Level-Domain), Subdomain und dem Protokolltyp:

http://	www.	bmw.	de
Protokoll	Subdomain	SLD	TLD

Die Subdomain ist nicht immer obligatorisch (z.B. http://antennewest.de). Die SLD kann gute Kandidaten für Firmennamen oder domainspezifische Terme beinhalten.
Vorstellbare Firmennamen innerhalb einer URL-Adresse können in drei Gruppen eingeteilt werden:

- **Vollständige Firmennamen**: Bei der Abstimmung mit Titel-Angaben muß er caseinsensitiv behandelt werden.
- **Variationen der Firmennamen**: z.B. wegen Umlauten (ä, ö, ü), Bindestrich (-) und Sonderzeichen (&)
- **Abkürzungen der Firmennamen**: z.B. GM, IBM

[3] Händlerbereich (Entity: ä → ä)
[4] URL (Uniform Resource Locator)

Die Tabelle 5.1 zeigt signifikante Terme aus URL- und Titel-Angaben. Mit verschiedenen Methoden können Firmennamen ebenfalls mit Hilfe von Firmennamenslisten erkannt werden, die Firmennamen, Abkürzungen und Variationen beinhalten sollten.

URL	**Terme aus URL**	**Original-Titel-Angaben**
http://www.bmw.de	bmw	BMW
http://www.autoscout24.de	autoscout24	AutoScout24 Europas Automarkt für[5] Gebrauchtwagen und Neuwagen
http://www.citroen.de	citroen	CITROËN AG
http://www.gm.com	gm	GM Cars - General Motors Corporate Website - GM Customer Service
http://www.ibm.com/de/	ibm	IBM Deutschland

Tabelle 5.1: Beispiele für signifikante Terme aus URL und Title-Angaben

5.4 Affixanwendung zur Erkennung der KGT

Zur Erkennung der KGT (Komplexe Generische Terme) wird in dieser Arbeit die Affix-Anwendung von EGT (Elementare Generische Terme) genannt und gebraucht. Bei der Affixanwendung gibt es Präfix-, Infix- und Suffixanwendung, ähnlich wie beim Derivationsprozess. Durch die Affixanwendung mit EGT können domainspezifische Terme identifiziert werden.
EGT müssen konkret (z.B. Auto, Wein, Banane) sein. Die Affixanwendung definiert mit einem Beispiel-Term 'Auto' wie folgt:

Suffix-	$W = W_1...W_n \quad W_n$ ist ein EGT (z.B. Auto).	Renault-Autos
Präfix-	$W = W_1...W_n \quad W_1$ ist ein EGT.	Autositze
Infix-	$W = W_1...W_n \quad W_{1+1}...W_{n-1}$ beinhaltet ein EGT.	Gebrauchtautomarkt

Tabelle 5.2: Definitionen von Affixanwendung

Zur Erkennung der domainspezifischen Terme (DST) können diese in den folgenden Punkten betrachtet und erkannt werden:

- selber EGT (z.B. Wagen, Auto)

- DST durch Suffixanwendung (z.B. Luxuswagen, Polizeiauto)
- DST durch Präfixanwendung (z.B. Wagenmiete, Autokonzern)
- DST durch Infixanwendung (z.B. Altautoentsorgung, Altautorecycling)
- DST durch die Erkennung mit Hilfe domainspezifischer Listen (z.B. Firmennamen: VW, Modellnamen: Golf)

Die Infixanwendung erzeugt mehr Fehltreffer als die Suffix- und Präfixanwendung. Trotzdem kann sie eine nützliche Methode zur Erkennung von DST sein.
Zusammengesetzte Einwortterme für die deutsche Sprache sind ein signifikanter Teil der DST, z.B. Mietwagen, Rent-A-Car. Solche domainspezifischen Einwortterme, z.B. 'Rent-A-Car', 'Car-Hifi' aus den englischen Mehrwortterm en "rent a car" und "car hifi" werden im Deutschen meist mit Bindestrich zusammengesetzt. Nach diesen Betrachtungen wird die sogenannte Affixanwendung entworfen und zur Erkennung der DST eingesetzt. Für die englische Sprache kann sie auch nützlich sein. Durch die erwähnte Affixanwendung kann 'car' im englischen Mehrwortterm "rent a car" als ein domain-spezifischer Teil erkannt werden.
Bei der Affixanwendung werden zuerst die Suffixanwendung, zweitens die Präfixanwendung und drittens die Infixanwendung überprüft - nicht rekursiv innerhalb eines Einworttermes. Die folgende Tabelle zeigt das Ergebnis ("www.auto.de" Stand: 22.08.2007) der Terminologie-Extraktion "CGI-Programm 1 mit sechs verschiedenen Quellen", die im Kapitel 5.6.2. vorgestellt werden:

S: Suffix-Anwendung, P: Präfix-, I: Infix-, Marke: Automarke <Title, Metakeyword, Metadescription, Ankertext, Body, URL> aus sechs verschiedenen Quellen in einer Webseite (z.B. <1,1,1,0,0,0>)
[3] Gebrauchtwagen <1,1,1,0,0,0> [$wagen; S$]
[2] Pkw/ PKW <0,1,0,1,0,0> [$Pkw; Acronym/Persoanlkraftwagen$]
[2] Deutschlands grosses Autoportal <1,0,1,0,0,0> [$Auto; P$]
[2] Suzuki Splash <0,0,0,0,2,0> [$Suzuki; Marke$]
[2] Autoverkauf <0,2,0,0,0,0> [$Auto; P$]
[2] Neuwagen <1,1,0,0,0,0> [$wagen; S$]
[1] kompakte Schrägheckli mousine <0,0,0,0,1,0> [$limousine; S$]
[1] Gebrauchtwagenmarkt <0,0,1,0,0,0> [$wagen; I$]

5.4.1 Suffix-, Präfix- und Infixanwendung

Die **Suffixanwendung** ist der Kern der Affixanwendung.
Durch die Suffixanwendung können die auf Produkte bezogenen Terme (z.B. Renault-Autos, Ein-Liter-Auto) effizient erkannt werden, weil EGT wie ein Grundwort bei den Komposita behandelt werden können.
Damit können hierarchische Strukturen zwischen Termen - Oberbegriffe und Unterbegriffe - identifiziert werden. Also ist 'Luxuswagen' beispielsweise ein Hyponym (Unterbegriff) von 'Wagen'. Co-Hyponyme von 'Wagen' sind z.B. Luxuswagen, Gebrauchtwagen, Neuwagen, Mietwagen, usw.
Die Struktur von Affixanwendung wird wie folgt dargestellt:

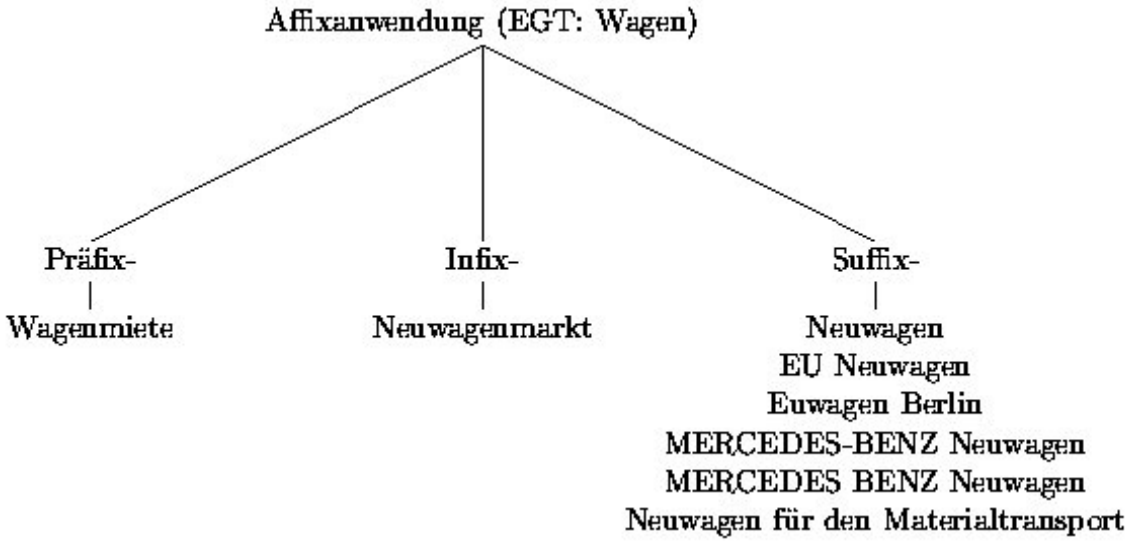


Abbildung 5.1: Struktur der Affixanwendung

Die **Präfixanwendung** kann für die Erkennung der auf Dienstleistungen bezogenen Terme unter der folgenden Voraussetzung signifikant angewendet werden. Solche Wörter, die aus einem EGT im Vorfeld- oder Mittelfeld und aus einem abstrakten Nomen für Dienstleistungen im Nachfeld bestehen, können als domainspezifisch für die auf Dienstleistungen bezogenen Terme stark berücksichtigt werden (z.B. Autovermietung, Autotuning, Autoteileverkauf, Auto-Truck-Service). 'X' bedeutet hier ein abstraktes Nomen für Dienstleistungen (z.B. Vermietung, Service, Tuning, Verkauf):

Präfix-	$W = W_1...W_n$	$W_1...W_{n-1}$ beinhaltet ein EGT. W_n ist ein X.	z.B. Autotuning

Tabelle 5.3: Präfixanwendung für die auf Dienstleistungen bezogenen Terme

Einwortterme mit Infixgebrauch von "Auto"

```
4149    Klimaautomatik/ KLIMAAUTOMATIK/ klimaautomatik/                [auto;I]
345     altautoannahme/ Altautoannahme/ ALTAUTOANNAHME                 [auto;I]
278     Altautoverwertung/ altautoverwertung/ ALTAUTOVERWERTUNG        [auto;I]
252     Altautorecycling/ altautorecycling/ alt-auto-recycling         [auto;I]
241     Altautoentsorgung/ altautoentsorgung/ Altauto-Entsorgung/      [auto;I]
169     Unfallautoverkauf/ UNFALLAUTOVERKAUF                           [auto;I]
140     Geldautomaten/ geldautomaten                                   [auto;I]
128     Altautoverordnung/ altautoverordnung                           [auto;I]
125     Gebrauchtautomarkt/ gebrauchtautomarkt/ Gebraucht-Automarkt    [auto;I]
113     Gebrauchtautoboerse/ GebrauchtautobÖRse/ GEBRAUCHTAUTOBOERSE   [auto;I]
```

Die Annotation "[auto;I]" bedeutet, dass der EGT "Auto" als Infix verwendet wird. Die obigen Top10-Einwortterme mit Infixgebrauch von "Auto" aus dem im Experiment erstellten Korpus im Automobilbereich werden als domainspezifisch erkannt. Der Infixgebrauch bringt mehr Fehltreffer als der Suffix- und Präfixgebrauch (z.B. Geldautomaten). Aber solche Fehltreffer können leicht korrigiert und verbessert werden. Mit dem Infixgebrauch können die domainspezifischen Einwortterme (z.B. Altautoannahme, Unfallautoverkauf) erkannt werden. Der Infixgebrauch wird im Experiment versuchsweise verwendet.

5.4.2 Affixanwendung für Mehrwortterme

Die Affixanwendung wird jetzt nur für Einwortterme betrachtet. Für Mehrwortterme ist sie prinzipiell gleich.
Zusammengesetzte Wörter mit Bindestrich werden als Einwortterme nach der Definition der Einwortterme in Kapitel 4.1. behandelt, weil sie kein Leerzeichen beinhalten. Im Deutschen gibt es eine starke Tendenz zu Kompositabildungen mit Hilfe von 'Bindestrich' für die Fachtermini. Obwohl die Fragestellung der zusammmengeschriebenen Wörter mit Bindestrich umstritten ist, werden in dieser Arbeit z.B. 'MERCEDES-BENZ' und "MERCEDES BENZ" als Einwort und Zweiwort unterschiedlich gezählt. Einwort-Termini

und Zweiwort-Termini spielen eine große Rolle für die deutsche Terminologie. Die meisten terminologischen Mehrwortterme sind Nominalphrasen:

```
Autositz (Nomen)
billige Autoversicherungen (Adjektiv + Nomen)
Baby Autositz (Nomen + Nomen)
ADAC Autositze (Abkürzung + Nomen)
Autositz für Kinder (Nomen + Präposition + Nomen)
Antennen für Autotelefone (Nomen + Präposition + Nomen)
Auto-Klimaanlagen und Heizungssysteme (Nomen + Konjunktion + Nomen)
```

Die Affixanwendung wird für jedes Wort in einem Mehrwortterm durchgeführt. Der Mehrwortterm wird nach domain-spezifischen Teilen abgesucht. Der domain-spezifische Teil von "Autositz für Kinder" kann im Automobilbereich "Auto" sein. Im Bereich Möbel kann "Sitz" als ein domain-spezifischer Teil betrachtet werden.
Wenn das Wort "Autositz" als EGT betrachtet wird, kann es durch "Maximum-Matching" als domain-spezifischer EGT erkannt werden.
Solche Wörter (z.B. Autositz, Autohaus), sollten als EGT behandelt und nicht weiter zerlegt werden, obwohl sie zum Komplex "Auto" gehören.

5.4.3 Affixanwendung mit Maximum-Matching

Die Affixanwendung mit Maximum-Matching bedeutet, daß längere EGT zur Erkennung der KGT bei der Affixanwendung zuerst benutzt werden.
Angenommen, beide Wörter "Autositz" und "Auto" sind EGT im Automobilbereich. Das Wort "Autositz" ist länger als das Wort "Auto". Deshalb wird das Wort "Autositz" als EGT im Ausdruck "Autositz für Kinder" identifiziert und nicht rekursiv weiterbearbeitet. Es wird durch syntaktische Analyse als Kopf der Mehrwortgruppe "Nomen_Präposition_Nomen" identifiziert.
Die Affixanwendung mit Maximum-Matching kann auf der folgenden Ebene sehr effizient verwendet werden:

- Bei der Kompositazerlegung der domainspezifischen Terme (z.B. Autositz, Babyautositz, Handschuhe)
- Bei der Erkennung des Kopfteils von domainspezifschen Termen (z.B. Lederhandschuhe, Gebrauchtwagen)

5.5 Abkürzungen und Firmennamen

Es gibt domainspezifische Listen, z.B. für Abkürzungen und Firmennamen (bzw. Marken), zur Erkennung der domainspezifischen Terme in allen Bereichen. Sie werden nicht durch die Affixanwendung, sondern durch die 'Lookup-Methode' erkannt, um Fehltreffer wegen Mehrdeutigkeiten zu vermeiden. Die 'Lookup-Methode' ist einfach: Wenn ein Term in einer Liste für Abkürzungen oder Firmennamen enthalten ist, ist er als "bereichsspezifisch" erkannt.
Verbreitete domainspezifische Abkürzungen (z.B. ADAC, Pkw, Lkw, PS, Kfz), die bei Kompositabildungen (z.B. Lkw-Kombi, Kfz-Technik) wie EGT benutzt werden, müssen für diese Affixanwendung zur Erkennung der "KGT" verwendet werden. Sie gehören zu EGT in dieser Arbeit und können semantisch unterschiedlich kodiert werden.
Wenn ein Term "Kfz-Technik" ein DST ist, muß die Abkürzung Kfz als EGT nach der Grundannahme der domainspezifischen Terme behandelt werden, weil das Wort "Technik" kein EGT ist. Solche Wörter (z.B. Technik, Test, Service, Dienst) sind domainneutral.
Firmennamen bzw. Marken (BMW, VW, Audi) sind abstrakt und domainspezifisch. Sie sollten ähnlich wie EGT behandelt werden.

BMW Deutschland	abgekürzte Marke von "Bayerische Motoren Werke AG"
VW Golf	abgekürzte Marke von "Volkswagen AG"
Fiat in Bayern	abgekürzte Marke von "Fabbrica Italiana Automobili Torino", was übersetzt bedeutet "Italienische Automobilfabrik"
Mercedes-Benz Pkws	Marke und die Pluralform von 'Pkw'
Kfz Technik	Abkürzung von 'Kraftfahrzeug'
ADAC-Test	Die Abkürzung ADAC muss als EGT behandelt werden.

5.6 Zwei CGI-Programme im Automobilbereich

Die typischen Aufgaben der Terminologieextraktion sind Termerwerb (term acquisition) und Termerweiterung (term enrichment). Die von mir geschriebenen CGI[6]-Programme in Perl werden dafür verwendet, domain-spezifische Terme zu erwerben und zu erweitern. Damit können Webseiten (bzw. Dokumente) in ihrem jeweiligen Bereich klassifiziert werden.
Die zwei CGI-Programme werden bisher in verschiedenen Bereichen, nämlich

[6] CGI / Common Gateway Interface

Autobranchen, Kosmetik, Schmuck, Wein, Möbel, Gastronomie (Restaurant), Kleidung, Büroartikel und Haushaltsgeräte experimentell eingesetzt. Daraus werden gute Ergebnisse für die Verbesserung gesammelt.
Wichtige Punkte für die Extraktion der "DST" werden am Beispiel des Automobilbereiches beschrieben.

5.6.1 EGT aus den semantischen Klassen im CISLEX

Die Qualität der EGT ist entscheidend für die Affixanwendung in allen Bereichen. Wie schon erwähnt sind die Eigenschaften von EGT domain-spezifisch und fast Einwortterme. Natürlich gibt es verschiedene Quellen für die EGT-Suche, z.B. Webseiten und Fachtexte im Automobilbereich.

Das CISLEX-Wörterbuchsystem wird für ein vollständiges, theorieneutrales elektronisches Wörterbuch des Deutschen von unserem Institut CIS ("Centrum für Informations- und Sprachverarbeitung") erstellt. Im CISLEX wurden 41.528 Lexeme für einfache Nomina von Stefan Langer semantisch manuell kodiert [Lan96]. Die Lexeme werden als Grundform eingetragen. Insgesamt sind sie in 429 semantischen Klassen hierarchisch gegliedert.
Die Klasse FAHRZEUGE (Unterklasse von ARTEFAKTE) umfaßt ca. 300 Lexeme [Lan96, S. 137].
Das Merkmal 'FZA' wird für Autos, PKWs und LKWs bei der semantischen Kodierung im CISLEX gebraucht. Durch dieses Merkmal werden 80 Lexeme identifiziert. Einige Wörter (z.B. Ambulanz, Deichsel, Vierspänner, Überwölbung) sind nicht ganz deutlich, aber sie können auf 'Autobranche' im übertragenen Sinn bezogen werden. Das Wort 'Ambulanz' kann auf Krankentransportwagen oder Rettungswagen referenziert werden. Für 'DST-Finder' werden die folgenden 80 Lexeme als EGT im Automobilbereich ausgewählt und für die Demoversion verwendet:

> Auto, Wagen, Bus, Taxi, Automobil, Mini, Mobil, Diesel, Limousine, Käfer, Ente, Kombi, Konvoi, Laster, Cabrio, Car, Wrack, Wrack, Jeep, Wohnmobil, Oldtimer, Jaguar, Cabriolet, Gespann, Karre, Kutsche, Karren, Ambulanz, Landauer, Truck, Trabant, Armatur, Spider, Karosse, Bolide, Bolid, Brummi, Pneu, Minna, Fuhre, Droschke, Cab, Landrover, Hardtop, Bulk, Deichsel, Borgia, Einsitzer, Anlasser, Zweitakter, Kabriolett, Sanka, Kabrio,

Viertakter, Vierspänner, Einspänner, Zweispänner, Sechsspänner, Überwölbung, Dreiachser, Kalesche, Camion, Kupee, Skooter, Achtachser, Achttonner, Autocar, Barutsche, Britschka, Coupé, Dreispänner, Fünfachser, Fourgon, Halbspänner, Sankra, Töfftöff, Triga, Vollspänner, Zweiachser

5.6.2 CGI-Programm 1 mit sechs verschiedenen Quellen

Domainspezifische Terme (DST) aus den schon erwähnten sechs verschiedenen Quellen in einer Webseite werden mit der schon vorhandenen Datenbank verglichen, die 2.988.819 Einträge (Metakeywords) beinhaltet. Neu gefundene Terme für die Erweiterung der DST werden in der jeweiligen neuerstellten DBM-Datenbank in Perl gespeichert.
Die aus sechs verschiedenen Quellen erkannten DST werden mit dem 6-dimensionalen Vektor wie folgt angezeigt:

[Frequenz] Term <Title, Metakeyword, Metadescription, Anckertext, Body, URL> [Semantische Annotation]
(z.B. [5] Gebrauchtwagen <1,1,1,1,1,0> [$wagen; S$])

Der Input des Programms [7] ist eine URL-Adresse. Der Output des Programms sind die erkannten DST mit verschiedenen Quellen und semantischen Annotationen. Durch das Programm wurden 156 DST aus 245 Wörtern in der Webseite - www.autoscout24.de - erkannt. Ein Term aus der URL-Adresse ist 'autoscout'. Die Top20-DST, die nach den Worthäufigkeiten sortiert wurden, werden im folgenden Beispiel gezeigt:

[7] CGI-Programm 1:
http://knecht.cis.uni-muenchen.de/cgi-bin/kimda/k04/mkw/termExtraction2WB_domain2_unitex_New2.pl

S: Suffix-Anwendung, P: Präfix-Anwendung, I: Infix-Anwendung, '= A': EGT
[5] Gebrauchtwagen <1,1,1,1,1,0> [$wagen; S$]
[4] Opel <0,0,0,0,4,0> [$Opel; Marke$]
[4] Automarkt <1,1,1,0,1,0> [$Auto; P$]
[4] AutoScout/ autoscout <1,0,0,0,2,1> [$Auto; P$]
[3] Autoverkauf <0,1,1,0,1,0> [$Auto; P$]
[3] Autokauf Autoverkauf <0,1,1,0,1,0> [$Auto; P$][$Auto; P$]
[3] Autokauf <0,1,1,0,1,0> [$Auto; P$]
[3] BMW <0,0,0,0,3,0> [$BMW; Marke$]
[3] Neuwagen <1,0,1,1,0,0> [$wagen; S$]
[3] Europas großer Automarkt <0,0,1,0,2,0> [$Auto; P$]
[2] Kia <0,0,0,0,2,0> [$Kia; Marke$]
[2] Toyota <0,0,0,0,2,0> [$Toyota; Marke$]
[2] großer Automarkt <0,0,1,0,1,0> [Auto;P]
[2] Gebrauchtwagen und Neuwagen <1,0,1,0,0,0> [$wagen; S$][$wagen; S$]
[2] Automarkt für Gebrauchtwagen <1,0,1,0,0,0> [$Auto; P$][$wagen; S$]
[2] Nissan <0,0,0,0,2,0> [$Nissan; Marke$]
[2] Auto <0,1,1,0,0,0> [$Auto; = A$]
[2] Peugeot <0,0,0,0,2,0> [$Peugeot; Marke$]
[2] Peugeot 308 RC Z <0,0,0,0,2,0> [$Peugeot; Marke$][$Z; Marke$]
[2] Skoda <0,0,0,0,2,0> [$Skoda; Marke$]

Tabelle 5.4: Top20-DST aus www.autoscout24.de (Stand:10.08.2007)

Algorithmus von CGI-Programm 1

Der folgende Algorithmus von CGI-Programm 1 zeigt die wichtigen sechs Schritte von i nach vii zur automatischen TE domainspezifischen Vokabulars aus einer Webseite und die Nutzung der vier verwendeten Listen für Stoppwörter, Firmennamen, Abkürzungen, EGT:

i. Eingabe einer oder mehrerer URL-Adressen

ii. Terme aus den fünf verschiedenen Quellen (außer Body-Teil) werden zuerst extrahiert.

iii. Nach der Entfernung von HTML-Tags und Skripten (z.B. JavaScript, Stylesheets) werden Terme aus dem Body-Teil extrahiert

iv. Stoppwörter werden durch die Stoppwortliste, die 1220 Stoppwörter für Deutsch und Englisch beinhaltet, eliminiert.

v. Firmennamen werden durch die Firmennamensliste, die 654 international bekannte Automarken beinhaltet, geprüft und annotiert.

vi. Abkürzungen werden durch die Abkürzungsliste, die 1692 domainspezifische Abkürzungen beinhaltet, geprüft und annotiert.

vii. Termkandidaten werden durch die Affixanwendung von den aus CISLEX ausgewählten 80 Lexemen überprüft und annotiert.

viii. Für erkannte domainspezifische Terme wird der 6-dimensionale Vektor erstellt und nach Frequenz sortiert.

N-Gramme mit Wortfolgen für Mehrwortterme

In der Tabelle 5.4 stehen erkannte DST für Mehrwortterme (z.B. Europas großer Automarkt). Für die Erkennung der Mehrwortterme im CGI-Programm 1 werden Satzzeichen und HTML-Tags als Begrenzer für Klumpen (Chunk, bzw. Nominalchunk) der Mehrwortgruppen benutzt.
Nicht als Zeichenfolge, sondern als Wortfolgen kann dieses N-Gramm in dieser Arbeit berücksichtigt und eingesetzt werden, um Mehrwortterme bzw. Nominalphrasen zu finden. Natürlich gibt es zahlreiche falsche Treffer dabei. Man kann solche falschen Treffer mit den folgenden strengen Betrachtungen erkennen und beseitigen:

a. Falsche Treffer, die mit Stoppwörtern oder Verben beginnen.

b. Falsche Treffer, die mit Stoppwörtern oder Verben enden.

c. Gute Treffer, die mit Nomen enden. (z.B. Ausnutzung der Groß- und Kleinschreibung für Deutsch)

Die Bi-, Tri- und Tetragramme innerhalb von Klumpen der Mehrwortgruppen werden für Mehrworttermextraktion in dieser Arbeit berücksichtigt. Nach den oben erwähnten Betrachtungen können folgende N-Gramme aus dem Beispielsatz
"*Auf Europas großem Automarkt findet jeder das passende Auto.*" extrahiert werden:

a. großem Automarkt (Bigramm)

b. passende Auto (Bigramm)

c. Europas großem Automarkt (Trigramm)

Damit kann man zumindest wichtige Kandidaten für Mehrwortterme extrahieren, indem zahlreiche falsche N-Gramme eliminiert werden.

5.6.3 CGI-Programm 2 mit Unitex für Einwort- und Mehrwortterme

Das Ergebnis für automatisch erkannte Einwortterme aus CGI-Programm 1 und CGI-Programm 2 kann wegen des gleichen Konzepts nicht unterschiedlich sein. Aber das Ergebnis für erkannte Mehrwortterme ist sehr unterschiedlich. Die statistischen Erkennungsmethoden und POS-Muster für Mehrwortterme bzw. Kollokationen erzeugen viel unnötige Termkandidaten und Fehltreffer. Bei den POS-Mustern geht es um die Qualität von POS-Taggern. Die Mehrdeutigkeiten von POS-Mustern (z.B. N/Nomen, A/Adjektiv) sind problematisch. Lokale Grammatiken mit Unitex für die Erkennung der domainspezifischen Einwort- und Mehrwortterme erzeugen wesentlich weniger unnötige Termkandidaten und Fehltreffer.
Die **Zielsetzungen von CGI-Programm 2 mit Unitex** ist, dass alle domain-spezifischen Einwort- und Mehrwortterme nach der Grundannahme von DST (Kapitel 4.3) in einem Text bzw. Korpus durch das Bootstrapping-Verfahren mit EGT fast ohne Fehltreffer erkannt werden können:

I. **Domainspezifische Einwortterme** können durch die Affixanwendung mit EGT und Überprüfung der domainspezifischen Listen erkannt werden.

II. **Domainspezifische Mehrwortterme** können durch die domain-spezifischen Graphen mit EGT erkannt werden.

Lokale Grammatiken mit Unitex

> Local grammars are finite-state grammars or finite-state automata[8] that represent sets of utterances of a natural languge. [Gro99, S. 229]

Lokale Grammatiken (Gross, 1997) werden durch endliche Automaten bzw. Transduktoren beschrieben. Sie werden nicht nur für die lexikalische Disambiguierung eingesetzt, sondern auch für die Erkennung von Mehrwortlexemen und Komposita genutzt[Bla97, S. 92].
Beispielsweise ist das Wort 'Golf' ambig, wenn man es ohne seinen Kontext betrachtet. Es ist unklar, ob es sich um Golfspiel, Automodell, Nachname usw. handelt. Deswegen sollte ein entsprechender Kontext für diese Disambiguierung im Text verwendet werden. Solche verschiedenen Umgebungen zu dem Wort 'Golf' können durch endliche Automaten erstellt werden:

Kontext mit 'Golf'	**Bedeutung**
Golf für Einsteiger	Golfspiel
Golf-Legende Arnold Palmer	Golfspiel
der neue Golf Plus bei VW	Automodell
VW Golf	Automodell
BMW Golf Cup	Golfspiel
Dr. S. Golf	Nachname

Diese oben erwähnten Umgebungen können durch lokale Grammatiken unterschiedlich typisiert werden.
Lokale Grammatiken können nicht nur für diese Disambiguierung, sondern auch für die Mehrworterkennung effizient eingesetzt werden.

Unitex

> Unitex is a corpus processing system, based on automata-oriented technology. The concept of this software was born at LADL (Laboratoire d'Automatique Documentaire et Linguistique), under the direction of its director, Maurice Gross. (aus UNITEX Homepage[9])

[8] Equivalent to regular or rational expressions, or Kleene languages.
[9] www-igm.univ-mlv.fr/~unitex [27.11.2006]

Unitex ist ein Korpusverarbeitungssystem und eine Sammlung von Programmen für die Behandlung eines natürlichsprachigen Korpustextes auf der linguistischen Ebene. Lokale Grammatiken werden mit Hilfe von Graphen visualisiert. Diese Graphen für lokale Grammatiken können durch das System 'Unitex' (oder INTEX[10]) erstellt werden.
Jeder Graph hat einen Anfangszustand, einen Endzustand, Zustandsübergänge und Aktionen (bzw. Ausgaben) wie bei endlichen Automaten. Diese Graphen werden von links nach rechts interpretiert, um bestimmte Sequenzen von Wörtern im Korpustext zu erkennen.
Unitex arbeitet mit Unicode, nämlich "UTF-16 Little Endian".

Bootstrapping-Verfahren

Im Jahr 1999 erklärte M. Gross in seinem Artikel "A Bootstrap Method for Constructing Local Grammars":

> A method for constructing local grammars around a keyword or equivalently around a semantic unit is presented. [Gro99, S. 229]

Im Artikel wurde der Schlüsselbegriff 'health' als ein Beispiel gebraucht. Empirische Kandidaten für Nominalphrasen aus einem Korpus, die auf einen Schlüsselbegriff 'health' bezogen sind, wurden für die Erstellung der lokalen Grammatiken schematisiert, um alle gesuchten Phrasen zu erkennen. Z.B. Hum (an individual human - z.B. a health minister), HumColl (a collective human body - z.B. a ministry fo health)[Gro99, S. 237].
Für unterschiedliche Bedeutungen der Schlüsselbegriffe 'health' wurden die folgenden acht Graphen im Artikel veröffentlicht:

HealthHum (humans and collective humans)	**HealthFood**
HealthPol (political activities involving health)	**HealthOther**
HealthMed (medicine and related activities)	**HealthAndN**
HealthEco (economic aspects of health)	
HealthPlace (places, texts, people)	

[10] INTEX (1994) is a linguistic development environment that includes large-coverage dictionaries and grammars, and parses texts of several million words in real time. (aus INTEX Homepage: http://intex.univ-fcomte.fr [27.11.2006])

Der Graph in der Abbildung 5.2 zeigt beispielsweise, dass das Wort 'health' als Schlüsselbegriff gebraucht werden kann und Nominalphrasen, die den Schlüsselbegriff 'health' beinhalten, durch das Muster "health and <N>" identifiziert werden können. Die Notation '<N>' bedeutet eine grammatikalische Kategorie des Nomens.

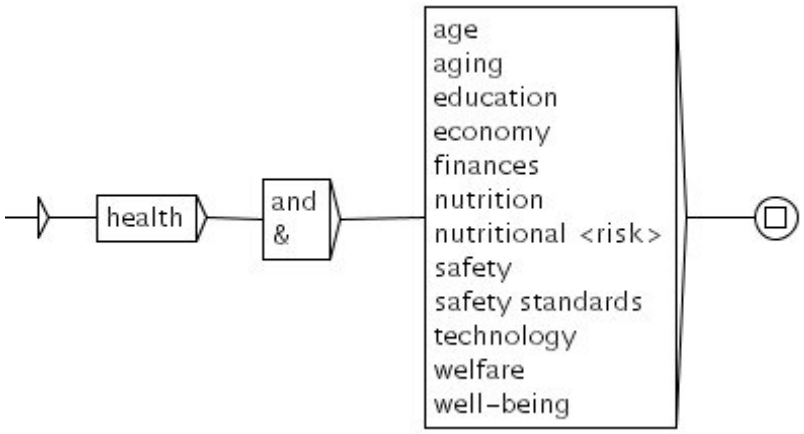


Abbildung 5.2: HealthAndN.grf aus Gross, 1999 [Gro99, S. 249]

Bootstrapping-Verfahren mit Automarken

Wie erwähnt sind Automarken domainspezifisch in der Automobilbranche. Sie spielen eine bedeutende Rolle in Fachtexten des Automobilbereichs. Im Folgenden werden die Nominalphrasen, die Automarken beinhalten, benutzt, um domainspezifische Mehrwortterme zu extrahieren. Dabei gehen wir von folgender Annahme aus:

> Eine Phrase mit Automarken ist domainspezifisch, weil Automarken domainspezifisch sind.

Um Qualität und Effizienz der Phrasen-Extraktion zu verbessern, können alle möglichen Kandidaten aus einem Korpus zuerst extrahiert und dann die Strukturen und Umgebungen der Phrasen schrittweise ermittelt werden. Aufgrund aller dieser möglichen Strukturen können dann weitere Kandidaten identifiziert werden. Diese Methode ist das schon genannte Bootstrapping-Verfahren.

Erstellung eines eigenen Lexikons für Automarken und Automodelle

Für das Bootstrapping-Verfahren wird eine semantische Einheit für Automarken in diesem Experiment berücksichtigt und verwendet.
Die 628 internationalen Automarken, die aus dem Web extrahiert und manuell nachgefiltert wurden, werden zuerst in einer Datei wie folgt definiert, um ein eigenes Lexikon für Automarken in Unitex zu erstellen:

Einträge für Automarken
Mercedes=Benz,Mercedes-Benz.AM
Mercedesbenz,Mercedes-Benz.AM
Mercedes,Mercedes-Benz.AM
Benz,Mercedes-Benz.AM
MB,Mercedes-Benz.AM
VW,VW.AM
Volkswagen,VW.AM
BMW,.AM

Mehrwortlexeme (z.B. Mercedes Benz, ALFA ROMEO) können im Lexikon direkt eingetragen werden. Das Metazeichen '=' bedeutet entweder einen Bindestrich oder ein Leerzeichen. Die Varianten für "Automarke" (z.B. VW, Volkswagen) werden wie flektierte Formen im Lexikon behandelt. Vor dem Komma stehen flektierte Formen (bzw. Varianten). Nach dem Komma stehen Lemmaformen (Grundformen). Wenn flektierte Form und Lemmaform gleich sind, kann die Lemmaform weggelassen werden (z.B. BMW,.AM). Nach dem Punkt steht eine semantische Einheit, in diesem Beispiel die Annotation "AM" für Automarken. Diese semantische Einheit wird bei der Erstellung von Graphen für diese erwähnte Nominalphrasen-Extraktion obligatorisch gebraucht.

Durch den folgenden Graph können alle Treffer für AM (Automarke) aus dem Korpus extrahiert werden, um Umgebungen für AM zu beobachten:

Abbildung 5.3: eine semantische Einheit "<AM>" für Automarke (AM.grf)

Innerhalb der spitzen Klammern (z.B. <N>, <DET>, <ADJ>) kann man syntaktisch kategorisierte Einheiten eines Lexikons in Unitex verwenden. Weil falsche Annotationen und Mehrdeutigkeiten nicht ausgeschlossen werden können, müssen sie bei der Graphen-Erstellung sehr vorsichtig verwendet werden.

Phrase für Automarken und Automodelle (bzw. konkrete Produktnamen)
In den meisten Fällen werden Automarken und Automodelle zur Verdeutlichung nacheinander im Text geschrieben (z.B. VW Golf). Mit der folgenden Struktur können Phrasen für Automarken und Automodelle erkannt werden:

<Automarke> + <Automodell> + <Zusatzinformationen für Automodell>
VW Golf Plus TDI
VW<Automarke> Golf<Automodell> Plus TDI <Zusatzinformationen für Automodell>

Tabelle 5.5: Struktur für Automarken und Automodelle

"Zusatzinformationen für Automodell" ist in dieser Struktur optional. Automarken und Automodelle sind obligatorisch. Diese Phrasen sind stark domainspezifisch. Im Experiment wurden die Phrasen, die mit einer Automarke beginnen und mit einem Automodell kombiniert sind, wie folgt extrahiert:

- *Alfa 159, Audi A2, BMW 3er, Alpina B6 Cabrio, BMW 325Ci Cabrio, Fiat 124 Sport Coupé*
- *Mercedes A-Klasse, Mercedes R-Klasse, Mercedes S-Klasse*
- *Porsche Carrera 4, Mitsubishi Lancer Evolution VIII, VW Golf Plus TDI*
- *Ford C-MAX Ambiente, Ford Focus Sport TDCi, Ford Galaxy Trend*
- *VW Transporter T3, VW T4/T5*
- *BMW X5 4.8i, OPEL ASTRA 1.7 DTL, VW Golf 2.0 TDI*
- *Seat Leon, VW Fox, VW Polo*

Für die Erkennung der letzten Fälle werden die 339 internationalen Automodelle, die im Internet sehr oft vorgekommen sind, für die Erstellung eines

eigenen Lexikons für Automodelle wie folgt formuliert:

Einträge für Automodelle
fox,.AMO
leon,.AMO
polo,.AMO

Graph für Automarken und Automodelle

Durch den Graph in der Abbildung 5.4 (am Ende dieses Kapitels) wurden die 637 signifikanten Nominalphrasen, die die oben erwähnten Phrasen beinhalten, für Automarken und Automodelle aus dem kleinen Test-Korpus erkannt. Die semantische Einheit '<AMO>' steht für Automodelle.

Graph mit Konjunktion und Präposition

Der bereits oben erwähnte Graph "AMAMO.grf" in der Abbildung 5.4 wird in dem Graphen mit Konjunktion und Präposition in der Abbildung 5.5 (am Ende dieses Kapitels) als Sub-Graph effizient verwendet. Die syntaktische Einheit "<DET>" steht für Determinator (engl. determiner). Sie wird optional im Graphen angewendet. "AMAMO" im Graphen verweist auf den Sub-Graphen "AMAMO.grf".

Dadurch wurden die erkannten folgenden 15 Treffer aus dem Test-Korpus fehlerfrei gefunden:

lus TDI DSG: Praxistest|Vergleichstest: *Alfa 147 vs. Seat Leon*|Opel Astra GTC Turbo: Test|Mitsu
nder: Fahrbericht|Nissan 350Z: Kurztest|*Alpina B3 S gegen BMW M3*|Smart Roadster Coupé: Fahrberi
: Cabrio gegen Motorrad|Vergleichstest: *BMW 325i vs. Audi A4 TFSI*|Vergleichstest: BMW 630i vs.
e-Modell.{S} Auf den Plätzen folgen der *BMW 3er und der Audi A4*.{S} Der Vorjahressieger Mazda 6
W 325i vs. Audi A4 TFSI|Vergleichstest: *BMW 630i vs. Mercedes CLS 350*|Geländetest: Nissan Pathf
bisher klar die Domäne der Konkurrenten *BMW 7er und Audi A8*.{S} Die neue S-Klasse hat mehr als
xus RX 400h: Praxistest|Vergleichstest: *BMW X3 vs. Alfa Crosswagon Q4*|BMW 525d touring: Test|Um
ahressieger, die Mercedes A-Klasse, der *Ford Focus und der Toyota Corolla* fielen dagegen ins Mi
nkt Arjeplog|Suzuki Grand Vitara : Test|*Mitsubishi Grandis vs. Mazda 5 MZR*: Vergleichstest|Pors
den wichtigsten Volumenmodellen wie dem *Opel Astra und Ford Focus* ist der Russfilter schon seit
A3 Sportback 2.0 TDI, dem BMW 120d, dem *Opel Astra 1.9 CDTI und dem VW Golf 2.0 TDI* messen
a Octavia vs. VW Passat|Vergleichstest: *Opel Zafira vs. Ford Focus* C-Max|Audi A4 Avant: Test|Me
nde im Crashtest: Topergebnisse für den *Peugeot 207 und den Alfa Romeo 159*. Der Chevrolet Aveo
t|Smart Roadster Brabus: Abschiedsfahrt|*Smart fortwo vs. Toyota Aygo*: Vergleichstest|Alfa 159:
t (Mercedes GST), erste Serienautos wie *Toyota Prius und Honda Civic* IMA werden bereits verkauf

Die Markierung {S} in Unitex bedeutet "Phrase Separator Symbol".
Die Abkürzung *vs.* steht für *versus* als Präposition (lateinisch für gegen). In

Unitex wurde diese Abkürzung als Satzendzeichen, nämlich Punkt(.), wie folgt fälschlich markiert:

Opel Zafira vs.{S} Ford Focus C-Max

In Unitex kann man solche falschen Markierungen verbessern, z.B. in dem Graphen 'abbreviation.grf' unter dem Unterverzeichnis "Graphs/Preprocessing/Sentence".

Master-Graph für Automarken und Automodelle

Die beiden schon erstellten Graphen "AMAMO.grf und AMAMOKonj.grf" werden im sogenannten Master-Graphen (Abbildung 5.6 - am Ende dieses Kapitels) für Automarken und Automodelle zusammengeführt. Dadurch wurden die insgesamt 653 qualifizierten Nominalphrasen aus dem Test-Korpus für die Gewinnung der Mehrwortterme in guter Qualität erkannt.

Bootstrapping-Verfahren mit EGT

In Kapitel 5.6.1. ("EGT aus den semantischen Klassen im CISLEX") wurden die 80 Lexeme als EGT aus CISLEX im Automobilbereich ausgewählt. Die in Kapitel 5.4. erklärte Affixanwendung mit EGT wurde für die Erkennung der domain-spezifischen Einwortterme effizient genutzt.
Diese Affixanwendung mit EGT wird für lokale Grammatiken verwendet, um domainspezifische Mehrwortterme zu erkennen.

Tokenisierung in Unitex

Die Tokenisierung bedeutet, dass ein Prozeß einen Text in Wörter und sonstige Texteinheiten aufspaltet. Tokens sind Grundeinheiten, die nicht zerteilt werden können. Die Extraktion der wort-ähnlichen Einheiten (word-like units) aus einem Text nennt man Tokenisierung [GT94].
Eine bekannte Schwierigkeit der Tokenisierung ist die Satzenderkennung, weil das Satzende oft nicht einheitlich ist:

- Abkürzungen (z.B., vs., Aug., U.S.A., Ltd., Inc.)
- Datum und Zeitangaben (07.12.2006, Do, 07.Dez.2006, 9.20 Uhr bis 13.25 Uhr)

- URLs und E-Mail-Adressen (http://www.autoscout24.de, info@tz-online.de)
- Zahlenangaben (1.200.200, 5.1.5.1.1.1 Kollokationen, 77.99 €, 1.000 - 10.000 €, 99.50 Euro, 1 EUR = 0.7028 LVL, 12.30% = 0.123)

Ein Token kann z.B. mit Bindestrich zusammengesetzte Wörter unterschiedlich definieren. Nach der Definition der Einwortterme in Kapitel 4.1. wurden zusammengesetzte Wörter mit Bindestrich als ein Token bzw. ein Wort in dieser Arbeit erkannt.
In Unitex werden die folgenden Tokens gezählt:

Einheit	**Anzahl der Tokens**
Mineralöl-Diesel	**3** (Mineralöl, -, Diesel)
VW-Käfer-Seife	**5** (VW, -, Käfer, -, Seife)
AutoScout24	**2** (AutoScout, 24)

Für die Identifizierung des Ausdrucks 'AutoScout24' werden die zwei verschiedenen Typen von Token, nämlich Wort (z.B. AutoScout) und Nummer (z.B. 24) in Unitex verwendet. Auf diese Zerlegung der Tokens in Unitex sollte bei der Graphenerstellung für die genannte Affixanwendung mit EGT geachtet werden.

Graphen für die Affixanwendung mit EGT in Unitex

Die Grundidee der Affixanwendung mit EGT ändert sich nicht. Beispielsweise wird der Graph für Präfix-Anwendung in der Abbildung 5.7 (am Ende dieses Kapitels) erstellt ('EDST' ist gleich 'EGT'). Die Struktur zwischen Präfix-, Suffix- und Infix-Anwendung ist gleich.

Die folgenden unterschiedlichen regulären Ausdrücke für die 80 ausgewählten EDST wurden bei "Morphological filters in Unitex" wie folgt verwendet:

Reguläre Ausdrücke	**Affixanwendung**
`<MOT><<^auto.*>>+<MOT><<^wagen.*>>+...`	Präfix (z.B. Autogas)
`<MOT><<auto.{0,2}$>>+<MOT><<wagen.{0,2}$>>+...`	Suffix (z.B. Gebrauchtwagen)
`<MOT><<auto>>+<MOT><<wagen>>+...`	Infix (z.B. Gebrauchtwagenbörse)

Tabelle 5.6: Reguläre Ausdrücke für die Affixanwendung mit EGT

Das Pluszeichen im Ausdruck "`<MOT><<...>>+<MOT><<...>>+`" in Unitex

steht für Vereinigung (Union). Durch den Ausdruck "<MOT><<auto>>" können Wörter, die das Wort 'auto' beinhalten, erkannt werden. Beide Ausdrücke, nämlich "<MOT>«wagen»" als 'case-insensitive' und "<MOT>«Wagen»" als 'case-sensitive' werden unterschiedlich behandelt.
Diese drei Graphen (EGTprefix, EGTsuffix, EGTinfix) erkennen die domainspezifischen Einwortterme (z.B. Gebrauchtwagen) und die folgenden Kombinationen mit Zahlen (z.B. ADAC-AutoMarxX 2006, Autohandel 50, Carrera 4). Damit können alle möglichen Umgebungen für wichtige domain-spezifische Nominalphrasen berücksichtigt werden.

Master-Graph mit EGT für die Erkennung der Einwort- und Mehrwortterme
Durch den Master-Graph in der Abbildung 5.8 (am Ende dieses Kapitels) können folgende signifikante Nominalphrasen mit Hilfe der Affixanwendung mit den 80 ausgewählten EGT erkannt werden, um domain-spezifische Mehrwortterme zu extrahieren.

Dadurch wurden die 2274 Nominalphrasen aus dem Test-Korpus wie in folgenden Beispielen erkannt:

{S} Es ist erstaunlich, wie langsam ein *Auto an Schwung* verliert - und wie weit man so kommt.{S
beeinflussen den Fahrstil.|Jedes fünfte *Auto auf Deutschlands Straßen* ist ein rollender Schrott
Holland Schweden|Top| Anbieten| Um ein *Auto in unserer Datenbank* zu inserieren, müssen Sie ein
.{S} Viele Kraftfahrer rüsten daher ihr *Auto mit einer Freisprechanlage* aus dem Zubehörregal na
|Flüssiggas zum Umbauen|Wer bereits ein *Auto mit Ottomotor* hat und die Umwelt schonen möchte, k
er Messerundgang:| AMI Leipzig 2005|15. *Auto Mobil International in Leipzig*|2. 10. April 2005|Ö
s Full-Service-Portal rund um das Thema *Auto und Motorrad* bringt AutoScout24 Verbraucher, Fahrz
er auch daran halten, wird es Zeit, das *Auto vom Winter-Dreck* zu reinigen.|Moderne Autos strotz
rkt vorbeisurfen.{S} Um möglichst viele *Auto-Anschaffungsplaner auf unser Portal* aufmerksam zu
ufinteressenten?| Über die Hälfte aller *Auto-Anschaffungsplaner in Deutschland nutzen* das Inter

Man kann unterschiedliche Tagging-Probleme aufzeigen, z.B. aus falsch getaggten Wörtern, unbekannten Wörtern und Wörtern, die verschiedenen Kategorien zuzuordnen sind. Mit Hilfe von CISLEX wurden beispielsweise die beiden Wörter 'nutzen' und 'ein' in den verschiedenen Kategorien (z.B. Nomen, Verben) grammatikalisch wie folgt in Unitex getaggt:

nutzen,.N+FF	ein,.DET:aeN:neM:neN
Nutzen,.N:aeM:deMT:geMT:neM	ein,.DET:aeNz:neMz:neNz
Nutzen,.N:aeM:deMT:neM	Ein,.EN+Hum+Nachname
Nutzen,.N:aeN:deNT:neN	ein,.VPART
nutzen,.V+intr+tr:1mGc:1mGi:3mGc:3mGi:OI	ein,einen.N+FF
nutzen,.V+refl(a)+tr:1mGc:1mGi:3mGc:3mGi:OI	ein,einen.V:eb
nutzen,.V:1mGc:1mGi:3mGc:3mGi:OI	

Daraus können unerwartete falsche Treffer gematcht werden. Deswegen sollte man bei der Graphen-Erstellung in Unitex darauf achten, die sog. POS-Muster anzuwenden. Die folgenden Wörter sind als '<EN>' (für Eigennamen) annotiert:

> A, Aber, Alle, Als, Auch, Aus, Anfang, Bank, Beginn, Bei, Das, Der, Die, Doch, Ein, Ende, Er, Firma, Geld, Glas, Im, In, Luft, Mit,...

Die Annotation '<EN>' erbrachte in Unitex 1.032 Treffer aus einer Test-Webseite (de.wikipedia.org/wiki/BMW), weil diese Annotation sehr mehrdeutig ist. Diese Annotation '<EN>' bei der Graphenerstellung sollte man daher beispielsweise wie '<EN+GEO>' in der Abbildung 5.8 (am Ende dieses Kapitels) nur für geographische Eigennamen (z.B. Berlin, China, Deuschland, Genf, Graz) eingrenzen.

Algorithmus von CGI-Programm 2 mit Unitex

Mit den Graphen 'AMAMOMaster.grf' für Automarken und Automodelle (Abbildung 5.6) und 'EGTmaster.grf' für die Affixanwendung (Abbildung 5.8) können signifikante Nominalphrasen im Automobilbereich erkannt werden. Um sonstige wichtige Nominalphrasen zu erkennen, wurde der schon verwendete Graph 'AM.grf' in der Abbildung 5.3 nach der folgenden Definition erweitert.

> Eine domainspezifische Nominalphrase beinhaltet zumindest ein EGT (z.B. BMW, Audi, Wagen, PKW).

Damit erhalten wir "AMnew.grf" in Abbildung 5.9 (am Ende dieses Kapitels) und können die folgenden Ausdrücke erkennen:

- *Golf Plus bei VW, Golf Plus bei den VW Händlern, Abenteuer mit dem kleinen Cabrio, Ausblick auf die Kleinwagenzukunft, Autogas im flüssigen Zustand wie Benzin oder Diesel, Strecken bei höherem Tempo, Auto-Bosse von Volkswagen*
- *Audi Gebrauchtwagen, TruckScout24 GmbH, AutoScout24 GmbH für den Automarkt, Stars der Leipziger Automesse*

Mit diesen oben erwähnten drei Graphen wird das CGI-Programm verbunden, um Einwort- und Mehrwortterme aus Webseiten im Internet zu extrahieren. Der Algorithmus von CGI-Programm 2 mit Unitex wird wie folgt konzipiert:

a. **Eingabe einer oder mehrerer URL-Adressen**

b. **Erstellung eines Input-Textes für Unitex**

c. **Verbindung mit Unitex für EGT**

d. **Erstellung von zwei Tabellen als Ausgabe für Einwort- und Mehrwortterme**

In den beiden Abbildungen 5.10 und 5.11 (am Ende dieses Kapitels) werden die Eingabemaske für URL-Adressen und ein Ergebnis aus dem "CGI-Programm 2 mit Unitex[11]" als Beispiel präsentiert.

CGI-Programm 2 mit Unitex und 'phpMyAdmin'

Das "CGI-Programm 2 mit Unitex" wird verbessert und mit der MySQL-Datenbank verbunden. SQL (Structured Query Language) bedeutet strukturierte Abfragesprache. In einer relationalen Datenbank werden Informationen in Tabellen gespeichert. Primärschlüssel (primary key) sind eindeutig, um relationale Verknüpfungen zwischen verschiedenen Tabellen herstellen zu können.

Nutzung von 'phpMyAdmin'

[11] CGI-Programm 2 mit Unitex liegt unter der URL-Adresse: http://knecht.cis.uni-muenchen.de/cgi-bin/kimda/k04/mkw/termExtraction2WB_domain2_unitex.pl

> phpMyAdmin is a tool written in PHP intended to handle the administration of MySQL over the Web. (Stand: 14.08.2007 [www.phpmyadmin.net/home

Die Nutzung von 'phpMyAdmin' ist sehr nützlich bei der Verbindung mit dem CGI-Programm, um Daten aus Datenbanken im Web anzuzeigen und zu manipulieren. Damit können domainspezifische Terme aus mehreren Bereichen in den jeweiligen Datenbanken gespeichert werden und neu erkannte Termkandidaten im Web zugeordnet und gespeichert werden.

Der **Algorithmus von CGI-Programm 2 mit Unitex und 'phpMyAdmin'** wird in den zwei Schritten "d und e" wie folgt zusätzlich verarbeitet:

a. **Eingabe einer oder mehrerer URL-Adressen**

b. **Erstellung eines Input-Textes für Unitex**

c. **Verbindung mit Unitex für EGT**

d. **Bewertung in Prozent, um Webseiten zu klassifizieren**

e. **Wird eine Webseite durch die Bewertung als domainspezifisch in einer Branche erkannt, so wird sie mit den MySQL-Datenbanken bzw. 'phpMyAdmin' verbunden, und es werden neue erkannte Terme im jeweiligen Bereich eingefügt.**

f. **Erstellung von zwei Tabellen als Ausgabe für Einwort- und Mehrwortterme**

Die zusätzlichen zwei Schritte (d und e) müssen vor der Verbindung mit den MySQL-Datenbanken bzw. 'phpMyAdmin' durchgeführt werden, um die Erkennung der domainspezifischen Terme aus EGT zu verbessern.
Die in Unitex erstellten Graphen für die Erkennung von Einwort- und Mehrworttermen mit EGT werden mit dem jeweiligen Fachtext entwickelt und verarbeitet. Diese domainspezifischen Graphen mit EGT können in einem domainneutralen Text mehr Fehltreffer erzeugen. Deshalb sollte eine Webseite für einen Bereich vorher durch solche Terme klassifiziert werden, die aus domainspezifischen Graphen mit EGT erkannt wurden. Dieses Verfahren wird in dieser Arbeit "EGT-Klassifikator" genannt.

5.6.4 EGT-Klassifikator

E-Commerce-relevante Webseiten können den jeweiligen Branchen mit Hilfe von EGT (Elementare generische Terme) maschinell zugeordnet werden. Die Qualität der EGT spielt eine entscheidende Rolle dafür. Im Kapitel 4.4 ("Elementare generische Terme (EGT)") wurden die Definition und Eigenschaften der EGT erklärt. Ein EGT kann zu einer Branche (z.B. Autobranche: Automobil, Fahrzeug, Wagen, Car) oder zu mehreren Branchen (z.B. Autobranche, Schmuck: Reifen) gehören.

Die jeweiligen domainspezifischen Graphen mit EGT für den EGT-Klassifikator in den zwei Branchen, nämlich Autobranchen und Haushaltsgeräte, werden im "CGI-Programm 2 mit Unitex und phpMyAdmin[12]" als Test erstellt und zur Erkennung der domainspezifischen Terme in einer Webseite durchgeführt. Dadurch wird die Prozentangabe zur Klassifikation der Webseiten berechnet.
Die Entscheidung des EGT-Klassifikators ist boolesch[13] ("domainspezifisch" oder "nicht domainspezifisch"), weil die Grenze dazwischen deutlich genug ist. Beim Test wird der Betrag von 3 Prozent für beide Bereiche als Grenze eingestellt. Sie ist beliebig einstellbar. Die Prozentangabe für eine Webseite wird wie folgt berechnet:

Die Anzahl der gesamten Wörter aus einer Webseite (GW)
Die Anzahl der erkannten Terme (ET)

$$Prozent = (100/GW) * ET$$

Eine höhere Prozentangabe bedeutet "domainspezifischer".
Bei fokussiertem Web-Crawling (Focused Web Crawling) kann der EGT-Klassifikator effizient eingesetzt werden.

In der Abbildung 5.12 wird das Ergebnis der Webseite - www.auto.de - gezeigt. Die 49 domainspezifischen Terme aus den 365 Wörtern in der Autobranche werden durch das "CGI-Programm 2 mit Unitex und phpMyAdmin"

[12] CGI-Programm 2 mit Unitex und phpMyAdmin liegt unter der URL-Adresse: http://knecht.cis.uni-muenchen.de/cgi-bin/kimda/k04/mkw/termExtraction2WB_domain2_unitex_New3Lynx2sql.pl
[13] engl. boolean

erkannt (Stand: 14.08.2007). Die Prozentangabe ist 13,42 %. Die Webseite wurde als domainspezifisch in der Autobranche betrachtet, weil die Prozentangabe höher als 3 % ist. Dann wird sie mit den MySQL-Datenbanken bzw. "phpMyAdmin" verbunden. Die durch die domainspezifischen Graphen mit EGT neu erkannten 15 Terme in der Autobranche werden mit der eckigen Klammer [New] im Web gekennzeichnet und in die entsprechende MySQL-Datenbank eingefügt.

5.6.5 Grundlagen der automatischen Klassifikation von Webseiten

> Ein Klassifikationssystem ist die strukturierte Darstellung von Klassen und der zwischen ihnen bestehenden Begriffsbeziehungen. Eine Klasse ist die Zusammenfassung derjenigen Begriffe, die mindestens ein identisches Merkmal (Klassem) haben. Ein Klassem (oder klassifikatorisches Merkmal) ist dasjenige Merkmal von Begriffen, das zur Bildung einer Klasse benutzt wird, und diese von anderen Klassen unterscheidet. Jede Klasse muss verbal durch eine Klassenbenennung bezeichnet werden. [Holger Nohr]

Schnell wachsende Webseiten, die in HTML geschrieben sind, stellen eine unvermeidlich unstrukturierte Informationsquelle dar. Das bedeutet, dass Webseiten in einer natürlichen Sprache (z.B. in Englisch, Deutsch, Französisch und Koreanisch) geschrieben sind.
Ein Klassifikationssystem ist absolut nötig, um relevante Informationen aus riesigen Webseiten zu finden. Durch Klassifikationssysteme können Webseiten jeweils entsprechenden Klassen zugeordnet werden, in denen gleichartige Objekte und Themen zusammengefasst sind. Die Zuordnung zu Klassen nennt man auch eine Gruppierung von verwandten Webseiten bzw. Dokumenten. Dafür werden die vier wichtigsten Komponenten - Baumindexierung, Vokabula (Lexikon), Morphologie, Semantik - verwendet.
Manuelle Klassifikation von Webseiten ist sehr zeitaufwendig und teuer, wie z.B. der Yahoo-Katalog. Menschen beurteilen die Webseiten und klassifizieren sie von Hand. Etwa 1,5 Millionen Webseiten kommen jeden Tag im WWW hinzu.

Die zunehmende Information kann nicht mehr manuell, sondern muss automatisch verarbeitet werden. Über die Notwendigkeit einer automatischen Klassifikation in Verbindung mit einer Suchmaschine schreibt T. Koch:

> Automatische Klassifikationsprozesse werden notwendig, wenn große robotergenerierte Dienste eine gute Navigationsstruktur für ihre Dokumente oder erweiterte Filtertechniken, sowie geeignete Anfragemöglichkeiten zur Verbesserung des Suchprozesses anbieten wollen. [Desire 1997]

Diese automatische Zuordnung von Objekten zu vorgegebenen Kategorien nennt man automatische Klassifikation (supervised learning)[14].

Kommerzielle Webseiten können inhaltlich automatisch analysiert werden und mit Hilfe von EGT zu vorher festgelegten Branchen automatisch zugeordnet werden.

Klassifikationsverfahren

Es gibt im wesentlichen folgende 3 Verfahren zur Dokumentklassifikation:

- **Statistische Verfahren** beschäftigen sich nur mit der Häufigkeit der Worte bzw. deren Position im Text.
- **Linguistische Verfahren** beschäftigen sich fast alle mit dem Auffinden von Phrasen bzw. Mehrworttermen. Sie sind morphologische, syntaktische sowie semantische Verfahren. Dafür werden Lemmatisierung (Grundformreduktion), Kompositazerlegung und Phrasenerkennung allgemein verwendet.
- **Begriffsorientierte Verfahren** abstrahieren nun die Bedeutung der vorgefundenen Wörter und versuchen, den Inhalt eines Textes zu erfassen. Es handelt sich z.B. bei den Termen "Computer" und "Rechner" um die sprachliche Repräsentation einer Bedeutung. Begriffsorientierte Verfahren spielen zur Zeit in der Praxis noch keine große Rolle [Nohr].

[14] Aufgabe des Clusterung:
automatische Gruppierung von Objekten zu vorher unbekannten Kategorien (unsupervised learning).

Die linguistischen und begriffsorientierten Verfahren sind mit hohem Implementierungsaufwand wegen des großen Speicherplatzbedarfs verbunden.

Statistische Klassifikationsalgorithmen
Es gibt mehrere Verfahren mit verschiedenen Vor- und Nachteilen:

- Maximum Entropy Modelling
- Decision Trees
- Vector Space Model
- Support Vector Machines (SVM)
- Entscheidungsbäume
- Neuronale Netze
- Latent Semantic Analysis (LSA)
- Bayes-Klassifikator (Naive Bayes)
- K-Nearest-Neighbour-Verfahren (KNN)

5.7 Schlußfolgerung

Bei dem "CGI-Progrmm 2 mit Unitex für Einwort- und Mehrwortterme" sind die verwendeten EGT in den jeweiligen Bereichen schon festgelegt, um domainspezifische Graphen zur Termerkennung zu erstellen. Dadurch können domainspezifische Terme aus EGT bei der Verbindung mit den MySQL-Datenbanken bzw. 'phpMyAdmin' in einer Webseite erkannt werden, um eine Terminologie in den jeweiligen Bereichen zu erstellen.

Die Anzahl der EGT in den jeweiligen Bereichen muss abzählbar sein. Schätzungsweise können es 5000 - 10000 EGT pro Bereich sein. Keine Auswertung von "Precision und Recall" wird ausgeführt, weil die als Test verwendeten ca. 90 EGT verbessert und erweitert werden müssen.

Im Test wird gezeigt, dass domainspezifische EGT in den jeweiligen Bereichen eine Basis sind für die in Kapitel 3.6 erwähnten drei Anwendungen, nämlich Erstellung von Fachwörterbüchern, Verbesserung von Suchmaschinen und Verbesserung der maschinellen Übersetzung.

Durch EGT und Firmennamen können domainspezifische Terme in den jeweiligen Bereichen erfasst werden.

Das "CGI-Programm 1 mit sechs verschiedenen Quellen" ist eine dynamische TE domainspezifischen Vokabulars aus einer Webseite. Durch die unterschiedliche Termgewichtung für Einwort- und Mehrwortterme aus den verschiedenen Quellen können EGT dynamisch erkannt und erweitert werden. Eine Kombination von CGI-Programm 1 und CGI-Programm 2 ist zur Qualitätsverbesserung und dynamischen Erkennung von EGT sehr hilfreich.

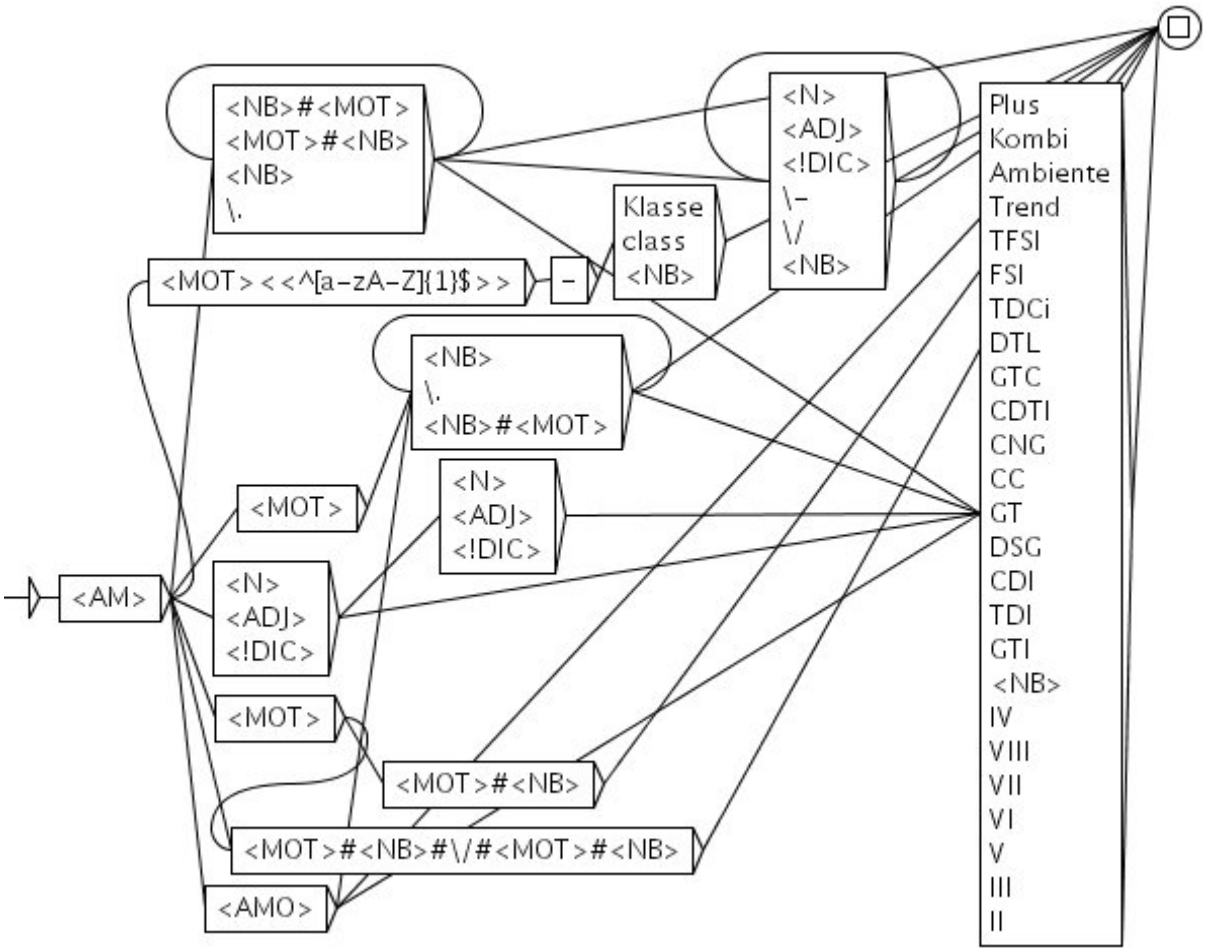


Abbildung 5.4: eine semantische Einheit "<AMO>" für Automodelle (AMAMO.grf)

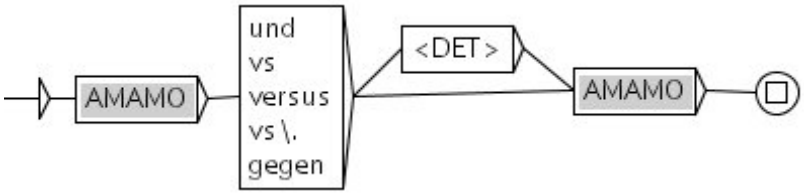


Abbildung 5.5: Graph mit Konjunktion u. Präposition (AMAMOKonj.grf)

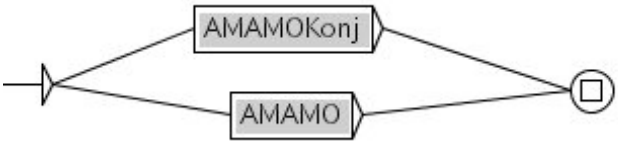


Abbildung 5.6: Master-Graph für Automarken und Automodelle (AMAMOMaster.grf)

Abbildung 5.7: EGTprefix.grf: Graph für die Präfix-Anwendung von EGT

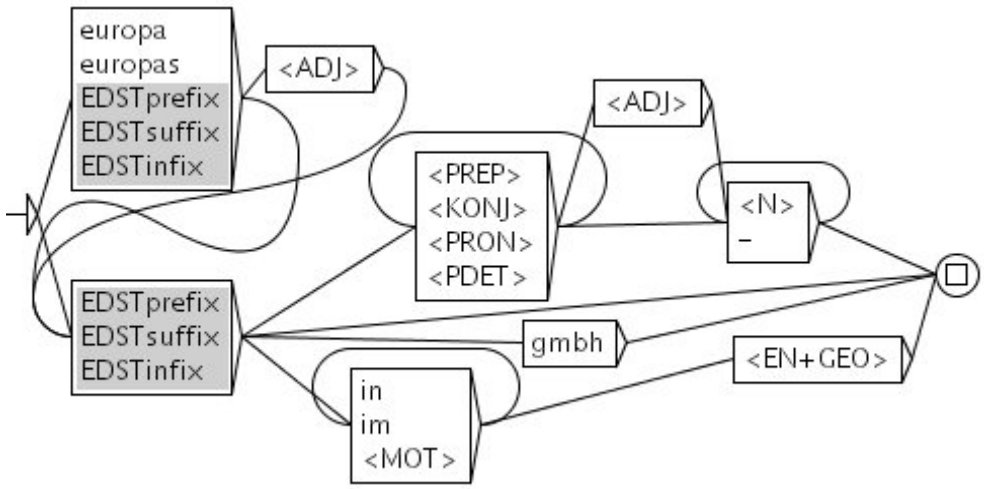


Abbildung 5.8: Graph für die Extraktion der Mehrwortterme (EGTmaster.grf)

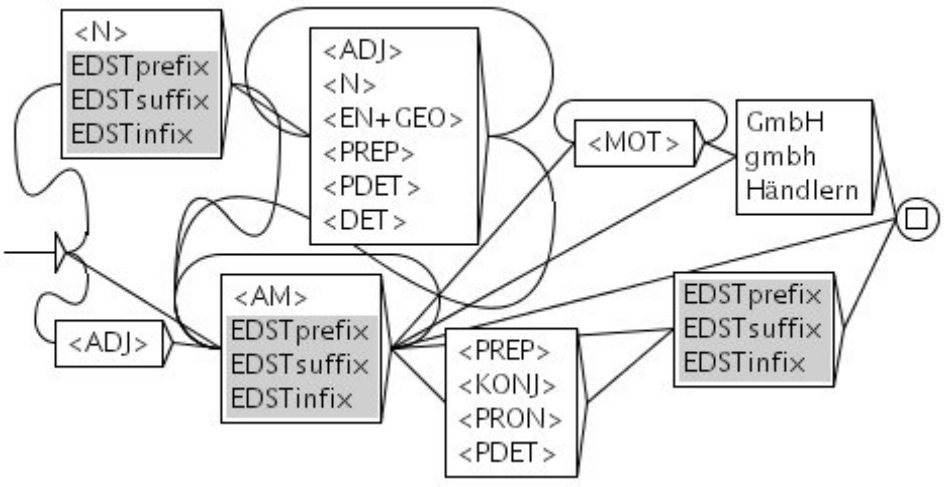
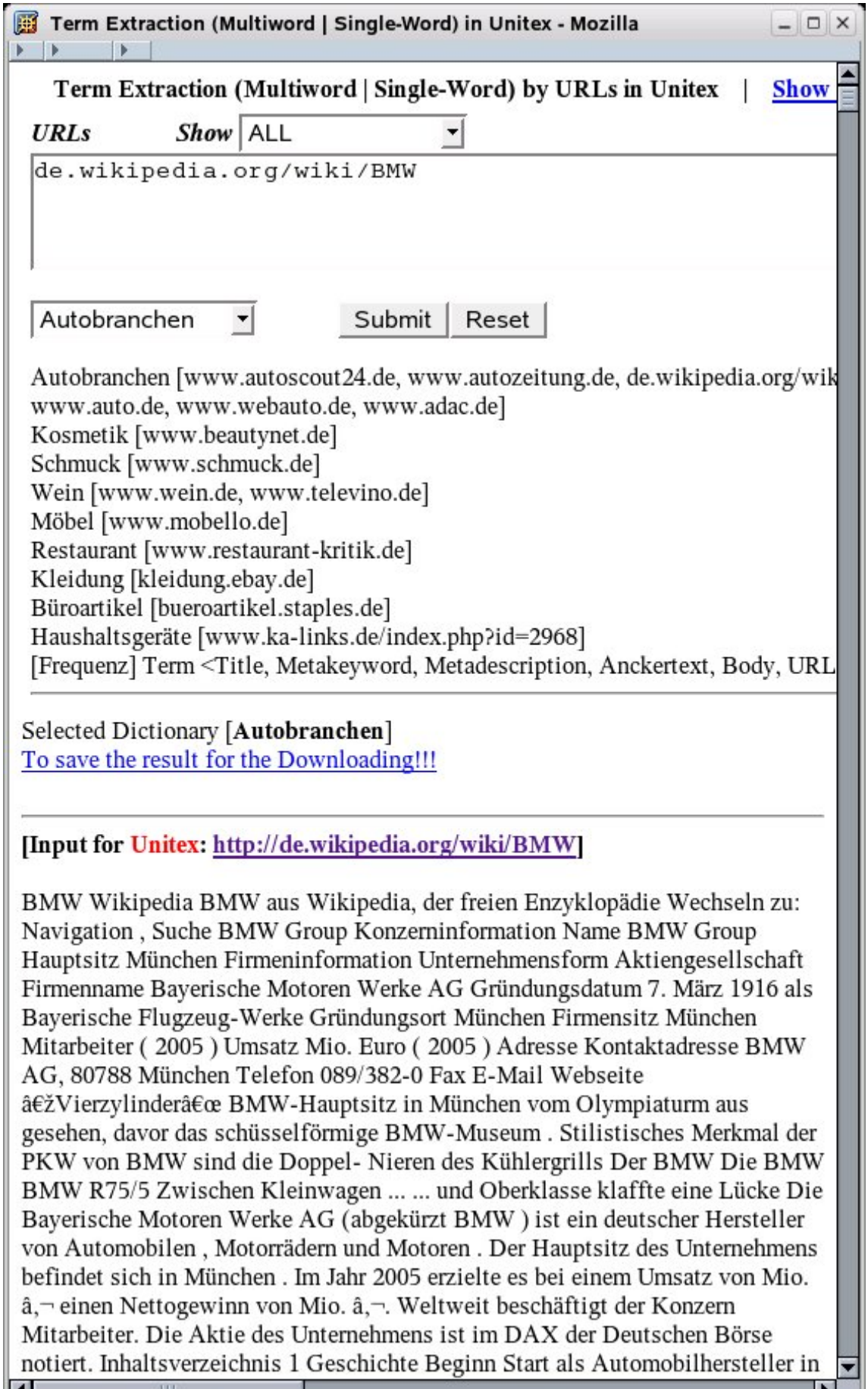


Abbildung 5.9: Graph für sonstige wichtige NP (AMnew.grf)


Term Extraction (Multiword | Single-Word) in Unitex - Mozilla
Term Extraction (Multiword | Single-Word) by URLs in Unitex | Show
URLs Show ALL
de.wikipedia.org/wiki/BMW
Autobranchen
Submit
Reset
Autobranchen [www.autoscout24.de, www.autozeitung.de, de.wikipedia.org/wik
www.auto.de, www.webauto.de, www.adac.de]
Kosmetik [www.beautynet.de]
Schmuck [www.schmuck.de]
Wein [www.wein.de, www.televino.de]
Möbel [www.mobello.de]
Restaurant [www.restaurant-kritik.de]
Kleidung [kleidung.ebay.de]
Büroartikel [bueroartikel.staples.de]
Haushaltsgeräte [www.ka-links.de/index.php?id=2968]
[Frequenz] Term <Title, Metakeyword, Metadescription, Anckertext, Body, URL
Selected Dictionary [Autobranchen]
To save the result for the Downloading!!!
[Input for Unitex: http://de.wikipedia.org/wiki/BMW]
BMW Wikipedia BMW aus Wikipedia, der freien Enzyklopädie Wechseln zu: Navigation , Suche BMW Group Konzerninformation Name BMW Group Hauptsitz München Firmeninformation Unternehmensform Aktiengesellschaft Firmenname Bayerische Motoren Werke AG Gründungsdatum 7. März 1916 als Bayerische Flugzeug-Werke Gründungsort München Firmensitz München Mitarbeiter (2005) Umsatz Mio. Euro (2005) Adresse Kontaktadresse BMW AG, 80788 München Telefon 089/382-0 Fax E-Mail Webseite â€žVierzylinderâ€œ BMW-Hauptsitz in München vom Olympiaturm aus gesehen, davor das schüsselförmige BMW-Museum . Stilistisches Merkmal der PKW von BMW sind die Doppel- Nieren des Kühlergrills Der BMW Die BMW BMW R75/5 Zwischen Kleinwagen ... ... und Oberklasse klaffte eine Lücke Die Bayerische Motoren Werke AG (abgekürzt BMW) ist ein deutscher Hersteller von Automobilen , Motorrädern und Motoren . Der Hauptsitz des Unternehmens befindet sich in München . Im Jahr 2005 erzielte es bei einem Umsatz von Mio. â,¬ einen Nettogewinn von Mio. â,¬. Weltweit beschäftigt der Konzern Mitarbeiter. Die Aktie des Unternehmens ist im DAX der Deutschen Börse notiert. Inhaltsverzeichnis 1 Geschichte Beginn Start als Automobilhersteller in

Term Extraction (Multiword | Single-Word) in Unitex - Mozilla

[Result with Unitex: http://de.wikipedia.org/wiki/BMW (All 261 domain specific terms found.)]

domain-specific single-word Terms (97 found.)	domain-specific multi-word Terms (164 found.)
Alpina	Aero 8
Austin	Audi A8
Auto	Autokonzern mit seinen Marken BMW
Autobauer	Automobile sowie BMW-Motorräder
Autoherstellung	Automobile und Motorräder
Autohändler	Automobilen und Motorrädern
Autokonzern	Automobilhersteller in Eisenach
Automobil-Hersteller	Automobilmarkt von Anfang
Automobile	Automobilsalon in Genf
Automobilen	Automodell 325
Automobilhersteller	BMW 1500
Automobilkonstruktion	BMW 1600-2
Automobilmarkt	BMW 1800 und BMW 2000
Automobils	BMW 1898â
Automobilsalon	BMW 1er
Automobilsektor	BMW 2000
Automobilwerk	BMW 2002
Autos	BMW 3/15 PS
Autotorâ	BMW 320d
BMW	BMW 328 Mille Miglia
BMW-Antriebskomponenten	BMW 3er
BMW-Reihensechszylinder-Dieselmotor	BMW 3er-Reihe
Beiwagen	BMW 501
Bentley	BMW 501/502
Benz	BMW 507
Bond-Car	BMW 5er
Borgward	BMW 600
Cabrio	BMW 645Ci
Cabriolet	BMW 6er
Coupé	BMW 700 LS Coupé

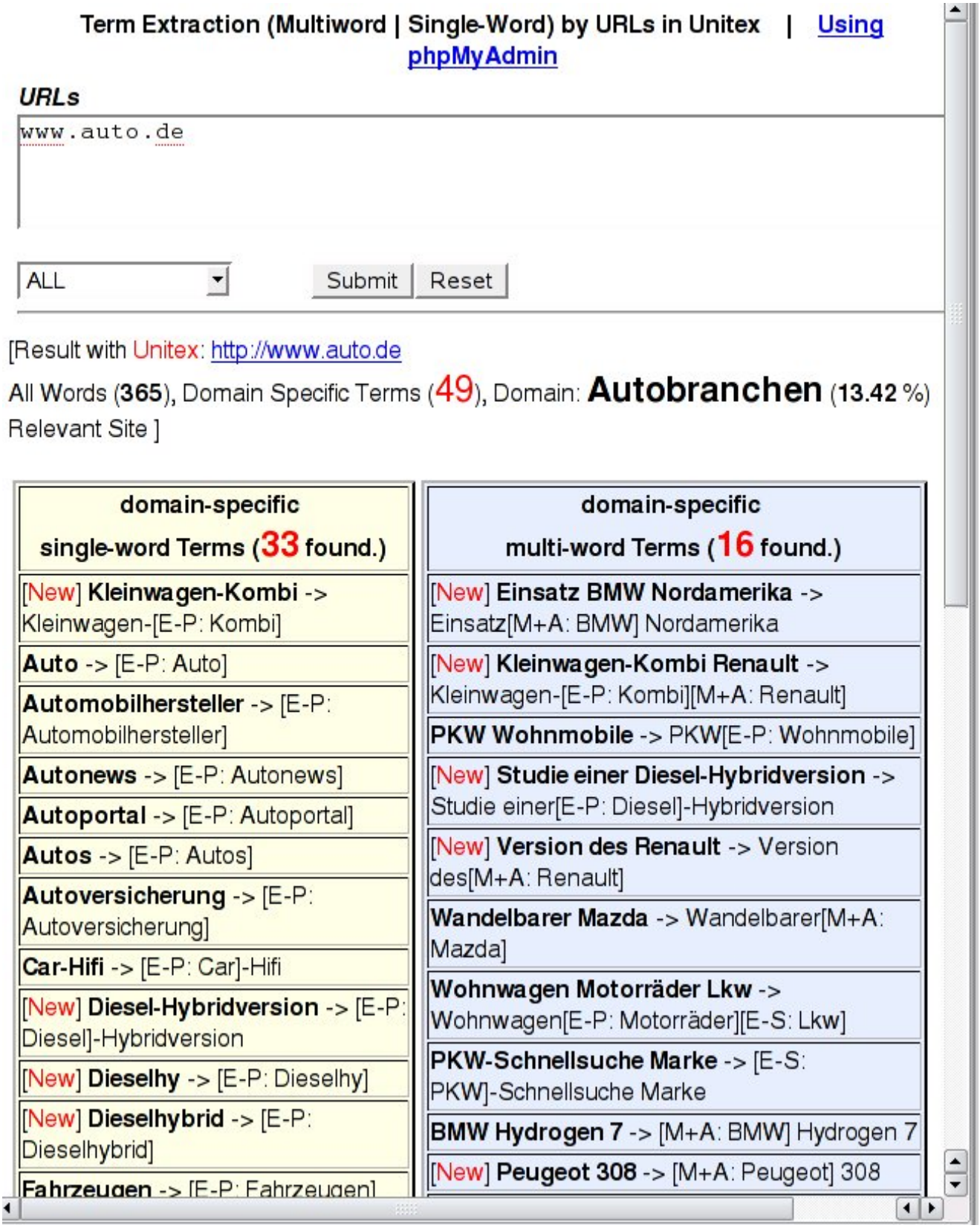


Abbildung 5.12: CGI-Programm 2 mit Unitex und phpMyAdmin (Stand: 14.08.2007)

Kapitel 6

Domainspezifische Korpora aus dem Web

6.1 Definition des Korpus

Es gibt verschiedene Definitionen für "Korpus". Nach "McEnery und Wilson" wird der Korpus wie folgt definiert:

> In principle, any collection of more than one text can be called a corpus. But the term "corpus" when used in the context of modern linguistics tends most frequently to have more specic connotations than this simple definition. These may be considered under four main headings: sampling and representativeness, finite size, machine-readable form, a standard reference. [KG03]

Für meinen Zweck kann ein Korpus als eine Dokumentensammlung aus einem Bereich betrachtet werden, um domainspezifische Einwortterme und Mehrwortterme in diesem Bereich zu erkennen und zu erweitern. Ein Korpus wird in dieser Arbeit wie folgt definiert:

> Ein Korpus ist eine Dokumentensammlung in einem Bereich. Sie beinhaltet eine Menge von bereichsspezifischem Vokabular, das in den jeweiligen Bereichen verschieden ist.

Das Web als Korpus für linguistische Daten ist zur Zeit sehr nützlich, weil es sich stetig verändert, sehr groß, maschinenlesbar und frei zugänglich ist. Die Anzahl der Webseiten (Hypertextdokumente) und Publikationen im Internet

wächst sehr schnell.
Das Web erfüllt also die vier oben erwähnten Hauptpunkte.
In den 60er Jahren wurde das erste bekannte Computerkorpus "Brown Corpus" aus 500 englischen Texten erzeugt. Es bestand aus 15 verschiedenen Bereichen (Genres), die ungefähr eine Millionen Wörter beinhalteten. Seine linguistischen Informationen wurden manuell annotiert. Wegen dieses hohen Aufwands war es fast fehlerfrei. Im Gegensatz dazu können domain-spezifische Korpora aus dem Web mit Hilfe von Tokenisierung, Lemmatisierung und POS-Tagging (engl. part-of-speech tagging / oder POS-Annotation) mühelos automatisch annotiert werden.

6.2 Aufbau der Korpora

Im Internet verfügbare Zeitungen ('news.google.de', 'de.news.yahoo.com' und 'de.news.yahoo.com'), Zeitschriften ('www.spiegel.de', 'focus.msn.de') und Fachtexte können für die deusche Sprache verwendet werden.

Das **Ziel der Korpora aus dem Web** ist, domainspezifische Korpora für die deutsche Sprache aus dem Web automatisch zu erstellen, um Einwort- und Mehrwortterme (bzw. Phrasen) zu erkennen und zu erweitern. Die Webseiten eines Korpus haben die folgenden Eigenschaften:

- **In HTML strukturiert**
- **Monolingual** (nur deutsche Webseiten)
- **E-Commerce-Seiten**
- **Automatisch annotierbar**

6.3 Dokumentensammlung

Die vier häufigsten MIME-Typen[1] sind 'text/html', 'image/gif', 'image/jpeg' und 'text/plain'.

[1]Multipurpose Internet Mail Extensions (MIME) ist ein Kodierstandard, der die Struktur und den Aufbau von E-Mails und anderer Internetnachrichten festlegt. MIME ermöglicht es, zwischen Sender und Empfänger Informationen über den Typ der übermittelten Daten auszutauschen (Content-Type) und

Im Experiment werden Webseiten mit dem Typ 'text/html' lokal heruntergeladen, um domainspezifische Terme (DST) zu extrahieren.
Die unterschiedlichen Kodierungen und HTML-Entitäten werden in der Kodierung "ISO 8859-1 (ISO Latin 1)" normalisiert.
Webseiten werden "case sensitive" für die deutsche Sprache lokal gesammelt und bearbeitet. Zahlreiche signifikante Nominalphrasen, die für Werbung und Produktnamen manuell verarbeitet werden, können in Webseiten zwischen HTML-Tags extra betont werden. Um solche manuell bearbeiteten Nominralphrasen in Webseiten zu erkennen, werden HTML-Tags beispielsweise wie folgt durch das Senkrecht-Zeichen '|' im Korpus ersetzt und können als Delimiter (Trennsymbol) neben Satzendzeichen (., ?, !) nützlich gebraucht werden. Sie sind gute Kandidaten für DST. Solche HTML-Strukturen können bei der Terminologie-Extraktion ausgenutzt werden.

Facelift BMX[2]X3	|Facelift BMX X3|
BMW 525 eA mit G-Kat	|BMW 525 eA mit G-Kat|
<td ...>DTM[3]-Saison 2006</td>	|DTM-Saison 2006|

Eigene Methoden zur Erkennung der verschiedenen Typen von Entitäten (Personen, Organisationen/Produktnamen, Lokationen, Temporalia und Ereignisse) können dann zusätzlich entwickelt werden.

Durch eine URL-Adresse wird auf eine HTML-Seite verwiesen. Das bedeutet, dass man diese Art von Korpus auch als Sammlung von URL-Adressen betrachten kann. Die folgenden zwei Methoden für Dokumentensammlung bzw. URL-Sammlung können angewendet werden:

a. Extraktion aus Startseiten (z.B. www.autoscout24.de, www.schmuck.de)

b. Extraktion mit Suchmaschinen (z.B. Google, Yahoo)

gleichzeitig eine für den verwendeten Übertragungsweg sichere Kodierung (Content-Transfer-Encoding) festzulegen. (aus 'Wikipedia')

[3]'BMX' ist ein Automodell von 'BMW'.

[3]DTM für "Deutsche Tourenwagen Masters"

6.3.1 Extraktion aus Startseiten

Ein Crawler (auch Spider oder Robot) braucht mindestens eine Startseite oder eine Liste für 'Start-Urls', um Daten aus dem Internet z.B. für Suchmaschinen zu sammeln.
Das Hauptziel der von mir erstellten speziellen Spider ist es, verschiedene URL-Adressen aus einem spezifischen Bereich zu sammeln und Webseiten ohne HTML-Tags nur inhaltlich in Dateien, in denen die Anzahl der Webseiten freilich beschränkt wird, zusammen zu setzen. Damit kann man identifizieren, aus welchen Seiten und welchen Teilen der Webseiten Terme aufgefunden wurden.
Der Spider mit einer Startseite besteht aus zwei Teilen, nämlich der Sammlung der verschiedenen URLs einerseits und der HTML-Analyse anderseits:

a. **Sammlung der verschiedenen URLs**

- Eine Startseite einlesen und Links crawlen
- Relative Linkpfade in absolute Linkpfade umwandeln
- Interne Links unter einer bestimmten Pfadangabe werden weiter verfolgt, indem verschiedene URLs aus einem gesuchten Bereich gesammelt werden.
- Webseiten können durch diese ausgewählten URLs nach Bedarf lokal heruntergeladen werden.

b. **HTML-Analyse**

- HTML-Tags werden durch das Senkrecht-Zeichen '|' im Korpus ersetzt, um als Trennsysmbol neben Satzendzeichen (.,?,!) zu dienen.
- Jede Webseite wird durch die o.g. 'HTML-Analyse' nach sechs schon beschriebenen Quellen einer Webseite analysiert und HTML-Tags werden entfernt.
- Webseiten ohne HTML-Tags werden inhaltlich in ein XML-ähnliches Format gebracht. Innerhalb dieser Dateien kann die Anzahl der Webseiten auf zum Beispiel 100.000 begrenzt werden.

6.3.2 Extraktion mit Suchmaschinen

Suchmaschinen liefern allgemein Suchergebnisse auf Grundlage von Suchbegriffen. Man braucht also spezifische Suchbegriffe, um domain-spezifische Dokumente mit Hilfe der bekannten Suchmaschinen zu sammeln.
Der Spider auf Basis von bekannten Suchmaschinen besteht auch aus zwei Teilen:

a. **Sammlung der verschiedenen URLs**

- Eine Datei, die ausgewählte Schlüsselwörter aus einem spezifischen Bereich beinhaltet, wird eingelesen, um domainspezifische Abfragen für Suchmaschinen automatisch zu erstellen, die z.B. aus zwei bis sechs Wörtern bestehen sollten.
  Beispielhafte Suchbegriffe im Automobilbereich:

  Autobranche Fahrzeuge
  Autobranche Fahrzeuge BMW
  Autobranche Fahrzeuge BMW KFZ
  Autobranche Fahrzeuge BMW KFZ PKW
  Autobranche Fahrzeuge BMW KFZ PKW LKW
  (N-Gramme mit dem Basiswort "Autobranche")

- Verschiedene URL-Adressen werden aus Suchergebnissen, die von bekannten Suchmaschinen geliefert werden, in einer Datei gesammelt.
- Webseiten können über gesammelte URLs nach Bedarf lokal gespeichert werden.

b. **Die HTML-Analyse bleibt gleich.**

Diese zweite Methode auf Basis von bekannten Suchmaschinen kann für die Anwendung der HTML-Analyse angemessen sein, weil die sechs Quellen der Webseiten vielseitiger sind als die der ersten Methode. Aber die erste Methode mit einer Startseite oder einer Liste für Startseiten ist einfach und zielgerichtet, um Korpora automatisch aufzubauen.

6.4 Beispiel für lokal gespeicherte Webseiten

Mit der HTML-Analyse können Wörter aus den folgenden sechs Quellen einer Webseite erkannt und im Korpus wie folgt gekennzeichnet werden. Dafür werden die folgenden Bezeichnungen definiert:

Bezeichnungen der HTML-Analyse
<**id...**> für ID-Nummer der Webseiten
<**u**>...</**u**> für URL-Adresse und Größe einer Webseite
<**t**>...</**t**> für Titelangabe
<**mk**>...</**mk**> für Metakeywords
<**md**>...</**md**> für "Meta Description"
<**at**>...</**at**> für Ankertext
<**b**>...</**b**> für Inhalt von 'HTML-Body'
</**id**> für Ende-Tag einer Webseite
HTML-Tags in der Original-Webseite werden durch das Zeichen '|', das als Satzendezeichen interpretiert werden kann, ersetzt.

Tabelle 6.1: Bezeichnungen der HTML-Analyse

```
<id="2">
<u>http://www.wlw.de/rubriken/laptop.html [Size: 12902 Bytes]</u>
<t>Laptop bei "Wer liefert was?" - der großen Lieferantensuchmaschine</t>
<mk>Laptop</mk>
<md></md>
<at>Diebstahlsicherungen für Notebooks|Laptop-Computer|Notebook-Akkus|Notebook-Halterungen
für Kraftfahrzeuge|Notebook-Koffer|Notebook-Reparaturen|Notebooks|Notebook-Ständer|
Notebook-Taschen|Notebook-Reparaturen|Notebook-Ständer|Notebook-Akkus|Diebstahlsicherungen
für Notebooks|Bohren|Schwamm|Chlormessgeräte|Hormone|Eiskratzer|Fittings</at>
<b>|Die Lieferanten- suchmaschine|Sie suchen|beruflich|nach|Laptop|? Bei uns finden Sie|
umfassende Anbieterinformationen|inklusive Angaben zur persönlichen Kontaktaufnahme.
|Ihre Treffer zu|Laptop|(9)|Laptop bei "Wer liefert was?"|
...
</b>
</id>
```

6.5 Schwierigkeiten beim Aufbau der Korpora

Die folgenden Schwierigkeiten beim Aufbau der Korpora sind in diesem Experiment aufgetaucht:

- Entfernung von Duplikaten und Quasi-Duplikaten
- Komprimierte Dateien aus dem Netz herunderladen
- Erkennung einer Cookie-Seite beim Herunterladen

6.5.1 Entfernung von Duplikaten und Quasi-Duplikaten

Die Vermeidung von Duplikaten ist für Suchmaschinen eine wichtige Aufgabe. In einem Korpus bzw. einer Dokumentensammlung sollte man diese zur Qualitätsverbesserung eines Korpus beseitigen. Die Definitionen von 'Duplikaten' und 'Quasi-Duplikaten' in dieser Arbeit sind folgende:

- Gleiche URL-Adressen sind 'Duplikate'.
- Gleiche Dateinamen und Größe sind 'Duplikate'.
- Fast ähnliche Webseiten, z.B. automatisch erstellte Serienbriefe mit zahlreichen verschiedenen Usernamen oder Produktnamen, werden als 'Quasi-Duplikate' betrachtet.

Eine Webseite hat mindestens eine URL-Adresse. Mehrere URL-Adressen können auf eine Webseite verweisen. Um Duplikate zu entfernen, wird folgende Überprüfung in drei Schritten durchgeführt:

a. Pfad-Alias und Links innerhalb von Webseiten werden normalisiert.
   z.B. gleiche Domainadressen mit verschiedenen Postfixen wie index.html, default.html, welcome.html, index.htm, index.asp, index.shtml, index.htm#firmennamen, etc.

b. URL-Adressen auf Gleichheit prüfen.

c. Bei gleichen Dateinamen auf gleiche Größe prüfen.
   Mit dem Zugriff auf die HTTP[4]-Header mit LWP[5]-Methoden kann man die jeweilige Größe einer Datei wissen, wenn sie in der Header-Information vorhanden sind. Keine Angabe bedeutet keine Größe. Deshalb wird eine Datei lokal temporär gespeichert, um ihre Größe sicher zu bestimmen.

[4]HTTP (HyperText Transfer Protocol)
[5]LWP (Library for WWW in Perl)

Bei der Entfernung der 'Quasi-Duplikate' müssen ähnliche für die Termextraktion unwichtige Pfadangaben z.B. bei Serienbriefen oder fast gleichen Mitteilungen berücksichtigt werden. (z.B. xxxlogin.asp?xxx, xxxsearch.asp?xxx,...) Nach dem Herunterladen der Webseiten kann man die Ähnlichkeit der Webseiten mit verschiedenen Algorithmen ausführlich vergleichen.

6.5.2 Komprimierte Dateien aus dem Netz herunderladen

Im Paket LWP (Library for WWW in Perl) gibt es zwei bekannte Module, um auf Webseiten zuzugreifen: LWP::Simple und LWP::UserAgent
LWP::Simple stellt nur drei Funktionen (get, getprint und getstore) zur Verfügung. Der erweiterte Zugriff mit LWP::UserAgent hat mehrere nützliche Funktionen. Deshalb wurde das Programm in dieser Arbeit mit LWP::UserAgent programmiert, um Dokumente aus dem Web lokal zu speichern. Der MIME-Typ 'text/html', der auf Webseiten angegeben ist, stellt sicher, daß es sich bei der jeweiligen Seite um eine HTML-Seite handelt (nicht um eine PDF- oder PS-Seite[6]). Nur die HTML-Seiten werden gecrawlt. Da HTML-Seiten auch komprimiert auftreten können, müssen sie vor dem Herunterladen dekomprimiert werden. Danach können die regulären Ausdrücke ausgeführt werden.

6.5.3 Erkennung einer Cookie-Seite beim Herunterladen

> Cookies[7] sind Dateien, die beim Besuch mancher Websites auf der Festplatte des Besuchers abgelegt werden. Cookies sind Text-Dateien und dienen der (Zwischen-)Speicherung von Daten. Cookies dürfen eine maximale Größe von 4 KB nicht uberschreiten.

Webseiten, die Cookies verlangen, stellen folgendes Problem dar:
Wenn der Webbrowser oder Crawler die Cookies nicht akzeptiert, kann nicht in die tieferliegende Seitenstruktur dieser Webseite vorgedrungen werden. Die Informationen auf den tieferliegenden Seiten gehen so verloren. Um dieses Problem zu beheben, muß der Crawler die Cookies erkennen und akzeptieren. Extrahierte URLs für ein Korpus aus dem Web können auf HTML-Seiten verweisen, die Cookies verwenden. Beim Herunderladen sollten sie erkannt werden,

[6] PDF (Portable Document Format), PS (PostScript)
[7] Stand: 19.12.2006 http://www.lexikon-suchmaschinenoptimierung.de/c.htm

um die HTML-Seite lokal zu speichern. Mit dem Perl-Modul HTTP::Cookies kann man Cookies erkennen und verarbeiten.

6.6 Extraktion der Einwortterme

Signifikante Einwortterme und Mehrwortterme aus einem Bereich (z.B. Automobilbereich) ergeben domain-spezifische Informationen, die manuell oder automatisch annotiert werden können. Diese Informationen (z.B. Automarken, Produktnamen, Abkürzungen und EGT) sollten automatisch erstellt und manuell verbessert werden. Sie können zur Dokumentenklassifikation und semantischen Analyse von Dokumenten sehr nützlich sein.

Einfache Worthäufigkeitsliste für Einwortterme:
Für das Experiment werden Korpora in den verschiedenden Bereichen aus dem WWW mit den zwei obengenannten Methoden automatisch erstellt. Einfache Worthäufigkeitslisten für Einwortterme werden mit Hilfe der Linux-Befehle angelegt.
Ein Programm 'frequence.pl', das von mir erstellt wurde, kann ein Korpus einlesen und tokenisieren, indem Stoppwörter eliminiert werden. Das Ziel des Programmes ist es, Kandidaten für Einwortterme als Output zeilenweise an die Shell zu liefern. Beispielweise wird der Befehl "frequence.pl ./ein_Korpus | sort | uniq -c | sort -n -r | less" in der Shell ausgeführt, um einfache Frequenzlisten zu erstellen.

6.6.1 Worthäufigkeitsliste mit Varianten

Die Worthäufigkeit spielt eine wichtige Rolle für die Termgewichtung auf statistischer und lingustischer Ebene. Zur richtigen Berechnung der Worthäufigkeit sollten mindestens Lemmatisierung und orthographische Varianten eines Wortes angewandt werden. Für die Lemmatisierung im Deutschen verwendet man allgemein Wörterbuch-basierte Verfahren, um linguistisch korrekte Grundformreduktion zu erreichen. In der folgenden Wortliste aus dem Automobilbereich, die mit Hilfe der oben beschriebenen einfachen Worthäufigkeitsliste automatisch erstellt wurde, werden Frequenzen der Varianten, bestehend

aus unterschiedlichen Schreibweisen, Umlauten und Bindestrichkomposita, addiert. Dabei wird kein Wörterbuch-basiertes Verfahren für die Grundformreduktion ausgeführt.

137192	Auto/ auto/ AUTO/ AUto
91348	BMW/ bmw/ Bmw/ BMw/ B-M-W/ bMW/ BmW
45343	Golf/ golf/ GOLF/ GOlf/ gOLF
36675	Gebrauchtwagen/ gebrauchtwagen/ GEBRAUCHTWAGEN/ gebrauchtwagen/ GEBRAUCHT-WAGEN/ Gebraucht-Wagen
36253	KFZ/ kfz/ Kfz/ KfZ/ KFz/ k-f-z/ K-F-Z
32128	Mini/ MINI/ mini/ mi-ni/ MiNi/ MI-NI
32112	Citroen/ CITROEN/ citroen/ Citrön/ citrön/ CITRÖN/ CItroen

Die ersten Wörter (z.B. Auto, BMW, Golf) sind die mit den höchsten Häufigkeiten. Sie sind Kandidaten für Grundformen der Varianten, denn richtige Grundformen werden allgemein öfter als andere gebraucht.

6.6.2 Eigenschaften der Worthäufigkeitsliste

Typische Eigenschaften der Worthäufigkeitsliste für die deutsche Sprache werden im Automobilbereich experimentell betrachtet, und sie können für Einwort- und Mehrwortterme in allen Bereichen des E-Commerce gültig sein. Eine Worthäufigkeitsliste im Automobilbereich kann typischerweise aus den folgenden Wortgruppen bestehen:

a. Stoppwörter und unwichtige Wörter
(*war, dann, wurde, EUR, Artikel, Preis, Deutschland, online,...*)

b. Geografische Namen und darauf bezogene Wörter
(*Deutschland, Berlin, Berlinerin, berliner-woche, Berlina,...*)

c. **Automarken und darauf bezogene Terme**
(*BMW, BMW Service, VW, Opel, ALFA ROMEO, Jaguar,...*)

d. **Automodelle und darauf bezogene Terme**
(*VW Golf, VW Golf Plus, VW Golf FanPage, VW Passat,...*)

e. **Abkürzungen und darauf bezogene Terme**
(*KFZ, KFZ-Service, LKW, PKW, ABS, Hifi, ADAC,...*)

f. **Überprüfte EGT**
(*Auto, Wagen, Bus, Taxi, Automobil, Mini, Mobil, Diesel,...*)

g. **Aus überprüften EGT zusammengesetzte Terme**
(*Gebrauchtwagen, Autoteile, Neuwagen, Jahreswagen, Mietwagen,...*)

h. **Kandidaten für DST**
(*Fahrzeuge, Fahrzeug, Motor, Motorrad, Reifen, Tuning, Kadett...*)

Mit Hilfe der automatisch erkannten Kandidaten für DST können kontrollierte EGT automatisch und manuell erweitert werden, ebenso wie Automarken, Automodelle und Abkürzungen im jeweiligen Bereich.
Erkannte Terme, die durch die oben erwähnten Wortgruppen von 'c' bis 'h' identifiziert werden, können als domain-spezifische Terme (DST) in diesem Experiment betrachtet werden.

6.6.3 Korpus aus dem Web im Automobilbereich

Ein Korpus aus dem Web im Automobilbereich wird erstellt, wie es in 6.2.2. "Extraktion mit Suchmaschinen" beschrieben wird. Die im Experiment benutzten EGT sind 80 Terme, die von Stefan Langer im CISLEX semantisch manuell kodiert wurden. Daraus werden die 1.471 Suchbegriffe, die aus drei bis sechs Wörtern zusammengesetzt sind, für die internationalen Suchmaschinen Google und Yahoo zusätzlich mit vier Termen nämlich "kfz, pkw, lkw und Fahrzeug", automatisch erstellt. Durch Ergebnisse der Suchmaschinen werden 82.930 verschiedene URL-Adressen extrahiert, die domain-spezifisch für den Automobilbereich sein können. Davon werden 24.109 Webseiten des Content-Typs 'text/html' und einer Dateigröße von weniger als 20 Megabyte als Korpus in den Dateien gespeichert. Die folgende Anzahl der verschiedenen Wörter ist für Einwortterme in der Worthäufigkeitsliste.

Einfache Worthäufigkeitsliste mit Varianten	1.945.764 Wörter
Worthäufigkeitsliste ohne Varianten	1.559.567 Wörter

Etwa 400.000 Wörter in der Liste sind orthographische Varianten.
Linguistische Lemmatisierung wurde dabei nicht ausgeführt. Das bedeutet, dass grammatikalische Grundformreduktion und Varianten zur besseren Kalkulation der Worthäufigkeiten durchgeführt werden sollten.

6.6.4 Vergleich der Korpora als "Background Filter"

Um Stoppwörter und unwichtige Wörter zu entfernen, werden die folgenden drei Korpora, nämlich Schmuck, Wein und Kleidung, mit der bereits vorgestellten Methode "Extraktion aus Startseiten" erstellt und verglichen.
In den Worthäufigkeitslisten werden die Wörter, die durch die 1220 Stoppwörter als Stoppwort identifiziert werden, zuerst eliminiert. Die folgenden drei Startseiten für die jeweiligen Korpora wurden für das Experiment verwendet:

- Schmuck - http://www.schmuck.de (Stand:18.10.2006) -
- Wein - http://www.germanwine.de (Stand:18.10.2006) -
- Kleidung - http://www.dooyoo.de/kleidung (Stand:18.10.2006) -

Domain	W (Worthäufigkeitsliste)	Anzahl der Wörter
Schmuck	Einfache W mit Varianten	16.509
	W ohne Varianten	16.292
Wein	Einfache W mit Varianten	18.176
	W ohne Varianten	16.757
Kleidung	Einfache W mit Varianten	23.794
	W ohne Varianten	21.188

Jede Worthäufigkeitsliste kann im jeweiligen Bereich übliche Stoppwörter, unwichtige und wichtige Wörter beinhalten. Zur Eliminierung der unwichtigen Wörter wird die folgende Annahme nach dem Kapitel 3.3. "Allgemeines Korpus als "Background Filter" verwendet:

> Die Überlappung zwischen einem Korpus (z.B. Wein) und einem Target-Korpus (z.B. Automobilbranche) wird als domainneutral identifiziert und eliminiert. Dieses Verfahren wird "Korpora als Background Filter" genannt.

Häufige Wörter, die nur in einem Bereich vorgekommen sind, können domainspezifisch sein. Aber sie können zahlreiche falsche Wörter, z.B. Tippfehler, enthalten. Darauf sollte man achten.
Unwichtige Wörter können in mehreren Bereichen domain-neutral häufig vorkommen. Manche Automarken (z.B. Volkswagen), Automodelle (z.B. Golf,

Kadett) und Abkürzungen (z.B. PS, ABS) können in mehreren Bereichen auftauchen, weil sie mehrdeutig sind. Deshalb werden sie in den domainspezifischen Listen für den jeweiligen Bereich gesammelt, um domain-spezifisch behandelt werden zu können.

Die gesamten von mir erstellten Perl-Programme werden schrittweise durchgeführt, um domain-spezifische Terme zu erkennen:

a. Aufbau der Korpora

b. Einlesen eines Korpus

c. Tokenisierung, dann Eliminierung der Stoppwörter

d. Erstellung der einfachen Frequenzliste mit Varianten für Einwortterme

e. Erstellung der Frequenzliste ohne Varianten für Einwortterme

f. Semantische Annotation der Frequenzliste für Einwortterme ohne Varianten

g. Erweiterung der EGT und der domain-spezifischen Wortlisten

Die beiden Schritte (f) und (g) können rekursiv ausgeführt werden. Die Qualität der EGT und der domain-spezifischen Wortlisten spielt eine wichtige Rolle zur Erkennung der DST.

Der **Schritt (f)** wird in drei weiteren Schritten überprüft:

1. Überprüfung der domain-spezifischen Wortlisten darauf, ob ein Term vorhanden ist.

2. Matching mit Hilfe der Affixanwendung von EGT.

3. Vergleich mit verschiedenen "Korpora als Background Filter", um Kanditaten für DST zu erkennen.

6.6.5 Semantische Analyse der Einwortterme im Automobilbereich

Das Ziel der semantischen Analyse von Einworttermen ist die Überprüfung, ob ein Term in einem Bereich domainspezifisch ist. Damit können Dokumente zu den jeweiligen Themenbereichen effizient klassifiziert werden.

Die beiden schon beschriebenen Wortgruppen "Stoppwörter und unwichtige Wörter" und "Geografische Namen und darauf bezogene Wörter" gehören nicht zu den domainspezifischen Termen (DST). Deshalb müssen sie zur Extraktion der DST identifiziert und eliminiert werden.

Stoppwörter und unwichtige Wörter
Die Stoppwortliste, die 1220 Stoppwörter für Deutsch und Englisch beinhalten, wird zuerst verwendet, um Stoppwörter zu entfernen. Trotzdem bleiben übliche Stoppwörter und unwichtige Wörter übrig. Ein Teil der unwichtigen Wörter ist die Wortgruppe "Geografische Namen und darauf bezogene Wörter". Sie sind domainneutral und somit keine DST. Um sie zu eliminieren, wird eine Liste, in der 13.987 verschiedene Stadtnamen (z.B. Berlin, München) in Deutschland eingetragen wurden, gebraucht.
Indem verschiedene Worthäufigkeitslisten aus den jeweiligen Bereichen verglichen werden, können fast alle Stoppwörter und unwichtige Wörter identifiziert werden.

Automarken und darauf bezogene Terme
Mit dem Vergleich der unterschiedlichen Korpora können domainspezifische Terme im jeweiligen Bereich (z.B. Autobranche) erkannt werden. Automarken sind stark domain-spezifisch. Sie können deshalb speziell behandelt werden. Manche Automarken (z.B. Jaguar, Volkswagen) sind mehrdeutig, aber relativ deutlich dem Automobilbereich zuzuordnen. Der bekannte Autohersteller 'Jaguar' ist im Automobilbereich kein Tiername, sondern sicher eine Automarke. Aus Webseiten werden 653 international anerkannte Automarken extrahiert und in einer Liste gesammelt, um einen Term als Automarke annotieren zu können. Abgekürzte und offizielle Formen, z.B. VW (Volkswagen) sind in der Liste enthalten.

Automodelle und darauf bezogene Terme

Internationale Automarken haben auf dem Markt ständig zahlreiche neue Automodelle herausgegeben. International bekannte Automodelle, die auf deuschen Webseiten öfter gebraucht werden, können in einer Liste wie Automarken gesammelt werden. Sie sind auch domainspezifisch. Die bekanntesten Automodelle der deutschen Autohersteller "VW (Volkswagen)" sind beispielweise Golf, Golf Plus, Golf III, Passat,etc. Mit der Liste für Automodelle können solche Terme als Automodell im Automobilbereich identifiziert werden. Automodelle können auf deutschen Webseiten sehr oft mit Automarken verbunden sein (z.B. VW Golf, Opel Arena). Mit Hilfe der beiden Listen für Automarken und Automodelle können solche Phrasen semantisch analysiert werden und als Automarke oder Automodell annotiert werden.

Abkürzungen und darauf bezogene Terme

Abkürzungen sind sehr mehrdeutig. Wichtige Abkürzungen im Automobilbereich sind domainspezifisch. Aus Webseiten werden 1.692 domainspezifische Abkürzungen extrahiert und in einer Liste gesammelt.
Abkürzungen (z.B. KFZ, LKW, PKW), die zur Kompositabildung (z.B. Kfz-Service, Kfz-Werkstatt, Kfzwerkstatt, PKW-Reifen, Gebrauchtpkw) fähig sind, gehören zu 'EGT' für die Affixanwendung.

Überprüfte EGT und darauf bezogene Terme

Überprüfte EGT sind ein Kernteil der domainspezifischen Phrasen.
Domainspezifische Phrasen nach der Definition der DST beinhalten mindestens ihre domainspezifischen Teile.
Für die folgende semantische Annotation wurden die semantisch schon annotierten 80 Lexeme aus CISLEX als überprüfte EGT im Automobilbereich ausgewählt. Aus EGT zusammengesetzte Terme sind domain-spezifisch.
Die Erweiterung der EGT ist notwendig zur Extraktion der DST. Die Qualität der EGT ist sehr wichtig.
Falsche EGT (z.B. Service, Dienst, Verkauf, Ankauf, Händler, Werbung) bringen viele falsche Treffer beim Gebrauch der Affixanwendung.

Kandidaten für DST

Wenn ein Term in den domainspezifischen Listen z.B. für Automarken, Automodelle und Abkürzungen im jeweiligen Bereich nicht vorhanden ist, durch die Affixanwendung von EGT nicht erkannt wird und nach dem Vergleich der verschiedenen Korpora nicht als unwichtiger Term erkannt wird, ist er ein Kandidat für DST.
Aus Kandidaten für DST können EGT im jeweiligen Bereich erweitert werden und nötige domainspezifische Listen zur Extraktion der DST erstellt werden. Die Erweiterung der EGT ist z.B. Fahrzeug, Kfz, PKW, LKW, Motor, Motorrad, Reifen. Die Erweiterung der domainspezifischen Listen ist z.B. Golf, Kadett, Corsa, Polo (als Automodell).

Semantische Annotation der Einwortterme

DST können durch die Affixanwendung der aus CISLEX ausgewählten 80 EGT mit Hilfe der oben genannten domainspezifischen Wortlisten und dem Vergleich der unterschiedlichen Korpora für die Eliminierung der weiteren Stoppwörter und unwichtigen Wörter erkannt und erweitert werden.

Automobilbereich (Korpus)	
Einfache Worthäufigkeitsliste mit Varianten	1.945.764 Wörter
Worthäufigkeitsliste ohne Varianten	1.559.567 Wörter

Diese "Worthäufigkeitslisten ohne Varianten" für Einwortterme im Automobilbereich werden experimentell semantisch annotiert. Ohne manuelle Manipulationen wird die Liste automatisch erstellt und nach Häufigkeit sortiert. Die weiteren Terme sind im "Anhang A" eingefügt. Die ersten Top-30 Terme werden nachfolgend aufgeführt. Die ersten Wörter (z.B. EUR, Auto, GmbH) mit den höchsten Vorkommen sind Kandidaten für Grundformen der jeweiligen Varianten:

Tabelle 6.2: **semantische Annotation im Automobilbereich**

Frequenz	**Varianten [semantische Annotation]**
368318	EUR/ Eur/ eur/ EuR [NO]
137192	Auto/ auto/ AUTO/ AUto [Auto;EDST]
131216	GmbH/ GMBH/ gmbh/ Gmbh/ gmbH/ GmBH/ GMbH [NO]

Tabelle 6.2: **semantische Annotation im Automobilbereich**

Frequenz	Varianten [semantische Annotation]
112108	war/ War/ WAR/ wAr/ WAr [NO]
98822	Audi/ AUDI/ audi/ AUdi/ AuDi [Audi;Automarke]
91348	BMW/ bmw/ Bmw/ BMw/ B-M-W/ bMW/ BmW [BMW;Automarke]
91158	dann/ Dann/ DANN/ DAnn [NO]
79255	www/ Www/ WWW/ wWW/ WwW [NO]
75382	dass/ Dass/ DASS/ daSS [NO]
71071	eBay/ ebay/ Ebay/ EBAY/ e-bay/ E-Bay/ EBay/ E-bay/ ebaY/ e-Bay/ E-BAY/ eBAY [NO]
67945	Renault/ RENAULT/ renault/ REnault [Renault;Automarke]
60709	Artikel/ artikel/ ARTIKEL [NO]
60040	Ford/ FORD/ ford/ FOrd [Ford;Automarke]
56591	Preis/ preis/ PREIS [NO]
55456	Fiat/ fiat/ FIAT [Fiat;Automarke]
53758	Alfa/ ALFA/ alfa/ ALfa [Alfa;Automarke]
53700	Opel/ OPEL/ opel/ OPel/ O-P-E-L [Opel;Automarke]
52495	Deutschland/ deutschland/ DEUTSCHLAND/ Deutsch-land/ DEutschland [NO]
51886	Uhr/ uhr/ UHR/ UHr [NO]
51764	finden/ Finden/ FINDEN [NO]
49113	wurde/ Wurde/ WURDE/ wur-de [NO]
46459	Mercedes/ mercedes/ MERCEDES [Mercedes;Automarke]
45343	Golf/ golf/ GOLF/ GOlf/ gOLF [vw;Automodell]
44947	Euro/ EURO/ euro/ EURo/ EUro/ EU-RO [NO]
43915	OLDTIMER/ Oldtimer/ oldtimer/ OLDTiMER/ OldTimer/ old-timer/ Old-Timer [OLDTIMER;EDST]
42835	Suche/ suche/ SUCHE/ SUche [NO]
42576	Fahrzeuge/ fahrzeuge/ FAHRZEUGE/ Fahr-zeuge [ET]
39174	Motor/ motor/ MOTOR/ MOtor [ET]
38708	Autos/ autos/ AUTOS/ aut-os/ auto-s/ Auto-S/ Au-tos/ AUTOs [auto;=A]
38601	online/ Online/ ONLINE/ on-line/ ON-Line/ On-Line/ ON-LINE/ OnLine/ On-line/ ON-line [NO]

Weitere Stoppwörter und unwichtige Wörter werden durch Vergleich der unterschiedlichen Korpora identifiziert und z.B. [NO] annotiert.
Überprüfte EGT und darauf bezogene Terme werden durch die Affixanwendung von EGT erkannt und z.B. in der Form [Auto;EDST], [auto;=A] annotiert. "EDST" und "EGT" sind gleich. Automarken und darauf bezogene Terme werden mit Hilfe der Liste für Automarken erkannt und annotiert (z.B. [Audi;Automarke]). Automodelle und darauf bezogene Terme werden auch mit Hilfe der Liste für Automodelle erkannt und festgehalten (z.B. [vw;Automodell]). Nach dem domainspezifischen Listen-Lookup und der Affixanwendung von EGT können Termkandidaten, die nur in einem bestimmten Korpus vorgekommen sind, durch den Vergleich von "Korpora als Background

Filter" domainspezifisch erkannt und in der Form [ET] annotiert werden. Sie sind eine gute Basis für die Erweiterung von 'EGT' und domainspezifischen Listen (z.B. Automarken).
Wenn die oben gezeigte Liste fehlerfrei ist, ist die Qualität für diese semantische Annotation im Automobilbereich sehr gut.

6.7 Termgewichtung

Die bekannte **TFIDF-Gewichtung** wird wie folgt berechnet [Sal89]:

$$Wij = tf * idf = tfij * \log_2 \frac{N}{dfi}$$

wobei gilt:

- **Wij** ist das berechnete Gewicht des Terms i im Dokument j
- **tfij** ist die Häufigkeit des Terms i im Dokument j
- **N** ist die Gesamtzahl an Dokumenten
- **dfi** ist die Anzahl der Dokumente, die den Term i enthalten

Ein Term, der öfter in einem Dokument vorkommt, andererseits jedoch seltener in einer Dokumentenkollektion auftaucht, wird als signifikanter Term durch diese 'TFIDF-Gewichtung' statistisch gewichtet. Beispielsweise bedeutet der 'TFIDF-Wert' 0, daß ein Term in allen Dokumenten vorkommt.
Für diese Termgewichtung habe ich selbst ein Perl-Programm entwickelt. Alle heruntergeladenen Webseiten (bzw. Dateien) unter einem oder mehreren Verzeichnissen werden somit rekursiv verfolgt und verarbeitet. Extrahierte Einwortterme werden durch diese 'TFIDF-Gewichtung' bewertet.

Zipfsches Gesetz und Wortlänge
Nach dem Zipfschen Gesetz (1949), das im Kapitel 2.3.1. erwähnt wird, können sowohl sehr häufige als auch sehr seltene Wörter als unwichtige Wörter betrachtet werden. Dabei wird der umgekehrte Zusammenhang zwischen Länge und Häufigkeit eines Wortes betrachtet. Für die Termgewichtung kann das Zipfsche Gesetz beeinflußt von der Wortlänge beispielsweise in Perl wie folgt kombiniert werden, so wie ich es in diesem Experiment eingesetzt habe:

$termgewichtung = $frequenz * $rang * (length ($term) / 1000);
$ran (Rang eines Wortes in einer Frequenzliste) length ($term) (Wortlänge eines Terms)

6.8 Normalisierung der Terme

Die Normalisierung der Terme kann auf folgenden drei Ebenen betrachtet werden:

a. Eliminierung der Stoppwörter und unwichtigen Wörter

b. Grammatikalische Grundformreduktion

c. Erkennung der orthographischen Varianten eines Wortes

Für die Ebene (a) wird i.a. eine Stoppwortliste eingesetzt, für die Ebene (b) i.a. wörterbuch- oder regelbasierte Verfahren.
Im folgenden wird die dritte Ebene (c) behandelt.
Orthographische Varianten, die aus unterschiedlichen Schreibweisen, Umlauten und Bindestrichkomposita bestehen, werden bisher in diesem Experiment wie folgt erkannt:

Häufigkeit	**Varianten**
36253	KFZ/ kfz/ Kfz/ KfZ/ KFz/ k-f-z/ K-F-Z
1009	Kraftfahrzeug/ KRAFTFAHRZEUG/ kraftfahrzeug/ KRAFTfahrzeug
34158	LKW/ lkw/ Lkw/ LkW/ LKw/ lKW
538	Lastkraftwagen/ LASTKRAFTWAGEN/ lastkraftwagen
46459	Mercedes/ mercedes/ MERCEDES
16793	Benz/ BENZ/ benz
29643	Mercedes-Benz/ MERCEDES-BENZ/ mercedes-benz/ Mercedesbenz/ MERCEDESBENZ/ mercedesbenz/ Mercedes-benz/ MercedesBenz/ mercedes-Benz/ Merce-des-Benz/ Mercedes-BEnz/ Mer-cedes-Benz

Die jeweils ersten Wörter (z.B. KFZ, Kraftfahrzeug, Mercedes-Benz) mit dem höchsten Vorkommen sind Kandidaten für Grundformen der jeweiligen Variationen. Die Kurzform “MB” steht im Automobilbereich für “Mercedes-Benz” (z.B. MB-Truck, MB Autos). In der Worthäufigkeitsliste für Einwortterme werden in dieser Arbeit wegen der Ambiguität die Wörter aus einem oder zwei Buchstaben als Stoppwort erkannt. Aber solche Varianten wie z.B. Mercedes, Benz, Mercedes-Benz, MB referenzieren auf ein gleiches Objekt im

Automobilbereich. Deshalb können sie untereinander als Varianten behandelt und die jeweilige Worthäufigkeit zur korrekten Berechnung addiert werden. Dafür kann man leicht eine Liste für Langform und Kurzform wie folgt erstellen:

Langform	Kurzform
Kraftfahrzeug	KFZ
Lastkraftwagen	LKW
Volkswagen	VW
Alfa Romeo	Alfa
Mercedes-Benz	Mercedes
Mercedes-Benz	Benz
Mercedes-Benz	MB

Verschiedene Kurzformen für 'Mercedes-Benz' kann man in einer Zeile wie folgt erstellen:

Langform	Kurzform
Mercedes-Benz	Mercedes/ Benz/ MB

Synonyme zwischen Termen können als Varianten betrachtet werden:

Häufigkeit	Varianten
137192	Auto/ auto/ AUTO/ AUto
38708	Autos/ autos/ AUTOS/ aut-os/ auto-s/ Auto-S/ Au-tos/ AUTOs
16764	Automobile/ AUTOMOBILE/ automobile/ AUTO-MOBILe/ Auto-mobile/ Automo-bile/ AUtomobile
5960	Automobil/ automobil/ AUTOMOBIL/ Auto-Mobil/ AUTO-MOBIL/ Auto-Mobil/ auto-mobil/ Auto-mobil/ AUTOmobil
16645	Wagen/ wagen/ WAGEN
29748	car/ Car/ CAR
42576	Fahrzeuge/ fahrzeuge/ FAHRZEUGE/ Fahr-zeuge
3	motorcars

Zur korrekten Berechnung der Worthäufigkeiten sollten zusätzlich Grundformreduktionen durchgeführt werden und Synonyme (bzw. Quasi-Synonyme) erkannt und ihre Häufigkeit addiert werden.
Eine Liste für Synonyme (bzw. Quasi-Synonyme) kann dafür sehr nützlich sein.
Das Beispielwort 'Auto' ist mehrdeutig, trotzdem läßt es sich relativ deutlich

dem Automobilbereich zuordnen. Es ist das Kurzwort für 'Automobil'. Eine sehr enge Synonymbeziehung im Deutschen haben Automobil, Fahrzeug, Gefährt, Wagen, Personenkraftwagen (PKW), Car, Vehikel, Karre und Schlitten. Ebenfalls als EGT für die deutsche Sprache müssen die englischen Synonyme automobile, car, motorcar (motor car) betrachtet werden, weil sie häufig auf deutschen bzw. internationalen Webseiten vorkommen. Die oben erwähnten Synonyme von 'Auto' können in dieser Arbeit als EGT betrachtet werden.

Kapitel 7

Extraktion der Mehrwortterme in NLP

7.1 Mehrwortterm versus Kollokation

Der Begriff der **Kollokation** ist sehr mehrdeutig. In der Statistik wird die Kollokation als ein statistisch assoziiertes Wortpaar allgemein verwendet. Sie zeigt an, welche Wörter in einem Korpus öfter in Kombination vorkommen. Die zwei bekannten Assoziationsmaße "Mutual Information (MI) und Log-Likelihood" wurden in Abschnitt 3.1. erwähnt.
In verschiedenen linguistischen Phänomenen können Mehrwortausdrücke erscheinen:

- **Eigennamen und dazugehörige Umgebungen**
  New York, Frankfurt am Main, ALFA ROMEO, AutoScout24 GmbH, Ford Galaxy und VW Sharan

- **Kollokation**
  Nicht einfach zusammengesetzte Wörter (z.B. Automobilsalon in Genf), sondern häufig zusammen auftretende Wörter (z.B. Schwarzes Brett) in verschiedenen Umgebungen werden i.a. als Kollokation betrachtet. Es gibt drei Arten von Kollokationen, die aus zwei oder mehreren Wörtern zusammengesetzt einen Sinn ergeben, [WM06, S. 23-24]:

  a. Nominalphrasen (*hellichter Tag, maschinelle Übersetzung*)
  b. Verbalphrasen [1] (*Kritik üben, Abschied nehmen*)
  c. feste Wendungen (*hin und wieder, an und für sich*)

[1] bzw. Stützverbkonstruktionen und Funktionsverbgefüge

- **Selektionspräferenzen in freien Syntagmen**
  drei Diesel (Adj_N), Flugmotor BMW (N_N), Wagen mit Vierzylinder-Motor (N_Präp_N)

- **Idiome**
  Jacke wie Hose, Morgenstund Gold im Mund

Mehrwortterm

In dieser Arbeit wird der Begriff "Kollokation" wegen dieser Mehrdeutigkeit nicht verwenden. Neben den Einworttermen sollte man die bereits erwähnten Mehrwortterme (bzw. Wortgruppen) erkennen, welche mehrere Wörter zu einem Ausdruck (z.B. AutoScout24 GmbH, Audi A4 2.7 TDI) zusammenfassen. Mehrwortterme können aus Mehrwortausdrücken extrahiert werden.
Aus Nominalphrasen werden im allgemeinen signifikante Mehrwortterme extrahiert. In dieser Arbeit wurde bereits die Termextraktion aus Nominalphrasen erwähnt. Die Extraktion der Mehrwortterme ist ein weiteres wichtiges Gebiet des linguistischen Ansatzes. Die POS-Muster mit Hilfe von POS-Taggern werden häufig zur Phrasenerkennung für Mehrwortterme gebraucht. Im 'NPtool' von Arppe [Arp95, S. 5] werden sie erfolgreich für die Erkennung der englischen Nominalphrasen eingesetzt. Die Qualität der POS-Tagger spielt eine entscheidende Rolle dabei. Zahlreiche unerwartete Nominalphrasen können durch solche syntaktischen Muster (POS-Muster) gefunden werden. Deshalb benötigt man innovative linguistische Methoden der Phrasenerkennung, um falsche Treffer effizient zu vermindern.
Es sind effiziente NLP-Techniken (Natural language processing) entwickelt worden. Sie sind dazu da, signifikante Mehrwortterme zu erkennen und zu erweitern. Im Bereich von 'NLP' werden die zwei effektivsten NLP-Tools für Termextraktion, nämlich "LEXTER" und "FASTR" vorgestellt, neben "Lokale Grammatiken mit Unitex". Die zwei eigenen Methoden "Mustererkennung in Perl" und "N-Gramme mit Wortfolgen" werden in dieser Arbeit experimentell ausgeführt:

- **LEXTER**
- **FASTR**
- **Lokale Grammatiken mit Unitex**

- **POS-Muster (Part-of-Speech-Muster)**
- **Mustererkennung in Perl**
- **N-Gramme mit Wortfolgen**

"Lokale Grammatiken mit Unitex" und "N-Gramme mit Wortfolgen" wurden für "Zwei CGI-Programme im Automobilbereich" im Kapitel 5.6. verwendet und vorgestellt. Die Grundideen dieser NLP-Tools können für die Erkennung und Erweiterung der Mehrwortterme im praktischen Teil dieser Arbeit angewendet werden.

7.2 LEXTER in NLP

LEXTER (Bourigault, 1994) ist die Abkürzung von "Logiciel d'EXtraction de TERminologie". 'LEXTER' ist ein Terminologie-Extraktionssystem für die automatische Erstellung von Terminologie aus französischen Fachtexten, die durch grammatikalische Kategorien (Nomen, Verben, Adjektive, etc.) getaggt wurden. Potentielle terminologische Einheiten (terminological units) sind Nominalphrasen (z.B. nom adj, nom de nom[2]).
'LEXTER' untersucht nur diese Nominalphrasen weiter.
Die Extraktion mit 'LEXTER' wird in zwei Schritten erarbeitet, um potentielle terminologische Einheiten zu extrahieren [Bou92, S. 979]:

a. **Analyse (Splitting)**
   Kategorisierte Texte werden durch endliche Automaten in maximalen Nominalphrasen (maximal-length noun phrases through finite state machines) zerlegt. Dabei werden Verben, Pronomen, Konjunktionen und Determinator als Grenze zwischen Nominalphrasen verwendet.

   <Kategorisierte Texte>
   UN TRAITEMENT DE TEXTE EST INSTALLE SUR LE

[2]

Kategorie	Erklärung
nom	Nomen und unbekannte Wörter
adj	Adjektive, sowie Partizip Perfekt und Partizip Präsens
de	die Präposition 'de'
prep	die Präpositionen (avec, contre, dans, par, pour, sous, sur, vers)
det	Determinator (la)

DISQUE DUR DE LA STATION DE TRAVAIL
<**Analyse**>
rules of frontier marker identification (z.B. verb, prep. (except 'de' et 'a') + det.)
<**Maximale Nominalphrase**>
TRAITEMENT DE TEXTE
DISQUE DUR DE LA STATION DE TRAVAIL

b. **Parsing**
Beim Parsing (zweite Phase) werden die maximalen Nominalphrasen mit Hilfe einer Regelbasis in kleinere Teilphrasen zerlegt und nach Head(H) und Modifier(M) analysiert. Diese Nominalphrasen sind entweder schon potentielle Fachtermini (z.B. TRAITEMENT DE TEXTE) oder sie enthalten Teilphrasen, die potentielle Fachtermini sind (z.B. "DISQUE DUR DE LA STATION DE TRAVAIL" enthält "DISQUE DUR" und "STATION DE TRAVAIL". [Bla97, S. 68]

<**Parsing rules**>

noun1 adj prep det noun2 prep noun3 →	noun1 adj noun2 prep noun3
DISQUE DUR DE LA STATION DE TRAVAIL →	DISQUE DUR (H: DISQUE, M: DUR) STATION DE TRAVAIL (H: STATION, M: TRAVAIL)

<**likely terminological units**>
TRAITEMENT DE TEXTE
DISQUE DUR DE LA STATION DE TRAVAIL
DISQUE DUR
STATION DE TRAVAIL

7.3 FASTR in NLP

'FASTR' von Christian Jacquemin ist ein NLP-Tool für die automatische Extraktion der Mehrwortterme aus großen Korpora. Damit können Terme normalisiert sowie Varianten erkannt und erweitert werden. Der Formalismus von 'FASTR' ist unifikations-basiert. Das Ziel ist, morphologische, syntaktische, semantische und pragmatische Term-Variationen durch kontrollierte Terme mit Hilfe einer Metagrammatik zu erkennen.

> The basic component of the parser is a metagrammar, which is a set of metarules describing acceptable linguistic transformations of terms.
> [Jac01, S. 2]

Die Termerkennung in 'FASTR' beruht nur auf einer partiellen und lokalen Analyse der Sätze wie folgt [Jac01, S. 161-167]:

Controlled term	Variant	Metarule
Beta effect	effect between beta	Permutation
Beta effect	effect on beta	Perm $(X_1 \rightarrow X_2X_3) \equiv X_1 \rightarrow X_3P_4X_2$
Beta effect	effect of beta	
Adult animal	adult obese animals	Modification or substitution
Adult animal	adult transgenic animals	Modif $(X_1 \rightarrow X_2X_3) \equiv X_1 \rightarrow X_2X_4X_3$
Blood plasma	blood flow and plasma	Coordination
Blood plasma	blood GPX and plasma	Modif $(X_1 \rightarrow X_2X_3) \equiv X_1 \rightarrow X_2X_4C_5X_3$
Anal sphincter	anal and urethral sphincter	Modif $(X_1 \rightarrow X_2X_3) \equiv X_1 \rightarrow X_2C_4X_5X_3$

7.4 Mustererkennung in Perl

Die Programmiersprache 'Perl' steht für "Practical Extraction and Report Language". Die Stärke von 'Perl' ist die hervorragende Mustererkennung (engl. pattern matching) mit regulären Ausdrücken (engl. regular expressions). Damit kann man Informationen aus Texten leicht extrahieren und manipulieren. Zur Extraktion der Mehrwortterme kann diese Mustererkennung gebraucht werden.

7.4.1 Phrasen für Automarken und Automodelle

Phrasen, die sich auf Automarken und Automodelle beziehen, sind sehr signifikant. Sie werden als "Konkrete Produktnamen (KPN)" im Kapitel 5.6.3.4. ("Bootstrapping-Verfahren mit Automarken") in dieser Arbeit berücksichtigt. Die folgenden Kombinationen von Automarken, Automodellen und dazugehörigen Zusatzinformationen können in kommerziellen Webseiten öfter vorkommen, um ein Automodell einer Automarke darzustellen:

```
<b>FORD Mondeo 2.0 TDCi NAVI</b>
<b>VW Sharan 1.9 TDI Comfort</b>
<b>AUDI A4 2.5 TDI</b>
<b><font size="4">BMW M3 GTR (E46) </font></b>
<b><a href="...">Opel Speedster 2.2</a></b>
<b>BMW 323i, weiß </b>
<a href="...">VW Multivan T5 2.5 TDI</a>
```

Solche manuell fein verarbeiteten Phrasen für Automarken und Automodelle können bei der HTML-Analyse als Nominalphrasen besonders behandelt

und erkannt werden, weil die HTML-Tags durch das Senkrecht-Zeichen '|' im Korpus ersetzt und als Trennsysmbol neben Satzendzeichen (.,?,!) effizient gebraucht werden. Automarken stehen vor Automodellen. Wir sehen z.B. bei (BMW 323i, weiß), dass die Kombination von Automarke und Automodell mit Hilfe eines Kommas von weiterer Information (z.B. weiß) getrennt und als eine Einheit behandelt wird. Diese Nominalphrasen (NP) für Automarken und Automodelle sind wie folgt definiert:

a. NP beginnen mit einer Automarke.

b. Automodelle und die dazugehörigen Zusatzinformationen haben eine angemessene Länge. (z.B. nicht mehr als 30 Zeichen)

Die beiden Schritte werden nacheinander überprüft. Die zwei folgenden Test-Korpora aus dem WWW werden dafür verwendet. Methode A steht für "Extraktion aus Startseiten", Methode B für "Extraktion mit Suchmaschinen":

Anzahl d. Webseiten	Korpus-Größe	gefundene Phrasen	Erstellungsmethode
2.930	15.287.372 Bytes	973	Methode A
24.109	459.140.633 Bytes	59.784	Methode B

Die 60.040 verschiedenen Phrasen, die aus den beiden Korpora erkannt wurden, sind von sehr guter Qualität. Damit kann man ein interessantes Abfragesystem für Automarken und Automodelle sowie deren Häufigkeiten leicht erstellen. Welche Automarken und Automodelle im Web kommen häufig vor? Welche Automodelle gibt es von VW, BMW oder Opel? Welche Automodelle von 'VW' sind populär im WWW?
Ebenfalls kann es nützlich sein, z.B. einen neuen Namen für ein neues Automodell zu konzipieren. Beispielsweise kann man mit Hilfe der Linux-Befehle 'grep' und 'head' die Liste, welche die 60.040 verschiedenen Phrasen enthält, wie folgt abfragen:

```
Lists_auto> grep -i "\[VW\]" automarkeModelle_both_2.list | head -10
1015     [VW] Bora Achsschenkel hi
1014     [VW] Passat Motor Bj
773      [VW] Passat
727      [VW] Polo
679      [VW] Golf
412      [VW] Käfer
```

```
404      [VW] Bus
381      [VW] Golf III
358      [VW] Bora
342      [VW] Golf IV

Lists_auto> grep -i "\[vw\] golf" automarkeModelle_both.list | head -10
679      [VW] Golf
381      [VW] Golf III
342      [VW] Golf IV
236      [VW] Golf II
124      [vw] golf
115      [VW] Golf 3
107      [VW] Golf V
98       [VW] Golf Diesel
97       [VW] Golf 4
95       [VW] Golf I

Lists_auto> grep -i "\[BMW\]" automarkeModelle_both_2.list | head -10
1190     [BMW] Z8
1090     [BMW] 6
1018     [BMW] 520i VBj
422      [BMW] 3er
352      [BMW] E30
298      [BMW] X5
230      [BMW] 3er-Reihe
216      [BMW] Cabrio
215      [BMW] 3
209      [BMW] Z3
```

In eckigen Klammern, z.B. [Fiat], [Maserati] werden Automarken formuliert. Zwischen "VW" und "BMW" ist der Unterschied der Automodelle leicht zu erkennen. Es gibt einige falsche Automodelle, z.B. [VW] Transporter, [BMW] Motorrad, [BMW] Gebrauchtwagen. Solche falschen Automodelle können automatisch erkannt und eliminiert werden, weil sie für alle Automarken gleich oft vorkommen. Einige Automodelle, z.B. "[Opel] Corsa" können von mehreren Firmen (z.B. [Maserati] MC12 Corsa, [Vauxhall] Corsa, [Fiat] 501 SS Corsa) sein. Aber Autohersteller versuchen natürlich, verschiedene Automodelle für ihre Firmen zu erfinden.
Für die genannte semantische Annotation werden die häufigsten 338 Automodelle, die als solche semantisch gekennzeichnet werden können, nach dem folgenden Kriterium aus diesen schon erwähnten 60.040 Phrasen für Automarken und Automodelle automatisch extrahiert:

> Automodelle, die nur für einen Autohersteller oder selten für zwei Autohersteller gültig sind, werden nach den häufigeren Vorkommnissen ausgewählt.

Z.B. golf [vw], bora [vw/ maserati], boxer [peugeot/ bmw], campo [isuzu/ opel], clk [mercedes/ brabus], crv [honda/ toyota], dino [fiat/ ferrari], fiesta [ford/ fiat], forfour [smart/ brabus], fox [vw/ nsu], kadett

[opel/ chevrolet], kalos [chevrolet/ daewoo], lupo [vw/ seat], panda [fiat/ seat], stratus [dodge/ chrysler], sunny [nissan/ datsun], vento [vw/ mercedes], viper [dodge/ chrysler], wrangler [jeep/ chrysler], abarth [fiat], accord [honda], actros [mercedesbenz], agila [opel], aklasse [mercedes]

Kapitel 8

Erkennung der Produktterme (PT) für E-Commerce

8.1 E-Commerce

E-Commerce steht für "Electronic Commerce". Zahlreiche Unternehmen investieren zunehmend in elektronischen Marktplätzen im Word Wide Web, um mehr Kunden zu gewinnen und so den Umsatz wesentlich zu erhöhen. Die zwei Akteure bzw. Rollen, nämlich Käufer[1] und Verkäufer[2] sind hierbei maßgeblich am Internet-Handel beteiligt. Dem elektronischen Handel können im allgemeinen die sechs Typen B2B[3], B2C[4], B2A[5], A2C[6], A2A[7], C2C[8] zugeordnet werden.
Beim Internet-Handel kommen überwiegend die beiden Typen B2B und B2C zur Anwendung. Die Entwicklung der elektronischen Handelstransaktionen ist in diesem Bereich besonders interessant.
Für den Erfolg einer E-Commerce-Plattform spielen Produktsuche und Kategorisierung eine wichtige Rolle. Für die Kommunikation zwischen Käufern und Verkäufern muss man domainspezifische Terme im jeweiligen E-Commerce-Bereich erfassen. Damit können sich Käufer u. Verkäufer gegenseitig informie-

[1] Käufer (bzw. Kunde oder Konsument)
[2] Verkäufer (bzw. Händler oder engl. Merchant)
[3] Business-to-Business: Handel zwischen Unternehmen
[4] Business-to-Consumer: Handel zwischen Unternehmen und Privatkunden
[5] Business-to-Administration: Handel zwischen Unternehmen und Staat (bzw. öffentlicher Verwaltung)
[6] Administration-to-Consumer (A2C): Handel zwischen Staat und Privatkunden
[7] Administration-to-Administration: Handel zwischen Staaten
[8] Consumer-to-Consumer: Handel zwischen Privatkunden

ren bzw. abfragen. Z.B. können Produkte (z.B. Wagen, Computer), Dienstleistungen (z.B. Autoteileverkauf, Auto-Truck-Service), semantische Beziehungen zwischen Termen (z.B. USB-Stick, USB Stick, Stick Usb, USB-memory, Speichersticks) und Preise effizient ermittelt werden.

8.2 Quellen der domainspezifischen Terme im E-Commerce-Bereich

Domainspezifische Terme (DST) im E-Commerce-Bereich können allgemein aus den folgenden 4 Quellen extrahiert werden:

- **Webseiten**: z.B. Title, Metakeywords, Ankertext, Inhalt der Webeite
- **Query Logs**
- **Produktklassifikationssysteme**: z.B. wand.com, gelbeseiten.de
- **Offline-Kataloge**: z.B. Yahoo, Open Directory Project (dmoz.de)

8.3 Eigenschaften der domainspezifischen Terme im E-Commerce-Bereich

Terme, die auf Produkte bezogen sind, werden in dieser Arbeit Produktterme (PT) genannt. Die Erkennung der Produktterme bzw. Produktnamen und Marken für E-Commerce ist die Hauptzielsetzung in dieser Arbeit.
DST im E-Commerce-Bereich können auf die folgenden 4 Ebenen bezogen werden:

- **Produkte**: z.B. Funkautos, Leichtauto, Renault-Autos, Ein-Liter-Auto, 3-Liter-Auto, ADAC Autositze
- **Dienstleistungen**: z.B. Autohändler, Autovermietung, Autoteilverkauf, Auto-Truck-Service, ADAC Autoverleih
- **Marken**: z.B. Mercedes, BMW, Audi, VW, Porsche, Ford, Opel, Nissan, Saab, Smart, Rolls-Royce, Land Rover, Jaguar

- **Branchenbezeichner**: Die meisten Branchenbezeichner sind abstrakt wie folgt:

```
konsumgüter#lebensmittel lebensmittelmarkt,lebensmittel,feinkost
dienstleistung#reinigen gebäudereinigung
konsumgüter#cafe kaffeehaus,cafes,cafes
gesundheit#allgemeinarzt allgemeinmediziner
finanz#inkasso inkassobüros
```

  Ausgenommen hiervon sind weniger abstrakte Branchenbezeichner (z.B. Auto, Computer).

8.4 Eigennamen

Gattungsnamen (Appellativa) verweisen auf eine Klasse, die gleichartige Objekte (z.B. Tier, Haus) beinhaltet. Im Gegensatz dazu können Eigennamen (Propria) eine Referenz auf einzelne Objekte sein. In den Zweifelsfällen ist es schwierig, sie zu unterscheiden. Die bekannten Produktnamen 'Tempo' und 'tesa' können im Text z.B. "Bitte ein Tempo!" anstelle der Gattungsnamen 'Papiertaschentuch' und 'Klebeband' einfach verwendet werden. Für die Einordnung der Eigennamen gibt es folgende bekannte unterschiedliche Klassen:

- **Eigennamen und Gattungsnamen sind die direkten Unterklassen der Substantive.**
  Konkreta und Abstrakta sind die direkten Unterklassen der Gattungsnamen (nach Helbig/Buscha[9] 2001, S. 206).

- **Eigennamen und Gattungsnamen sind die direkten Unterklassen der Konkreta**.
  Konkreta und Abstrakta sind die direkten Unterklassen der Substantive. Siehe hierzu Tabelle **??** aus der traditonellen Grammatik nach WIMMER 1973 und Duden 4 (1984) [Koß96, S. 35]:

Bei der Eigennamenerkennung (engl. Named Entity Recognition) können die folgenden Kategorien erkannt werden:

[9] Helbig, G. / Buscha, J: Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. Leipzig 2001, Langenscheidt

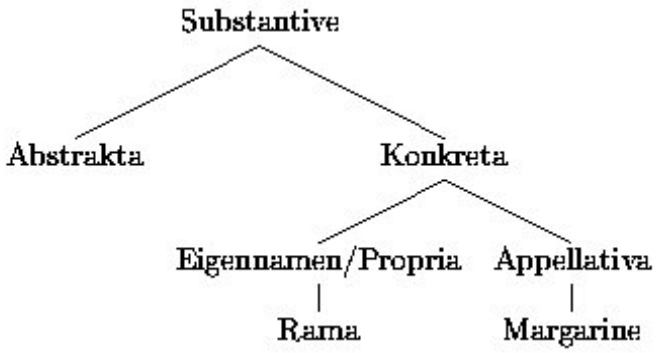


Abbildung 8.1: Klasse der Substantive (nach WIMMER 1973 u.a.)

- **Organisationen - Firmennamen**
- **Produkte**
- Personen
- Orte
- Zeitangaben
- Datumsangaben
- Währungsangaben

Firmennamen sind im Allgemeinen eine Unterklasse der Organisationsnamen. Firmennamen und PT sind domainspezifisch im E-Commerce-Bereich. Sie können durch EGT und domainspezifische Listen (z.B. Firmennamen) in den jeweiligen E-Commerce-Bereichen erkannt werden (z.B. AEG FOEN 1600 Pianissimo, BMW AG). Bekannte Firmennamen, die für die Kompositabildung (z.B. BMW-Hersteller, Audi-Kunden) möglich sind, können wie EGT behandelt werden. Firmennamen sind abstrakt, sie können Produkttermen zugeordnet werden.

8.5 Produktterme (PT)

Eigennamen bzw. Produktnamen sind nach der Klassifikation der Substantive nach WIMMER 1973 und Duden 4 (1984) “etwas Konkretes”.

Konkrete Produktnamen (z.B. Rama, AEG FOEN 1600 Pianissimo) und produktbezogene Terme (z.B. Margarine, Haartrockner) im E-Commerce können anhand von EGT erkannt werden. Wegen der allgemeinen Unterscheidung zwischen Eigennamen und Appellativa werden PT in dieser Arbeit bezeichnet, weil es sich um die Erkennung der PT in dieser Arbeit handelt.
Das Wort 'Rama' alleine hat keine Bedeutung und kann vielseitig (z.B. Casino Rama, Lord Rama) gebraucht werden. Im folgenden Werbetext im Internet kann 'Rama' als Margarine und Produktname durch den Ausdruck "Rama Margarine" bzw. "Margarine Rama" erkannt werden.

> Rama Margarine, 500g
> Rama - gut, gesund und mit dem bewährt guten Geschmack - nun erhältlich im preisgünstigen 500g Behälter.

Mit Hilfe der produktbezogenen Terme bzw. EGT (z.B. Margarine) können solche Ausdrücke (z.B. Rama Margarine) im Text maschinell dahingegend identifiziert werden, ob ein Wort (z.B. Rama) als Produktname, Dienstleistung oder Branchenbezeichner verwendet wird.
Produktnamen und produktbezogene Terme sind der Hauptbestandteil der domainspezifischen Terme in den jeweiligen E-Commerce-Bereichen. In den kommerziellen Webseiten, Produktklassifikationssystemen und "Offline Katalogen" wird ermittelt, wer was liefert. Zur Zeit gibt es tatsächlich verschiedene Lieferantensuchmaschinen, z.B. "Wer liefert Was?"[10] für deutsche Firmen, "Local.com"[11] für amerikanische und englische Firmen, um dazu effiziente Kommunikation zwischen Unternehmen und Kunden im E-Commerce zu ermöglichen. Es handelt sich um die Erkennung der Produktterme. Die produktbezogenen Terme bzw. EGT werden in dieser Arbeit der Unterklasse der "Generischen Produktterme" zugeordnet. Die Produktterme werden in zwei Unterklassen aufgeteilt:

- Generische Produktterme (GPT): z.B. Wein, Rotwein, Handschuhe
- Konkrete Produktnamen (KPN): z.B. BMW 645 Ci

[10] Wer liefert Was? : http://www.wlw.de [27.03.2007]
[11] Local.com: http://www.local.com [27.03.2007]

Einfache und komplexe GPT (z.B. Wein, französischer Wein, Handschuhe, Lederhandschuhe, Auto, Luxus-Auto) beinhalten mindestens einen EGT.
Die GPT können durch die Affixanwendung von EGT automatisch erkannt werden. Man kann domainspezifische Lexika (z.B. Firmennamen, Abkürzungen und Modellnamen) in den jeweiligen Bereichen gebrauchen, um konkrete Produktnamen (z.B. BMW 645 Ci, Mercedes-Benz 280 SL) zu identifizieren.

8.5.1 Struktur der Prudukttерme

Die Elementaren Generischen Terme (EGT) sind eine signifikante Basis für die Erkennung der Produktterme (PT). Generische Produktterme (GPT) werden in zwei Unterklassen, nämlich EGT und Komplexe Generische Terme (KGT) wie folgt eingeteilt:

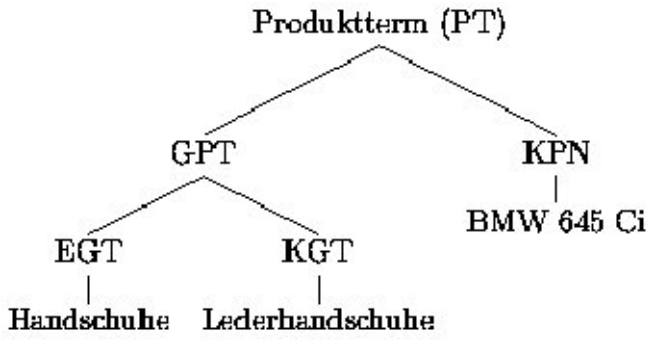


Abbildung 8.2: Struktur der Produktterme

8.5.2 Konkrete Produktnamen (KPN)

Während ein GPT auf mehrere Dinge referenziert, bezieht sich ein KPN auf ein Ding. Konkrete Produktnamen (KPN) sind nicht generisch und auf ein Produkt bezogen (z.B. BMW 645 Ci Cabrio, Mercedes Benz 280 SL Pagode Hardtop). Heutzutage nimmt die Anzahl der KPN von Firmen sehr schnell zu. Alle Firmen wollen ihre Produkte mit Namen, die sehr kommerziell, attraktiv und verkaufsfördernd sind, beim Kunden bekanntmachen und verkaufen. Dafür gibt es schnell wachsende Webseiten für E-Commerce. KPN spielen eine große Rolle dabei.

Suchmaschinen und Webkataloge für E-Commerce (z.B. wand.com, gelbseiten.de, goyellow.de) werden im Internet erstellt, um elektronische Kommunikationsbeziehungen (z.B. B2B, B2C, C2C, C2B) effizient und qualitativ zu ermitteln. Die KPN können dazu beitragen, dass ein Produkt eindeutig identifiziert werden kann.
Ein KPN als Eigenname referenziert auf ein einzelnes Produkt. Wenn man ins Suchfeld "Auto" eintippt, ist das mehrdeutig. Aber wenn der Name "BMW 645 Ci Cabrio" ins Suchfeld eingetippt wird, ist das eindeutig. BMW (Bayerische Motoren Werke AG) ist ein bedeutender deutscher Hersteller von Autos, Motorrädern und Motoren in Bayern. Der KPN "BMW 645 Ci Cabrio" wird auf einen speziellen Artikel (hier "Auto") referenziert. Man benötigt domainspezifische Listen (bzw. Lexika) für die Erkennung von KPN (z.B. Listen für Automarken, Automodelle, Abkürzungen etc.). Die KPN in der Autobranche wurden in Kapitel 5.6.3.4. "Bootstrapping-Verfahren mit Automarken" vorgestellt. Sie können **Marken**, **Modelle** und dazugehörige **Zusatzinformationen** beinhalten (z.B. BMW 645 Ci, BMW 645 Ci Cabrio, BMW 645 Ci Cabrio Schwarz Welly, BMW 645 Ci Cabrio silber, BMW 645 Ci Cabriolet blau Maisto, BMW 645 Ci Coupe Safetycar Moto GP, BMW 645 Ci silber).

8.5.3 Erkennung der Produktterme

Im Bereich der Eigennamenerkennung werden hauptsächlich Personen-, Firmen- und Produktnamen erkannt. Für die Erkennung der KPN kann man konventionelle Methoden für Eigennamenerkennung (z.B. Mustermatching, Namenslisten) anwenden. Im vorherigen Kapitel 5.6.3.4. ("Lokale Grammatiken mit Unitex") wurden die Kombinationen mit Automarken und Automodellen (z.B. BMW 645 Ci) in einem Text erkannt.
Die folgende Grammatik wird verwendet, um die Generischen Produktterme (GPT) zu erkennen. GEO steht für geografische Namen:

- x ist ein GPT, wenn:
  i. x ∈ EGT (z.B. Wein, Handschuhe)
  ii. x ∈ KGT (z.B. Weinaromen, Lederhandschuhe)
- x ist ein möglicher GPT, wenn:

a. **Suffixanwendung**: x hat die Form "AB" und B ist ein EGT.
Z.B. Lederjacke (Jacke), französicher Wein (<GEO> EGT)

b. **Präfixanwendung**: x hat die Form "AB" und A ist ein EGT.
Z.B. Autohersteller (Auto), BMW München (EGT <GEO>)

c. **Infixanwendung**: x hat die Form "ABC" und B ist ein EGT.
Z.B. Altautoannahme (Auto), Unfallautoverkauf (Auto)

- In EGT gibt es eine Regel zum Generieren: x > y (x generiert y.)
Z.B. Wagen > Gebrauchtwagen, Wein > Rotwein, Wein > Weißwein

Solche Terme, die nicht weiter zerlegt werden können (z.B. Wagen, Handschuhe, Autohaus), werden als EGT in dieser Arbeit betrachtet. Andere Terme, wie z.B. Rotwein, Weißwein, werden als KGT in dieser Arbeit betrachtet, weil sie als Hyponyme von Wein berücksichtigt werden können und zum Bereich "Wein" gehören. Dies ist eine Konvention. Die folgende hierarchische Struktur zwischen Wein und Rotwein wird vorgeschlagen:

Wein
|
Rotwein
|
Französischer Wein
|
Französischer Rotwein

Tabelle 8.1: Struktur von "Wein" und "Rotwein"

8.6 Nicht-domainspezifische Terme

Die folgenden domainneutralen Terme werden in dieser Arbeit als "nicht domainspezifisch" betrachtet, weil sie nicht auf den vier Ebenen - Produktnamen, Dienstleistungen, Marken und Branchenbezeichner - bezogen sind:

- **Berufsbezeichner**: z.B. Projektleiter/in, Sales Assistent, Vorstandsassistent/in, Sekretärin der Geschäftsleitung, Vertriebsmitarbeiter - Automotive, Niederlassungsleiter/in, Betriebsleiter/in

- **Abteilungsnamen von Firmen**: z.B. Entwicklungsabteilung, Abteilung Business Cooperations, Abteilung Customer Care, Abteilung Gebäudemanagement
- **Geographische Namen**: z.B München, Deutschland

8.7 Semantische Merkmale von Produkttermen

Die folgenden semantischen Merkmale von PT können verwendet werden:

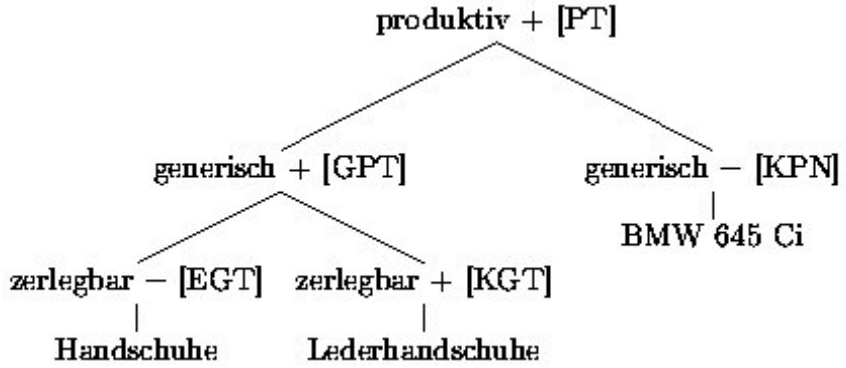


Abbildung 8.3: Semantische Merkmale von Produkttermen

8.8 Erkennung der Produktterme in der Autobranche

Im Kapitel 6 ("Domainspezifische Korpora aus dem Web") wurden die genannten zwei Methoden schon vorgestellt und von mir selbst programmiert, um domainspezifische Korpora aufzubauen. DST der Autobranche werden aus dem automatisch erstellten Korpus als Test erkannt, weil "Autobranche" zu den E-Commerce-Bereichen gehört.
Die Art der Extraktion von DST in der Autobranche in Kapitel 6. wird über wichtige Verfahren zur Erkennung von Einworttermen übersichtlich zusammengefasst. Die Methode für die Erweiterung der neuen EGT wird vorgestellt:

i. **Ein Korpus (z.B. Automobilbereich) aufbauen**:

   a. Extraktion aus Startseiten (z.B. www.autoscout24.de, www.schmuck.de)

b. Extraktion mit Suchmaschinen (z.B. Google, Yahoo)

ii. **"Frequenzliste 1 mit Varianten" erstellen**:
Die Stoppwortliste, die 1220 Einträge enthält, wird dabei zuerst verwendet, um Stoppwörter zu eliminieren.

99467 GmbH
99323 Auto
84249 Audi
82312 BMW

iii. **"Frequenzliste 2 ohne Varianten" aus Frequenzliste 1 erstellen**:
Die Lemmatisierung wird im Test nicht durchgeführt. Nur einfache orthographische Varianten eines Wortes werden betrachtet. Die getroffenen Frequenzen werden entsprechend addiert:

137192 Auto/ auto/ AUTO/ AUto
131216 GmbH/ GMBH/ gmbh/ Gmbh/ gmbH/ GmBH/ GMbH
98822 Audi/ AUDI/ audi/ AUdi/ AuDi
91348 BMW/ bmw/ Bmw/ BMw/ B-M-W/ bMW/ BmW

vi. **"Frequenzliste 3 mit der semantischen Annotation" aus Frequenzliste 2 erstellen**: Die Überprüfung der domainspezifischen Listen, die Affixanwendung von EGT und drei andere "Korpora als Background Filter" werden zur semantischen Annotation für die Erkennung der DST schrittweise von a nach g wie folgt verwendet:

a. 652 Automarken (z.B. BMW, Audi, VW)

b. 1692 Abkürzungen (z.B. PKW, ADAC)

c. Suffixanwendung von EGT (z.B. Gebrauchtwagen)

d. Präfixanwendung von EGT (z.B. Wagenheber)

e. Infixanwendung von EGT (z.B. Mietwagenservice)

f. 471 im Internet bekannte Automodelle (z.B. Golf, Astra)

g. Danach werden Termkandidaten mit den unterschiedlichen Korpora verglichen, um weitere Stoppwörter und domainneutrale Wörter zu beseitigen. In diesem Experiment wurden die drei Korpora - Wein, Schmuck und Kleidung als "Background Filter" - erstellt.

Bei der Annotation wurden weitere Stoppwörter und unnötige Wörter mit 'NO' markiert:

368318 EUR/ Eur/ eur/ EuR [NO]
131216 GmbH/ GMBH/ gmbh/ Gmbh/ gmbH/ GmBH/ GMbH [NO]
112108 war/ War/ WAR/ wAr/ WAr [NO]
91158 dann/ Dann/ DANN/ DAnn [NO]
79255 www/ Www/ WWW/ wWW/ WwW [NO]
75382 dass/ Dass/ DASS/ daSS [NO]
71071 eBay/ ebay/ Ebay/ EBAY/ e-bay/ E-Bay/ EBay/ E-bay/ ebaY/e-Bay/E-BAY/ [NO]
60709 Artikel/ artikel/ ARTIKEL [NO]
56591 Preis/ preis/ PREIS [NO]
52495 Deutschland/ deutschland/ DEUTSCHLAND/ Deutsch-land/ DEutschland [NO]

v. **Erweiterung von EGT**: Nach dem Vergleich der Korpora werden Termkandidaten für neue DST ausgewählt und 'ET' bei der Annotation markiert. Sie sind eine gute Basis für die Erweiterung der 'EGT'. Die Top10-Termkandidaten werden wie folgt angezeigt:

42576 Fahrzeuge/ fahrzeuge/ FAHRZEUGE/ Fahr-zeuge [ET]
39174 Motor/ motor/ MOTOR/ MOtor [ET]
34748 Motorrad/ motorrad/ MOTORRAD/ Motor-rad [ET]
29735 Ersatzteile/ ersatzteile/ ERSATZTEILE/ Ersatz-Teile/ ErsatzTeile/ ErsatZteile [ET]
25762 Fahrzeug/ fahrzeug/ FAHRZEUG/ FAhrzeug/ Fahr-zeug [ET]
24455 inserieren/ Inserieren/ INSERIEREN [ET]
23539 AMG/ amg/ Amg/ am-g [ET]
21914 Coupe/ coupe/ COUPE [ET]
21289 Tuning/ tuning/ TUNING [ET]
20682 Romeo/ ROMEO/ romeo [ET]

8.8.1 Erweiterung von EGT

In diesem Experiment wurden nur die 80 EGT, die von Stefan Langer mit dem semantischen Merkmal 'FZA' für Autos, PKWs und LKWs manuell kodiert wurden, automatisch selektiert und für die Affixanwendung von EGT verwendet. Diese 80 EGT im Automobilbereich genügen nicht und müssen erweitert werden. Die Qualität der verwendeten EGT für die Affixanwendung ist absolut wichtig für die Erkennung von jeweiligen DST. EGT können automatisch erweitert und mit Fachkenntnissen manuell verbessert werden.
Die zwei folgenden Punkte werden gemäß den Eigenschaften von EGT berücksichtigt, um Termkandidaten für die Erweiterung von EGT automatisch zu erstellen. Dann können die automatisch erkannten Termkandidaten für neue

EGT manuell ausgewählt werden.

a. nicht zerlegbar und generisch (z.B. Fahrzeug, Motor)

b. geeignet als Grundwort zur Kompositabildung (z.B. Nutzfahrzeuge)

Dafür wird der reguläre Ausdruck für die Suffixanwendung beim Perl-Programm verwendet. Die Pluralbildung (z.B. Fahrzeug, Fahrzeuge, Firmenfahrzeugen) werden dabei beachtet. Die Inputdatei im Experiment enthält die als 'ET' annotierten 190.689 Wörter aus den 224.292 semantisch annotierten Termen im Beispiel-Korpus "Autobranche". Durch das Beispielwort 'Fahrzeuge', das nicht zu den 80 EGT gehört, wurden die 199 zusammengesetzten Wörter wie folgt getroffen. Das Wort 'Fahrzeuge' wird wegen der oben genannten zwei Punkte als Termkandidat für neue EGT im Automobilbereich ausgewählt:

1325 Neufahrzeuge/ neufahrzeuge/ NEUFAHRZEUGE/ Neu-Fahrzeuge [ET]
1188 Gebrauchtfahrzeuge/ GEBRAUCHTFAHRZEUGE/ gebrauchtfahrzeuge/ Gebraucht-Fahrzeuge [ET]
1101 Schlachtfahrzeuge/ schlachtfahrzeuge [ET]
990 Kraftfahrzeuge/ KRAFTFAHRZEUGE/ kraftfahrzeuge [ET]
903 Fahrzeuges/ fahrzeuges/ FAhrzeuges [ET]
665 Erdgasfahrzeuge/ erdgasfahrzeuge/ ERDGASFAHRZEUGE/ Erdgas-Fahrzeuge [ET]
502 Unfallfahrzeuge/ unfallfahrzeuge/ UNFALLFAHRZEUGE/ Unfall-Fahrzeuge [ET]
429 Unfallfahrzeugen/ UNFALLFAHRZEUGEN/ unfallfahrzeugen/ Unfall-Fahrzeugen [ET]
331 Kraftfahrzeugen/ KRAFTFAHRZEUGEN/ kraftfahrzeugen [ET]
314 Einsatzfahrzeuge/ einsatzfahrzeuge/ EINSATZFAHRZEUGE [ET]

Als Resultat wurden 2.769 Termkandidaten der neuen EGT aus den 190.689 Wörtern, die als 'ET' annotiert sind, automatisch erstellt.
Die Top17-Termkandidaten für neue EGT werden wie folgt angezeigt. In der linken Klammer steht keine Worthäufigkeit, sondern die Häufigkeit der zusammengesetzten Wörter:

```
Input [190689 ./result/inputFileedst84_1471term/freVariationAnnotation190689ET.list]
[199] Fahrzeuge;EGT
[145] Motor;EGT
[31] Motorrad;EGT
[18] Ersatzteile;EGT
[339] Fahrzeug;EGT
[53] Tuning;EGT
[51] Getriebe;EGT
[24] Computer;EGT
[154] Fahrer;EGT
[22] Polizei;EGT
[28] Unfall;EGT
[18] Finanzierung;EGT
[26] Trans;EGT
[49] Transporter;EGT
[33] Werkstatt;EGT
[22] Felgen;EGT
[181] Halter;EGT
```

Die oben automatisch erstellte Liste für die Kandidaten der neuen EGT kann bei der manuellen Auswahl für die Erweiterung der EGT sehr nützlich sein (z.B. Fahrzeug, Motor, Motorrad, Getriebe, Trans, Transporter, Felgen). Die Wörter 'Tuning' und 'Finanzierung' sind auf Dienstleistungen bezogene Wörter, die bei der manuellen Auswahl auch gesammelt werden können. Solche Wörter (z.B. Ersatzteile, Computer, Fahrer, Polizei) sind unwichtig in der Autobranche. Bei der manuellen Auswahl für EGT kann man zusätzlich auf zwei Punkte achten:

- Auf Dienstleistungen bezogene Wörter sammeln
- weitere Stoppwörter und domainneutrale Wörter sammeln

8.8.2 Ergebnisse des Autobranche-Korpus

In der folgenden Tabelle 8.2 werden die Ergebnisse dieses Experiments im Automobilbereich angezeigt:

Bei der Erstellung von "Frequenzliste 3" wurde die Worthäufigkeit beschränkt. Wenn die Worthäufigkeit eines Wortes kleiner als 6 ist, wird es nicht verarbeitet. Wenn Wörter Nicht-Wortzeichen beinhalten, werden sie als Stoppwörter betrachtet und nicht verarbeitet. Die 25.256 Wörter unter den 224.292 semantisch annotierten Termen aus der "Frequenzliste 3" im Beispiel-Korpus Automobilbereich werden bei der semantischen Annotation 'NO' gekennzeichnet. Während die 'ET'-annotierten Wörter eine gute Quelle für die

Korpus (Mit Suchmaschinen)	24.109 Webseiten
Frequenzliste 1	1.945.764 Terme
Frequenzliste 2	1.559.567 Terme
Frequenzliste 3	224.292 Terme
'NO' annotiert (Frequenzliste 3)	25.256 Terme
'ET' annotiert (Frequenzliste 3)	190.689 Terme
Kandidaten für EGT	2.769 Terme

Tabelle 8.2: Erkennung der Einwortterme im Autobranche-Korpus

neuen DST und EGT sind, stellen die 'NO'-annotierten Wörter eine gute Quelle für weitere Stoppwörter und unwichtige Wörter dar. Die Affixanwendung von EGT spielt eine entscheidende Rolle für die Erkennung der domainspezifischen Terme in den jeweiligen Bereichen.
Die "Frequenzliste 3 mit der semantischen Annotation" ist von guter Qualität. Die ersten Top-30 Terme wurden im Kapitel 6.4.5. ("Semantische Annotation der Einwortterme") angezeigt.

8.9 Hierarchische Struktur der Produktterme (PT)

Im Kapitel 4.8 ("GermaNet - Semantisches Wortnetz") wurden die folgenden zwei unterschiedlichen semantischen Relationen erklärt:

- **Lexikalische Relationen**: Synonymie, Antonymie (Ist-Gegenteil-Von)[12]
- **Konzeptuelle Relationen**: Hyponymie ('is-a'), Hyperonymie, Meronymie (Teil-Ganzes-Relation), Holonymie

Hyponymie/Hyperonymie

Die Hyponymie ist die wichtigste konzeptuelle Relation für Nomina und bildet eine hierarchische Struktur. Die hierarchische Struktur zwischen Wörtern (z.B. Autositz, Babyautositz, Autokindersitz, Kinderautositz, Baby adac autositz) kann durch den schon erwähnten 'Suffix-Gebrauch' zum Beispiel wie folgt festgestellt werden, wenn die Wörter einen gemeinsamen Kopf (z.B. Sitz) haben. Im Automobilbereich sollten solche Wörter (z.B. Autositz) als EGT

[12]Synonymie (Bedeutungsgleichheit), Antonymie und Opposition (Bedeutungsgegensatz), Homonymie (Ein Homonym ist ein Lexem, das unterschiedliche Bedeutungen haben kann, wie z.B. Bank (Sitzmöbel) und Bank (Geldinstitut).)

behandelt werden, nach der "Grundannahme für domainspezifische Terme" in Kapitel 4.3. Das bedeutet, dass der EGT 'Autositz' selber ein Kopf ist, also nicht weiter zerlegt werden kann.

A	Kopf (z.B. Sitz)	'A' ist ein Hyperonym von 'BA'.
BA	(Autositz)	'BA' ist ein Hyponym von 'A'.
BEA	'E' für eine Ergänzung (z.B. Autokindersitz)	'BEA' ist ein Hyponym von 'BA'.
EBA	(Kinderautositz)	'EBA' ist ein Hyponym von 'BA'.

Tabelle 8.3: Hyponymie-Beziehung mit dem Suffix-Gebrauch

Die drei Terme "Autokindersitz, Kinderautositz und Babyautositz" sind Unterunterbegriffe von Sitz. Außerdem sind sie untereinander synonym und Kohyponyme von Autositz. Viele verschiedene Kohyponyme von "Wagen" aus dem automatisch erstellten Korpus von Autobranche werden durch die Suffixanwendung von Wagen beispielsweise wie folgt erkannt. Dieses Korpus aus dem Web ist dafür sehr nützlich:

> Gebrauchtwagen, Gebraucht-Wagen, Neuwagen, Jahreswagen, Wohnwagen, Mietwagen, Sportwagen, Unfallwagen, Kleinwagen, Rennwagen, Kastenwagen, Lastwagen, Traumwagen, Lieferwagen, Leihwagen, Kinderwagen, Personenwagen, Dienstwagen, Krankenwagen, Lastkraftwagen, Streifenwagen, Abschleppwagen, Firmenwagen, Rettungswagen,...

Nach der folgenden Definition von John Lyons kann das Synonymieverhältnis als symmetrische Hyponymie definiert werden (z.B. X = fabrikneuer Wagen, Y = 0-km-Wagen) [Lyo68, S. 466]:

> Wenn X ein Hyponym von Y ist und wenn Y auch ein Hyponym von X ist, dann sind X und Y Synonyma.

Durch den Suffix-Gebrauch können Hyponyme, Hyperonyme und Kohyponyme automatisch erkannt werden, wenn sie einen gemeinsamen Kopf haben.

8.9.1 Hierarchieextraktor

Ein CGI-Programm für die Extraktion der hierarchischen Struktur von Produkttermen wurde für diese Arbeit erstellt. Es wird "Hierarchieextraktor[13]"

[13]'Hierarchieextraktor' http://knecht.cis.uni-muenchen.de/cgi-bin/kimda/k04/mkw/egtHierarcy.pl

genannt. Beispielsweise gibt es die folgenden vier Wörter für die Extraktion der hierarchischen Struktur:

Sommerreifen
Winterreifen
Regenreifen
Autoreifen

Das Wort 'Reifen' als EGT und der Kopf der oben genannten Komposita müssen bei der Zerlegung automatisch erkannt werden, weil KGT mindestens einen EGT beinhaltet.
Sommerreifen, Winterreifen, Regenreifen und Autoreifen sind Kohyponyme von 'Reifen'. Für das CGI-Programm über 'Hierarchieextraktor' (engl. hierarchy extractor) werden die folgenden zwei Methoden entwickelt:

- **Längste gemeinsame Zeichenkette** im Suffixbereich
- **Lexikon-Lookup** mit Hilfe von Maximum-Matching[14] im Suffixbereich

Die längste gemeinsame Zeichenkette zwischen "Sommerreifen" und "Winterreifen" ist "erreifen". Nach der ersten Methode muss die zweite Methode 'Lexikon-Lookup' nochmal durchgeführt werden, um als ein Wort bzw. Nomen identifiziert zu werden. Deswegen ist der 'Lexikon-Lookup' für den Zweck dieser Arbeit geeignet. In der Demo-Version werden die 41.528 einfache Nomen von CISLEX beim Lexikon-Lookup verwendet, um die EGT der zusammengesetzten Wörter im Suffixbereich zu erkennen.

Der folgende Algorithmus für 'Hierarchieextraktor' wird schrittweise im CGI-Programm ausgeführt:

a. Eingabetext in ISO-Latin normalisieren

b. Tokenisieren und dann Stopp-Wörter eliminieren

[14]Ein maximales Matching (Maximum-Matching) ist ein Matching, welches nicht mehr erweitert werden kann.

c. Um vorhandene EGT der jeweiligen zusammengesetzten Wörter im Eingabetext zu erkennen und bis zur Stufe der Unterunterbegriffe zu suchen, wird die folgende Annahme angewendet:

Wenn ein Wort die Form 'AB' hat und ein Wort der Form 'B' im Eingabetext vorhanden ist, ist die Form 'B' ein EGT.

```
Z.B. Eingabetext: Autositz, Autokindersitz, Babyautositz, Kinderautositz, Sitz

[S] sitz [n] -> autositz [Kopf: sitz]
[S] sitz [n] -> autositz -> babyautositz [Kopf: sitz]
[S] sitz [n] -> autositz -> kinderautositz [Kopf: sitz]
[S] sitz [n] -> autositz -> autokindersitz [Kopf: sitz]

[S] für Suffixgebrauch, [n] für einfache Nomen, ' ->' für Hyponym-Relation
```

d. **Lexikon-Lookup** mit Hilfe von Maximum-Matching im Suffixbereich

Wenn ein Wort die Form 'AB' hat und ein Wort der Form 'B' im Eingabetext nicht vorhanden ist, wird die Form 'B' als ein EGT durch den Lexikon-Lookup mit Hilfe von Maximum-Matching im Suffixbereich erkannt.

```
Z.B. Eingabetext: Autokindersitz, Babyautositz, Kinderautositz
[S] sitz [n] -> kinderautositz [Kopf: sitz]
[S] sitz [n] -> babyautositz [Kopf: sitz]
[S] sitz [n] -> autokindersitz [Kopf: sitz]
```

Ein Beispiel für 'Hierarchieextraktor':
Durch den Schritt [c] im oben genannten Algorithmus wurde der folgende Fehltreffer im Beispiel getroffen:

[S] sitz [n] -> autositz -> autobesitzer [Kopf: sitz]

In solchen Fällen soll der Schritt [d] im oben genannten Algorithmus vor dem Schritt [c] wie folgt durchgeführt werden:

Ein Thesaurus[15] bzw. Wortnetz ist in der Dokumentationswissenschaft ein kontrolliertes Vokabular, dessen Begriffe durch Relationen miteinander verbunden sind. Die Demo-Version soll mit Hilfe eines Thesaurus weiter entwickelt werden:

[S] besitzer [N] -> autobesitzer [Kopf: besitzer]

Das Wort 'Autobesitzer' ist kein DST, weil es entweder auf Produktnamen oder auf Dienstleistungen bezogen ist. Solche unwichtigen Wörter (z.B. Besitzer, Fahrer, Inhaber, Liebhaber, Fan) können durch den Vergleich von Korpora erkannt und entfernt werden. In der Abbildung 8.4 (am Ende des Kapitels) wird ein Beispiel-Ergebnis des CGI-Programms 'Hierarchieextraktor' gezeigt.

8.10 Semantische Klassen für E-Commerce im CISLEX

Die 41.528 Lexeme für einfache Nomina im CISLEX werden bei der semantischen Kodierung anhand der Merkmale (z.B. FZA) von Stefan Langer manuell wie folgt gekennzeichnet. Sie sind als Grundform eingetragen:

```
Auto;20454;n;S2;P6;#01;=FZA;;
Automobil;1622;n;S2;P2;#01#03;FZA&FZM;%Auto;
Car;503;m;S2;P6;#01;FZA&FZM;;
Ricercar;9;n;S2;P33;#01;XXX;;
Autocar;0;m;S2;P6;#01;FZA&FZM;;
Wagen;5986;m;S2;P0;#01;FZA|FZH;;
Beiwagen;17;m;S2;P0;#n;FZH;;
Leuwagen;1;m;S2;P0;#01;XXX;;
```

Das Feld 5 (z.B. P6) steht für die Pluralbildung. Vom Feld 1 bis Feld 6 werden morphosyntaktische Informationen annotiert. Im Feld 7 und 8 werden semantische Informationen (z.B. FZA (Autos, PKWs und LKWs), XXX (nicht klassifiziert, weil nicht bekannt)) annotiert.
Insgesamt sind ca. 429 semantische Klassen hierarchisiert, die für einfache Nomina des CISLEX semantisch manuell kodiert wurden. Davon werden 236 Klassen für E-Commerce (z.B. abf (Abfälle), ael (Elektrische Bauteile), agt

[15] In Wikipedia - http://de.wikipedia.org/wiki/Thesaurus

(Spektakel-/Theatertypen), ake (Kleineisenwaren), amb (Blasinstrumente), amc (Schlagzeug)) manuell ausgewählt. In diesen 236 Klassen werden 23.921 Lexeme, die schon semantisch kodiert wurden, identifiziert.
Die 23.921 Lexeme können als EGT für die Erkennung der DST verwendet werden, wenn sie den entsprechenden E-Commerce-Bereichen richtig zugeordnet sind.

8.10.1 Zuordnung der semantischen Klassen durch die Suffixanwendung

Durch die Suffixanwendung von 23.921 EGT für E-Commerce, die im CISLEX schon semantisch manuell kodiert wurden, können erkannte DST den jeweiligen semantischen Klassen zugeordnet werden. Dafür wird die Grammatik für die Suffixanwendung benutzt:

> **Suffixanwendung**: x hat die Form "AB" und B ist ein EGT. x gehört zu einer oder mehreren semantischen Klassen von B.

Die verwendete Metakeyword-Liste enthält die automatisch gesammelten 4.778.097 Einträge für Einwortterme aus dem Web. Zum Experiment ohne Infixgebrauch wurden Suffix- und Präfix-Gebrauch mit den schon semantisch kodierten 23.921 EGT bzw. einfache Nomina im CISLEX durchgeführt. Dabei wird die folgende Idee umgesetzt, um gute Wörter zu erkennen:

> Gute Wörter, die durch die Affixanwendung von EGT bzw. einfachen Nomina im Lexikon (z.B. CISLEX) erkannt werden.

Davon werden 2.125.737 Terme erkannt. Die 441.052 Terme darunter, die durch die Suffix- und Präfixanwendung von 23.921 EGT für E-Commerce identifiziert wurden, werden als gute Termkandidaten wie im folgenden Beispiel betrachtet:

```
‘‘S’’ für Suffix, ‘‘P’’ für Präfix

campingwagen [2;beide] [wagen-A][camping]
Wagen;5986;m;S2;P0;#01;FZA|FZH;;
Camping;168;n;S2;0;#01;AKT;;
```

```
funkautos [auto-S][funk]
Auto;20454;n;S2;P6;#01;=FZA;;
Funk;1303;m;S2;0;#01;ITU;;

autohändlern [auto-P][händler]
Auto;20454;n;S2;P6;#01;=FZA;;
Händler;5448;m;S2;P1;#01;=BVK;;
```

Im Experiment wurden die 236 semantischen Klassen von Stefan Langer verwendet, um die dazugehörigen semantischen Klassen eines Terms zu ermitteln. Das semantische Merkmal 'AKT' steht für "Aktionen". Das Merkmal 'FZA' steht für "Autos, PKWs und LKWs" und FZH für "Fahrzeuganhänger". Wörter (z.B. Campingwagen), deren semantisch zusammengesetzte Bedeutung (z.B. "AKT" und "FZA|FZH") richtig ist, können durch die Methode den jeweiligen semantischen Klassen (z.B. "FZA|FZH") zugeordnet werden.

Solche Wörter wie Funkauto, Kinderwagen gehören nicht zur Autobranche. Nach der Identifizierung der Affixanwendung von EGT können sie dann als Fehltreffer gefiltert werden. Das Wort 'Funkauto' wird durch Suffixgebrauch von "Auto" erkannt und das Präfix [funk] wird auch als ein EGT erkannt. Das Wort 'Funkauto' wird auf die zwei semantischen Klassen, nämlich FZA und ITU (Institutionen) referenziert. Das Wort 'Funkauto' gehört zu der Klasse FZA. In diesem Fall wird das Wort "funk" nicht richtig semantisch kodiert, und die Analyse ist falsch.
Die zusammengesetzten Klassen von 'Autohändlern' sind FZA (Autos) und BVK (Verkäufer).

Wenn die zusammengesetzte Bedeutung eines Wortes richtig ist, kann sie wie ein Kompositum[16] behandelt werden.

Solche kommerziellen EGT (z.B. Notebook, Laptop, Bildschirm, Keybords, DVD, PC, MP3) sind im CISLEX noch nicht semantisch kodiert, weil sie dort noch nicht vorhanden sind.

[16]Unter Zusammensetzung (Komposita) verstehen wir Wörter, die ohne Ableitungsmittel aus zwei oder mehreren selbständig vorkommenden Wörtern gebildet sind. Die Bestimmungswörter stehen links, sie erklären das Grundwort näher. Das letzte Wort ist das Grundwort, das die Wortart der ganzen Zusammensetzung festlegt. [Duden 4]

Neue EGT für E-Commerce müssen erweitert werden. Sie können in zwei typischen Bereichen - Produktnamen und Dienstleistungen - sowie in Sektoren und Branchen semantisch kodiert werden. Damit kann man DST im jeweiligen Bereich identifizieren und semantisch analysieren, ob ein Term zu Produktnamen, Dienstleistungen oder zu einer anderen Branche gehört.

8.11 Erkennung der auf Dienstleistungen bezogenen Terme

Dienstleistungen im E-Commerce-Bereich sind persönliche Leistungen, um Geld zu verdienen. Die Grammatik für "Erkennung der auf Dienstleistungen bezogenen Terme" (z.B. Autovermietung, Autoversicherung, PKW-Service, Autotuning, Schuhreparatur) sieht wie folgt aus:

> x ist ein möglicher auf Dienstleistungen bezogener Term, wenn x die Form "AB" hat. Wenn 'A' ein EGT oder KGT ist, ist 'B' ein Abstraktum für Dienstleistungen (z.B. Service, Tuning).

Ein Beispiel mit dem EGT 'Schuh'
Die aktuell verwendete Metakeyword-Liste enthält 2.988.819 Einträge. Die Einträge sind Unigramme. Durch Suffix- und Präfix-Gebrauch mit dem EGT von Schuh wird die folgende Anzahl von Treffern in der Metakeyword-Liste angezeigt:

EGT (Schuh)	Trefferanzahl von 2.988.819
Suffix-Gebrauch	2.001
Präfix-Gebrauch	744

Der Präfix-Gebrauch ist gut geeignet für die Erkennung der auf Dienstleistungen bezogenen Terme (z.B. Schuhmacher, Schuhtechnik, Schuhmode, Schuhgeschäft, Schuhpflege, Schuhreparatur, Schuheinlagen).
Der Suffix-Gebrauch ist geeignet zur Erkennung der Produktnamen (z.B. Handschuhe, Damenschuhe, Herrenschuhe, Kinderschuhe, Sportschuhe, Sicherheitsschuhe, Maßschuhe, Mädchenschuhe)

Abstrakte Basiswörter für Dienstleistungen

Die abstrakten Basiswörter für Dienstleistungen können für die Erkennung der auf Dienstleistungen bezogenen Terme effizient gebraucht werden. Im Gegensatz zu EGT sind sie in den jeweiligen Bereichen domainneutral. Sie müssen für alle Domänen gleich verwendet werden, wie das folgende Beispiel zeigt:

> *M*obilfunkservice, Pannenservice, Autoservice, Dolmetscherservice, Aircraftservice, Elektroservice, Computer-Reinigungs-Service, DVD-Brennservice, Notservice, Filmservice, EDV-Service

Um die abstrakten Basiswörter für Dienstleistungen zu extrahieren, kann der Präfix-Gebrauch von EGT genutzt werden. Die 2.940 Terme aus den 224.292 semantisch annotierten Termen im Beispiel-Korpus von Autobranche wurden durch den Präfix-Gebrauch von EGT erkannt. Dadurch kann die in Tabelle 8.4 dargestellte Wortliste mit den Frequenzen der Kandidaten für abstrakte Basiswörter (Dienstleistungen) aus einem Beispiel-Korpus von Autobranche automatisch erstellt werden. Schließlich können qualifizierte Basiswörter für Dienstleistungen manuell wie EGT selektiert und verbessert werden.

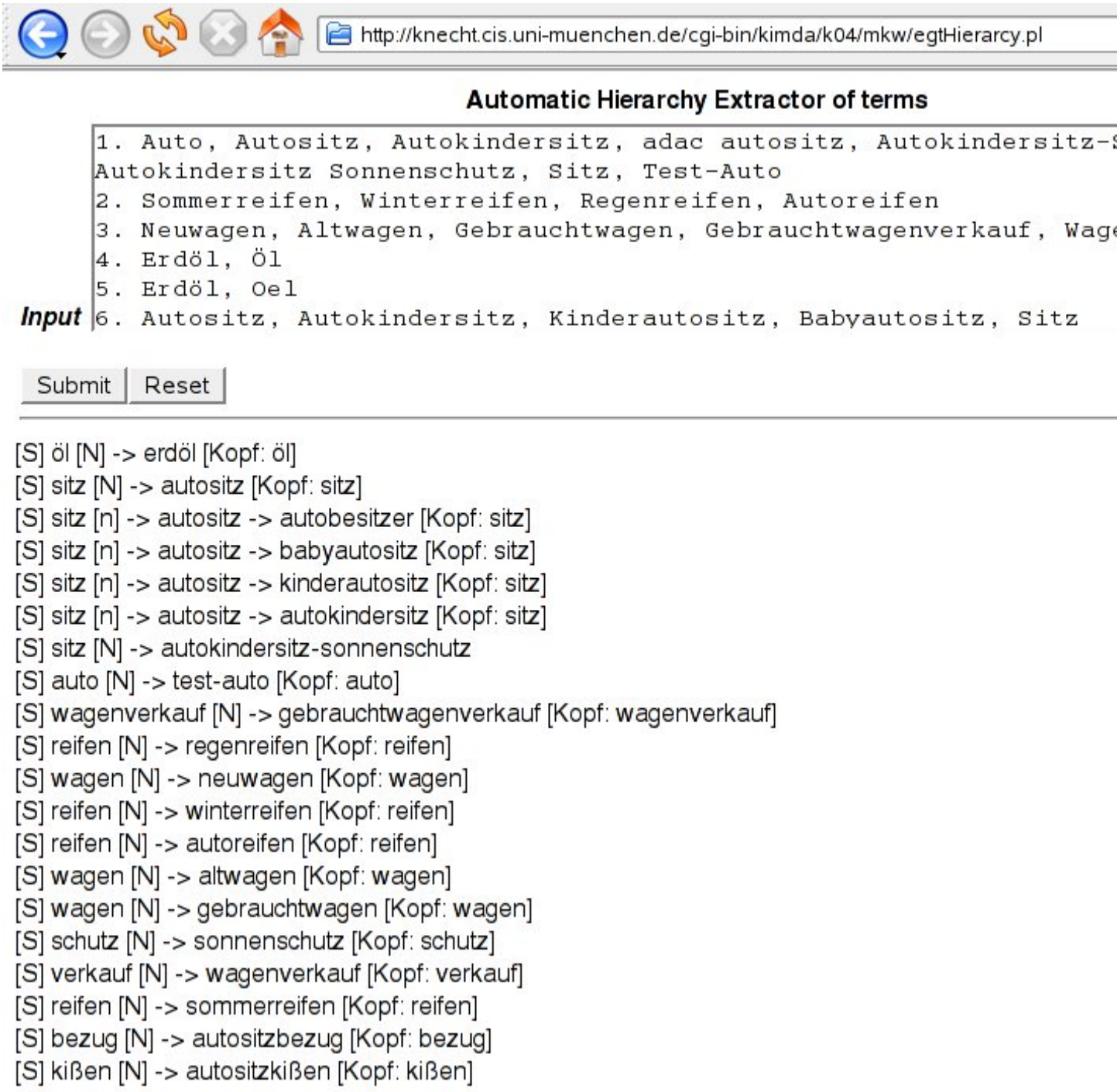


Abbildung 8.4: Ein Beispiel von 'Hierarchieextraktor'(Stand: 29.03.2007)

Frequenz	Kandidat	manuell selektiert
13	fahrer	
7	vermietung	+
7	modell	
7	markt	
6	verleih	+
6	modelle	
6	fahren	+
5	verkehr	+
5	tuning	+
5	teile	
5	shop	
5	reisen	+
5	preise	
5	pflege	+
5	fan	
5	club	
5	betrieb	+
4	world	
4	versicherung	+
4	ver	
4	umbau	+
4	service	+
4	reinigung	+
4	reiniger	+
4	museum	
4	messe	
4	hersteller	
4	handel	+
4	forum	
4	fahrten	
4	fahrt	
4	fahrern	
4	dach	
3	zeitung	
3	werbung	
3	welt	
3	vermittlung	+
3	verkauf	+

Tabelle 8.4: abstrakte Basiswörter für Dienstleistungen durch die Präfixanwendung

Kapitel 9

Automatische Gewinnung von branchenspezifischem Vokabular (AGBV)

Webseiten in den jeweiligen E-Commerce-Branchen können durch domainspezifische Ausdrücke bzw. Nominalphrasen automatisch klassifiziert werden. Die domainspezifischen Ausdrücke beinhalten mindestens einen domain-spezifischen Bestandteil, der in dieser Arbeit EGT oder Marke genannt wurde. Wenn eine Webseite nur aus den folgenden zwei Sätzen besteht, können Schlüsselwörter (Keywords) durch statistische Verfahren für Schlüsselwortextraktion (engl. keyword extraction) erkannt werden. Aber nicht alle Schlüsselwörter sind domainspezifisch. Durch den Abgleich der Frequenzlisten in den jeweiligen E-Commerce-Branchen und der semantischen Annotation für branchenspezifisches Vokabular können Wörter in einem Bereich als branchenspezifisch erkannt werden. Nur domainspezifische Nominalphrasen, wie zum Beispiel im Bereich "Automobil", werden in dieser Arbeit berücksichtigt:

> Bei MAN und Scania stehen die Zeichen auf Fusion. Beim schwedischen Lkw-Bauer Scania hat die Hauptversammlung VW-Chef Martin Winterkorn wie erwartet zum Aufsichtsratschef gewählt.

Die folgenden branchenspezifischen Nominalphrasen aus den oben genannten Sätzen werden beispielsweise manuell erstellt:

- **Schlüsselwörter**:
  MAN, Scania, Zeichen, Fusion, Lkw-Bauer, Hauptversammlung, VW-Chef Martin Winterkorn, Aufsichtsratschef

- **branchenspezifische Nominalphrasen**:
  MAN und Scania, Lkw-Bauer Scania

Berufsbezeichner (z.B. Aufsichtsratschef) sind nicht branchenspezifisch. Eine Nominalphrase wie z.B. VW-Chef Martin Winterkorn (Berufsbezeichner + Personenname) ist zwar branchenspezifisch, jedoch werden Personennamen in dieser Arbeit nicht behandelt.
Branchenspezifisches Vokabular, das auf Produkte und Dienstleistungen bezogen ist, kann in dieser Arbeit hauptsächlich automatisch identifiziert werden. Ausdrücke wie z.B. myAudi, myBMW werden in dieser Arbeit nicht behandelt. In diesem Kapitel werden die folgenden Zielsetzungen experimentell durchgeführt:

I. **Automatische Erstellung der deutschen Korpora für E-Commerce-Branchen**

II. **Erstellung des Terminologie-Extraktionssystems AGBV**:
  Automatische **G**ewinnung von **B**ranchenspezifischem **V**okabular aus den erstellten Korpora

9.1 Überprüfung der erkannten Wörter in einer Branche

Branchenspezifische Wörter in den jeweiligen Bereichen maschinell zu erkennen ist das Hauptziel des Experiments. Dafür kann ein Wort in einer Branche unter der folgenden Annahme überprüft werden:

> Ein Wort in einem Text ist für eine Branche entweder ein "branchenspezifisches Wort" oder ein "Stoppwort" (bzw. branchenneutrales Wort).

Aufgrund der oben genannten Annahme wird die Struktur zur Überprüfung der erkannten Wörter in einer Branche abgebildet. Dabei werden die folgenden Abkürzungen verwendet:

Abkürzung	Bemerkung
BW	Branchenspezifisches Wort
F	falsch erkanntes BW
S1	allgemeine Stoppwortliste
S2	Stoppwortliste für geographische Namen
S3	Stoppwortliste für "Branchenneutrale Stoppwörter"

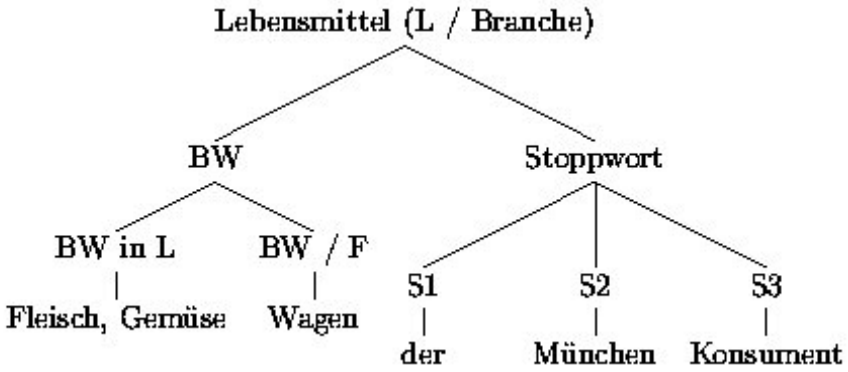


Abbildung 9.1: Überprüfung der erkannten Wörter in einer Branche

Branchenspezifische Wörter, die durch die 'AGBV' als 'branchenspezifisch' automatisch erkannt werden, sind Basis für die Erstellung der "Branchenspezifischen Wortlisten" in den jeweiligen Branchen.
Zur Qualitätsverbesserung sind die "Branchenspezifischen Wortlisten", die zusätzlichen Wortlisten von "BW / F (falsch erkanntes BW)" und "Stoppwörter" sehr nützlich.

"BW / F (falsch erkanntes BW)" wird ein Wort genannt, das in einer Branche häufig als Fehltreffer vorgekommen ist, aber als 'branchenspezifisch' in anderen Branchen betrachtet werden kann.
Die Liste von "BW / F (falsch erkanntes BW)" wird verwendet, um die korrekten branchenspezifischen Wörter in einer entsprechenden Branche zu erkennen. Diese Liste gilt für eine Branche.
Das Beispielwort 'Wagen' ist branchenspezifisch und ein EGT in der Autobranche. In den anderen Bereichen, z.B. für Lebensmittel und Computer, ist das Wort 'Wagen' nicht bereichsspezifisch. Solche Wörter können zur Qualitätsverbesserung der Liste manuell korrigiert werden.

Die Datei '`Computer_BVGap_2007-06-13_17.35.54_.list`' im Beispielbereich von 'Computer' enthält die besten automatisch erkannten branchenspezifischen Wörter mit Großschreibung, deren Abstandswert höher als '0.90' ist. Bei der manuellen Verarbeitung dieser Datei können solche Fehltreffer als "BW / F (automatisch falsch erkanntes BW)" gekennzeichnet und gesammelt werden.
Die oben erwähnten drei Stoppwortlisten gelten für alle Branchen.
Sie werden in dieser Arbeit erfolgreich eingesetzt und können erweitert werden.

9.2 Branchenspezifische Wörter (BW) pro Branche

Nicht immer kann ein Wort einer Branche eindeutig zugeordnet werden. Ein Beispielwort ist "Reifen". Das Wort "Reifen" wird im Automobilbereich (z.B. Winterreifen, Sommerreifen), aber auch z.B. in der Schmuckbranche als EGT betrachtet (z.B. Armreifen, Stirnreif, Goldreifen, Haarreifen, Halsreifen, Jacquardstreifen, Magnetarmreifen, Metallreifen, Oberarmreifen, Omega-reifen, Silberarmreifen).
Jede Branche (B) hat "Branchenspezifische Wörter (BW)", die zu dieser Branche gehören. Die folgende Tabelle zeigt dafür:

Branchen	B_1	...	B_n
Branchenspezifische Wörter	BW_1	...	BW_n

Tabelle 9.1: Branchenspezifische Wörter (BW) pro Branche

9.3 Branchenspezifische Wortlisten

In den in Kapitel 9. erwähnten zwei Beispielsätzen können "MAN" und "Scania" als Firmennamen erkannt werden, wenn die beiden Wörter in der Liste von Firmennamen vorhanden sind.
"MAN" ist der größte deutsche Hersteller von LKWs und Spezialfahrzeugen in der Autobranche. "Scania" ist ein schwedischer LKW-Hersteller.
Aber das Wort "MAN" könnte darüber hinaus auch ein Firmenname sein,

z.B. für Möbel oder Computer, wenn das Wort "MAN" in der Firmenliste (z.B. von Möbeln oder Computern) vorhanden ist. Firmennamen können also auch mehrdeutig sein.
Jeder Bereich hat die jeweiligen branchenspezifischen Wortlisten.
Die folgenden branchenspezifischen Wortlisten werden in dieser Arbeit berücksichtigt. Sie können automatisch erkannt und zur Vermeidung von Fehltreffern manuell korrigiert werden. Sie werden in folgende zwei Gruppen eingeteilt:
Unter der Gruppe "Branchenspezifische Wörter in einer Branche (z.B. Autobranche)" liegen vier Wortlisten, nämlich Marken- bzw. Firmennamen, EGT, KPN und Abkürzungen, die für die automatische Erkennung von branchenspezifischen Einworttermen und Mehrworttermen nützlich sind. Sie sind eine Basis für die semantische Analyse und Webseitenklassifikation:

a. **BW in einer Branche (z.B. Autobranche)**

- **Marken- bzw. Firmennamen**: *Opel, BMW, Audi, VW*
- **EGT**: *Wagen, Auto, Fahrzeug*
- **branchenspezifische Abkürzungen mit ähnlichen Eigenschaften wie EGT**: KFZ, PKW, LKW, ADAC
- **KPN** (Konkrete Produktnamen): *VW Polo 1.2*

b. **BW / F (automatisch falsch erkanntes BW)**

"KPN (Konkrete Produktnamen)", die im Kapitel 8.5.2. beschrieben wurden, können Marken, Modelle und Zusatzinformationen beinhalten. Solche konkreten Ausdrücke (z.B. BMW 116i Advantage, VW Polo 1.2) werden als 'KPN' erkannt. Branchenspezifische Abkürzungen mit ähnlichen Eigenschaften wie EGT (z.B. KFZ, PKW) werden in dieser Arbeit als 'EGT' behandelt, weil sie fähig zur Kompositabildung sind.
Die manuell erstellte Wortliste für "BW / F (falsch erkanntes BW)" kann nur in einer entsprechenden Branche (z.B. 'Autobranche') benutzt werden.

9.4 Branchenneutrale Stoppwörter

Es gibt verschiedene Stoppwort-Listen für die jeweiligen Anwendungen. In dieser Arbeit wurde die allgemeine Stoppwortliste, die 1220 Einträge enthält, bei

der Erstellung von den jeweiligen Frequenzlisten in allen Bereichen verwendet. Typische Elemente dieser allgemeinen Stoppwortliste sind die folgenden:

- **Inhaltsleere (Funktions-)Wörter** wie Artikel, Präpositionen, Pronomina und Konjunktionen für Deutsch und Englisch

- **Branchenneutrale Wörter**: Z.B. erhöht, gibt, kurz, Türen, Insassen, reichen, Liter, bleibt, nennt, Euro, EUR

Diese 1220 Stoppwörter sind sehr nützlich für allgemeine Anwendungen. Für die Terminologie-Extraktion von branchenspezifischem Vokabular ist diese Menge an Stoppwörtern zu gering. Nach der Nutzung von 1220 Stoppwörtern werden Frequenzlisten für Einwortterme in allen Bereichen erstellt. Die Frequenzlisten enthalten noch weitere Stoppwörter und branchenneutrale Wörter, die nicht als branchenspezifisch betrachtet werden können. Branchenneutrale Stoppwortlisten können zur Beseitigung weiterer Stoppwörter und branchenneutraler Wörter effizient eingesetzt werden. Die branchenneutralen Wörter werden in dieser Arbeit behandelt wie Stoppwörter.

Branchenneutrale Stoppwörter gelten für alle Branchen.
Für das Experiment werden 20 E-Commerce-Branchen ausgewählt (vgl. Tab. 9.2). Im Experiment werden die branchenneutralen Stoppwörter, die mehrmals vorgekommen und in mehr als 15 Branchen aufgetaucht sind, in einer Datei (z.B. '`Computer_Birrelevant_*`' im Beispielbereich von 'Computer') in den jeweiligen Branchen automatisch gesammelt, um bei der Eliminierung der Stoppwörter verwendet zu werden. Die Qualität der branchenneutralen Stoppwörter muss letztlich manuell verbessert werden.

9.5 Auswahl von E-Commerce-Branchen

Es gibt zahlreiche Branchen im Internet-Handel, die noch nicht vereinheitlicht sind. Die Zuordnung von Waren zu Branchen ist deshalb ein großes Problem, wenn Waren in den Produktklassifikationssystemen und "Offline-Katalogen" aufgefunden werden sollen. Für die automatische Gewinnung von branchenspezifischem Vokabular (**AGBV**) werden die zwei Kategorien "Produkte" und

"Dienstleistungen" unterschieden. Die folgenden 20 E-Commerce-Branchen für Produkte und Dienstleistungen sind in dieser Arbeit relevant. Mit dieser exemplarischen Auswahl an Branchen wird das Experiment durchgeführt. Bei der Kodierung steht 'P' für 'Produkte' und 'D' für 'Dienstleistungen':

Kodierung	E-Commerce-Branchen
P1	Autobranche
P2	Haushaltsgeräte
P3	Möbel
P4	Kleidung
P5	Büroartikel
P6	Kosmetik
P7	Wein
P8	Computer
P9	Lebensmittel
P10	Schmuck
D1	Reisen
D2	Bank
D3	Versicherung
D4	Gesundheit
D5	Hotel
D6	Immobilien
D7	Finanzen
D8	Restaurant
D9	Ärzte
D10	Altenpflege

Tabelle 9.2: E-Commerce-Branchen für Produkte und Dienstleistungen

9.6 Korpora für E-Commerce-Branchen

Im Kapitel 6 ("Domain-spezifische Korpora aus dem Web") wurden die folgenden zwei Methoden für die Erstellung der domain-spezifischen Korpora vorgestellt:

a. Extraktion aus Startseiten (z.B. www.autoscout24.de, www.schmuck.de)

b. Extraktion mit Suchmaschinen (z.B. Google, Yahoo)

Die Methode [b] ist für das Experiment besser geeignet, weil mehr unterschiedliche Quellen für die "HTML-Analyse", die im Kapitel 6.2. ("Dokumentensammlung") vorgestellt wurde, gefunden werden.

9.6.1 Masterprogramm für den Aufbau der Korpora

Auf der beigefügten DVD sind alle verwendeten Programme, die im Rahmen dieser Arbeit implementiert werden, die verwendeten Korpora und Frequenzlisten gespeichert.
Für die Methode [b] "Extraktion mit Suchmaschinen (z.B. Google, Yahoo)" werden bisher Perl-Skripten und nützliche Linux-Befehle verwendet. Diese Perl-Skripten sind als Standalone-Programme modular aufgebaut und erfordern Eingaben bzw. Parameter. Die folgenden drei Programme erfüllen die wichtigen Aufgaben von URL-Sammlung, Korpusbau und Erstellung der Frequenzlisten:

Programm	Eingabe (Parameter)
`BV_urls_crawler.pl`	File Maximum DB[1] Zeitinterval Language Domain SE[2]
BVfloskelnCrawlerKorpus.pl	File
`BV_frequence.pl`	`Directory_Name`

Das Masterprogramm für den Aufbau der Korpora verbindet diese Perl-Skripten durch Linux-Befehle. Folgende Einstellungen können anhand des Masterprograms vorgenommen werden:

- Datenbanken, um verschiedene und neue URL-Adressen zu sammeln.
- Sprache (z.B. 'a' für "Any Language", 'g' für 'German')
- Domäne (z.B. 'any' für "any domain", de, com)
- Suchmaschinen (z.B. 'A' für 'All', 'G'(Google), 'Y' (Yahoo))

Für mein Ziel werden Korpora für die deutsche Sprache im E-Commerce-Bereich aufgebaut. Die folgenden zwei Voraussetzungen für das Masterprogramm werden benötigt, um ein Korpus aufzubauen:

[2] 'DB' für 'Datenbank'
[2] 'SE' für 'Search Engine'

- **Eine Datei** enthält branchenspezifische Wörter als "Startwörter", um ein Korpus aufzubauen. In dieser Datei steht in jeder Zeile ein Wort.
- **Ein Wort** wird für alle maschinell erstellten Suchbegriffe für Suchmaschinen , die aus zwei und sechs Wörtern bestehen, gebraucht. Beispielsweise wird das Wort "Haushaltsgeräte" im Bereich "Haushaltsgeräte" als ein Basis-Wort für alle Suchbegriffe verwendet, die wie folgt zusammengeschrieben werden: z.B. "Haushaltsgeräte Miele".

Input und Output des Masterprogramms

Der Input für das Masterprogramm sind die oben erwähnten zwei Parameter, nämlich ein Dateiname und ein Basis-Wort. Es folgt ein Beispiel für ein Korpus im Bereich von 'Haushaltsgerät', um das Masterprogramm auszuführen:

`BV_masterKorpus.pl ./suchbegriffe/Haushaltsgeraete` Haushaltsgeräte
(Dateiname: './suchbegriffe/Haushaltsgeraete', Basis-Wort: 'Haushaltsgeräte')

Nach der Ausführung des Masterprogramms werden die elf Dateien im Beispiel-Bereich 'Computer' als Output wie folgt maschinell erstellt:

```
SecondaryLevelUrls.urls
SecondaryLevelUrlsForStarturls.urls
allunique.urls
fre.list
freVariation.list
frezipfsize.list
korpus_suchbegriffeForComputer_2007-05-18_16.57.08_10000.urls_finished
korpus_suchbegriffeForComputer_2007-05-19_05.26.59_10001.txt
logfile.log
runtime_doubles.info
urlfile_Computer_2007-05-18_16.29.12_10000.urls
```

Die folgenden sechs Dateien sind besonders relevant:

- 'allunique.urls': **Sammlung der verschiedenen URLs**
- 'korpus_suchbegriffeForComputer_*': **Dateien für ein Korpus**. Verschiedene Webseiten werden mit Hilfe der gesammelten verschiedenen URLs in den Dateien lokal gespeichert, indem sie durch die 'HTML-Analyse' analysiert werden. Es wird als Test beispielweise die Einschränkung von 10000 HTML-Seiten pro Datei gemacht. Wenn diese Datei

10000 HTML-Seiten beinhaltet, wird der Dateiname mit der Zeitangabe wie folgt automatisch geändert:

korpus_suchbegriffeForComputer_2007-05-18_16.57.08_10000.urls_finished
korpus_suchbegriffeForComputer_2007-05-19_05.26.59_10001.txt

- '`runtime_doubles.info`': Laufzeit des Korpusbaus und Informationen für die Gesamtgröße der Original-Webseiten und Duplikate, die unterschiedliche Pfadangaben haben, wo aber Dateiname und Dateigröße gleich sind.

- '**fre.list**': Nach der Eliminierung der Stoppwörter, die durch die 1220 Einträge lange Stoppwortliste identifiziert wurden, wird die Frequenzliste 'fre.list' für Einwortterme automatisch erstellt. Der Lemmatisierungsprozess wird für das Experiment nicht verwendet. Wörter werden "case sensitive" behandelt, da auf diese Weise die für die weitere Verarbeitung relevanten Nomen aufgrund der Großschreibung leichter erkannt werden können. Die Worthäufigkeit und das Wort mit orthographischen Varianten werden zeilenweise wie folgt geschrieben:

47131 Computer
42669 computer
32513 Dell
23146 laptop
21639 Laptop

- '**freVariation.list**': Die Datei "freVariation.list" wird aus der Datei "fre.list" erstellt, indem orthographische Varianten eines Wortes zusammengefaßt werden. Die jeweilige Worthäufigkeit wird addiert. Z.B. ist die Worthäufigkeit von 'laptop' höher als die Worthäufigkeit von 'Laptop'. Dadurch wird das Wort 'laptop' als Grundform nach dem Vergleich der Frequenzen vom Programm ausgewählt. Die Auswahl 'laptop' als Grundform ist richtig. Wenn die Qualität des Korpus gut ist, muß die Auswahl für die Grundform richtig sein. Die häufigsten orthographischen Varianten werden als Grundform (bzw. Normalform) eines Wortes ausgewählt und an erster Stelle nach Worthäufigkeit wie folgt geschrieben:

94359	Computer/ computer/ COMPUTER/ CompUTer/ COMputer/ COmputer/ com-puter
46547	laptop/ Laptop/ LAPTOP/ LapTop/ lap-top/ laPtoP/ lapToP/ Lap-top/ LAPtop
40699	Dell/ DELL/ dell/ DeLL/ DEll

- '**SecondaryLevelUrlsForStarturls.urls**': Für die Erweiterung des Korpus werden gute Kandidaten von Start-Urls für die Methode [a] "Extraktion mit Startseiten" (z.B. www.autoscout24.de, www.schmuck.de) selektiert, indem Terme aus der jeweiligen URL mit dem verwendeten branchenspezifischen Vokabular für Suchmaschinen verglichen werden. Die Start-Urls sind nach Frequenz wie folgt sortiert:

```
4 www.jeans-preisguenstig.de [jeans]
3 www.x-jeans.de [jeans]
3 www.wilder-westen-web.de [westen]
2 www.westen-usa.de [westen]
2 www.kinder-kleider-basar.de [kleider]
```

9.6.2 Algorithmus für Masterprogramme

Es folgt eine Grobdarstellung des Algorithmus für den Aufbau der Korpora:

Masterprogramm für die Methode [b] "Extraktion mit Suchmaschinen"
(a) Manuelle Erfassung der Startwörter
(b) maschinelle Erstellung der Suchbegriffe für Suchmaschinen
(c) Sammlung der verschiedenen URL mit Hilfe der Suchmaschinen
(d) lokale Speicherung der verschiedenen Webseiten
(e) automatische Erstellung der Frequenzlisten
(f) Liste für "Secondary Level URLs" maschinell erstellen

Masterprogramm für die Methode [a] "Extraktion aus Startseiten", das von Schritt (c) bis Schritt (e) gleich ist
(i) Manuelle Auswahl der Start-URLs
(ii) Extraktion der internen Links, dann verschiedene URL-Adressen sammeln

Tabelle 9.3: Algorithmus für den Aufbau der Korpora

9.6.3 Suchbegriffe für Suchmaschinen

Branchenspezifische Suchbegriffe in den E-Commerce-Branchen zu gewinnen, ist nicht schwierig. Es gibt verschiedene Arten, dabei vorzugehen.
Ein Suchbegriff ist der erste Schritt, um erwartete Webseiten mit Hilfe der Suchmaschinen aufzufinden. Deshalb sollte man einen relevanten Suchbegriff richtig auswählen. Die einfachste Lösung ist wohl mit Hilfe der bekannten

Produktklassifikationssysteme, Kataloge und Suchmaschinen, um schon erkanntes branchenspezifisches Vokabular in den jeweiligen Bereichen zu finden. Die Lieferantensuchmaschine "Wer liefert Was?"[3] für deutsche Firmen ist ein nützliches Hilfsmittel dafür. Damit kann man nach Produkten und Dienstleistungen suchen. Die Treffer eines Suchbegriffs werden den passenden branchenspezifischen Begriffen im Klassifikationssystem maschinell zugeordnet, um die erweiterte Suche zu präzisieren.
Branchenspezifische Wörter als Startwörter für die zweite Methode "Extraktion mit Suchmaschinen" werden in den jeweiligen E-Commerce-Branchen benötigt. Für die ausgewählten 20 E-Commerce-Branchen werden die folgenden branchenspezifischen Wörter für "Test 1" verwendet. In Klammern steht die Anzahl der verwendeten branchenspezifischen Wörter:

Tabelle 9.4: **Branchenspezifische Wörter als Startwörter für "Test 1"**

Branche	**Startwörter für "Test 1"**
P1 (14)	Porsche, Fahrzeuge, BMW, KFZ, PKW, LKW, Auto, Wagen, VW, Automobil, Gebrauchtwagen, Neuwagen, ADAC, Mercedes
P2 (35)	Miele, Akkusauger, Bodenstaubsauger, Bügeleisen, Bügelstationen, Bosch, Dampfreiniger, Handstaubsauger, Nähmaschinen, Trockner, Waschmaschinen, Privileg, Braun, Heizlüfter, Klimageräte, Luftbefeuchter, Luftentfeuchter, Luftkühler, Luftreiniger, Ventilatoren, Allesschneider, Brotbackautomaten, Dampfgarer, Elektro-Grills, Fondue-Geräte, Fritteusen, Handrührer, Küchenmaschinen, Küchenwaagen, Mikrowellengeräte, Raclette-Grills, Sandwichtoaster, Stabmixer, Standmixer, Mixer
P3 (18)	Schrank, Stuhl, Regal, Schubladen, Sofas, Schlafsofas, Sessel, Hocker, Esstische, Esszimmerstühle, Barstühle, Sideboards, Vitrinen, Betten, Nachttische, Kleiderschränke, Matratzen, Kommode
P4 (23)	T-Shirts, Hosen , Jeans , Bademode , Kleider , Röcke , Blusen, Jacken, Langarmshirts, Tuniken, Babydolls, Hosenanzüge, Pullover, Poloshirts, Gürtel, Strickjacken, Shirtjacken, Taschen, Sweatshirts, Kostüme, Hemden, Westen, Mäntel
P5 (21)	Locher, Datenträger, Tacker, Heftgeräte, Heftklammern, Papierkörbe, Mappen, Scheren, Kalender, Sichthüllen, Aktenhüllen, Ordner, Stifte, Bleistifte, Kugelschreiber, Textmarker, Edding, Lineale, Haftnotiz, Tinte, Toner
P6 (26)	Eyeliner, Eyeshadow, Lipstick, Lipliner, Lipgloss, Cover, Mascara, Fluid, Powder, Brush, Rouge, Eyefect, MakeUp, MakeUp-Entferner, Puder, Rouge, Fußnägel, Nagellacke, Nagelpflege, Lippen, Lippenpflege, Lippenstifte, Körperpflege, Haarpflege, Sonnenpflege, Damenduft
P7 (26)	Rotwein, Weißwein, Winzer, Weingüter , Champagner, Sekt, Prosecco, Bordeaux-Weine, Auxerrois, Bacchus, Chardonnay, Clevner, Ehrenfelser, Elbling, Freisamer, Gewürztraminer, Grauburgunder, Gutedel, Huxelrebe, Kerner, Moriomuskat, Muskateller, Müller-Thurgau, Weißburgunder, Acolon, Blauburger

[3] http://www.wlw.de/start/DE/de/index.html

Tabelle 9.4: **Branchenspezifische Wörter als Startwörter für "Test 1"**

Branche	Startwörter für "Test 1"
P8 (23)	Notebook, Computerperipheriegeräte, Computer-Reinigungsmittel, Computer-Schriften, Laptop, IBM, Dell, Samsung, Computertische, Computertomographen, Drucker, Rechner, Militärcomputer, Minicomputer, Musikcomputer, Notebook-Akkus, Notebook-Halterungen, Notebook-Koffer, Notebook-Ständer, Notebook-Taschen, Notepad-Computer, Maus, Tastatur
P9 (35)	Kaffee, Tee, Kakao, Säfte, Konzentrate, Wein, Spirituosen, Honig, Konfitüre, Zucker, Süssmittel, Müesli, Flocken, Getreide, Reis, Pasten, Dörrgemüse, Pilze, Gewürze, Kräuter, Öl, Essig, Schokolade, Snacks, Nüsse, Dörrfrüchte, Lidl, Haribo, Gurken, Käse, Getränke, Milch, Milchprodukt, Brot, Kartoffel
P10 (22)	Anhänger, Armbänder, Armreife, Broschen, Colliers, Creolen, Ketten, Ohrhänger, Ohrstecker, Panzerketten, Piercing, Körperschmuck, Ringe, Schmucksets, Trauringe, Uhren, Accessoires, Edelsteine, Goldschmuck, Perlen, Silberschmuck, Modeschmuck
D1 (29)	Pauschalreisen, Frühbucherreisen, Städtereisen, Länderinfos, Flughafeninfos, Währungsinfos, Partnerangebote, Kreuzfahrten, Ferienhäuser, Mietwagen, Flughafentransfer, Reiseshop, Reiseversicherung, Reisedatum, Reiseziel, Reiseführer, Pauschalreisen, Lastminute-Reisen, Reisesuchmaschine, Lastminute, Ferienwohnungen, Badeferien, Studienreisen, Sprachreisen, Geschäftsausflüge, Mitsegeln, Reiseliteratur
D2 (38)	Commerzbank, Konto, Privatkunden, Girokonten, Komfortkonto, Startkonto, Online, Kreditkarten, Reisezahlungsmittel, Geschäftskunden, Business-Konto, TagesGeld-Konto, MasterCard, Kartenzahlungs-terminals, Kontostand, Dispokredit, BLZ, Stadtsparkasse, Bankenautomation, Geldinstitute, Partnerkarte, Kontoauszug, Kreditkartenabrechnung, Guthabenverzinsung, Internet-Konto, Kredit, Postbank, KFZ-Leasing, Anlegen, Sparen, Bausparen, Baufinanzierung, Vermögensberatung, Zahlungsverkehr, Kartenakzeptanz, Leasing
D3 (34)	Bauleistungsversicherungen, Betriebshaftpflichtversicherungen, Betriebsunterbrechungsversicherungen, Elementarschadenversicherungen, Sozialversicherungsrecht, Gebäudeversicherungen, Versicherungsschäden, Industrieversicherungen, Kraftfahrzeugversicherungen, Krankenversicherungen, Kreditversicherungen, Lebensversicherungen, Gruppenversicherung, Maschinenversicherungen, Rechtsschutzversicherungen, Reiseversicherungen, Rückversicherungen, Sachversicherungen, Versicherungswesen, Versicherungsdienstleister, Transportversicherungen, Unternehmensversicherungen, Vermögensschaden-Haftpflichtversicherungen, Versicherungsformulare, Versicherungsmakler, Versicherungsvergleiche, KFZ-Versicherung, Wohngebäudeversicherung, Unfallversicherung, Hausratversicherung, Haftpflichtversicherung, Risiko-Lebensversicherung, Berufsunfähigkeit, Rentenversicherung, Tierhalterhaftpflicht
D4 (34)	Gesundheitsberatung, Gesundheitswesen, Ernährung, Formuladiäten, Blut, Puls, Zuckermessgeräte, Hörgeräte, Sehhilfen, Mundpflege, Atemwege, Akupunktur, Reizstrom, Bandagen, Pflaster, Kissen, Matratzenauflagen, Fieberthermometer, Licht, Wärme, Bücher, Fit, Schlank, Gewichtsreduktion, Fitness, Sport, Waagen, Fettanalysen, Pflege, WC-Hilfen, Inkontinenz-Hilfen, Körperpflege
D5 (36)	Hamburg, Container-Hotels, Garni-Hotels, Hotelbäder, Hotelbedarf, Hotelbetten, Hotelbusse, München, Hoteldrucksachen, Hotelgeräte, Hotelgeschirr, Hotelglas, Frankfurt, Hotelhygieneprodukte, Hotelkerzen, Hotelkleiderbügel, Hotelmarketing, Hotelpensionen, Hotelporzellan, Hotelreservierung, Hotelschiffe, Kaiserslautern, Hotel-TV-Systeme, Hotel-Video-Systeme, Hotelwäsche, Hotelwäsche-Leasing, Kurhotels, Landhotels, Luxushotels, Düsseldorf, Messehotels, Sporthotels, Wellness-Hotels, Berlin, Stuttgart, Dresden

Tabelle 9.4: **Branchenspezifische Wörter als Startwörter für "Test 1"**

Branche	Startwörter für "Test 1"
D6 (29)	Immobilienwirtschaft, Handelsvermittlung, gewerbliche, landwirtschaftliche, private, Immobilienfinanzierung, Immobilienmarketing, Immobilienverkauf, Immobilienvermietung, Immobilienverpachtung, Immobilienverwaltung, Immobilienverwertung, Leasing, Wohnung, Büro, Immobilienmakler, Vermietung, Mieten, WGs, Kaufen, Versteigerungen, Wohnanlagen, Projekte, Bauen, Grundstücke, Typenhäuser, Fertighausanbieter
D7 (29)	München, Finanzberatung, Unternehmen, Finanzdienstleistungen, Finanzplanung, Rechnung, Kredit, Steuern, Zoll, Dresden, BMF, Aktien, Anleihen, Genussscheine, Investmentfonds, Spezialitätenfonds, Riester-Rente, Zertifikate, Marktanalysen, Business, Stuttgart, Börse, Branchen, Gewerbe, Buchhaltung, E-Commerce, Währungsrechner, Management, Organisation, Steuererklärung, Berlin
D8 (21)	München, Gastronomie, Betriebsrestaurants, Deutsche, Tresore, Vegetarische, Berlin, Frankfurt/Main, Hamburg, Düsseldorf, Nürnberg, Dresden, Garmisch-Partenkirchen, Stuttgart, Rothenburg, Heidelberg, Baden-Württemberg, Brandenburg, Bremen, Hansestadt, Hamburg, Hesse
D9 (41)	München, Krankenhaus, Operationskleidung, Ärztebedarf, Kranken, Rettung, Notarztwagen, Deutsche, Software, Tierärzte, Zahnarzt-Formulare, Zahnarztstühle, Allgemeinärzte, Berlin, Augenärzte, Chirurgen, Frauenärzte, Internisten, Kinderärzte, Hamburg, Psychiater, Krankheiten, Medizin, Medikamente, Frankfurt/Main, Düsseldorf, Nürnberg, Dresden, Garmisch-Partenkirchen, Stuttgart, Rothenburg, Heidelberg, Baden-Württemberg, Brandenburg, Bremen, Hansestadt, Hamburg, Hesse
D10 (37)	München, Krankenpflege, Hubbadewannen, Notrufsysteme, Pflegearbeitswagen, Seniorenpflegemittel, Pflege, Berlin, Therapie, Ernährung, Textil, Bekleidung, Küche, Hauswirtschaft, Hamburg, Raumeinrichtungen, Gebäudetechnik, Dienstleistungen, Facility, Management, Kommunikationstechnik, Organisation, Verwaltung, Frankfurt/Main, Düsseldorf, Nürnberg, Dresden, Garmisch-Partenkirchen, Stuttgart, Rothenburg, Heidelberg, Baden-Württemberg, Brandenburg, Bremen, Hansestadt, Hamburg, Hesse

Geographische Namen (z.B. Düsseldorf, Nürnberg, Dresden) sind eigentlich branchenneutral. Sie werden aber als Hilfsmittel für automatisch erstellte branchenspezifische Suchbegriffe gebraucht (z.B. Restaurant Hamburg). Branchenspezifische Suchbegriffe für Suchmaschinen mit den vorgestellten branchenspezifischen Wörtern werden in den jeweiligen E-Commerce-Branchen automatisch erstellt.

9.7 Ergebnis der erstellten Korpora für "Test 1"

Das Ergebnis der erstellten Korpora für E-Commerce-Branchen ist für die Ziele dieser Arbeit sehr geeignet. Die Größe der jeweiligen Korpora ist unterschiedlich. Wenn man ein sehr großes Korpus aufbauen will, sollten mehrere

Suchbegriffe für Suchmaschinen ausreichend berücksichtigt werden. Die automatisch erstellten Suchbegriffe für Suchmaschinen können z.B. mit Hilfe von Großstadtnamen, geographischen Namen und Postleitzahl leicht maschinell erweitert werden (z.B. Büroartikel Locher Tacker Berlin). Die Nutzung kann für die Sammlung der einsprachigen Webseiten sehr geeignet sein.
Lokale Speicherung von Webseiten in den drei Branchen "Lebensmittel", "Gesundheit" und "Altenpflege" wurde nach zwei Wochen Laufzeit manuell abgebrochen, weil bei der lokalen Speicherung Probleme auftraten. Eine Ursache dafür war die Dateigröße. Deshalb wurde im Programm die Dateigröße auf maximal 2 Megabyte beschränkt. Die bisher heruntergeladenen Webseiten in den drei Branchen sind für das Experiment ausreichend.
Die Tabelle 9.5 faßt den grundsätzlichen Aufbau der Korpora übersichtlich zusammen. In der Tabelle werden die folgenden Abkürzungen verwendet:

Abkürzung	Bemerkung
B	Kodierung der Branchen
BS	Anzahl der verwendeten branchenspezifischen Startwörter
BSS	Anzahl der automatisch erstellten Suchbegriffe für Suchmaschinen
URL	Anzahl der verschiedenen URL-Adressen
Web	Anzahl der verschiedenen lokal gespeicherten Webseiten
OG	Original-Größe der Gesamt-Webseiten
KG	Größe des automatisch erstellten Korpus
F	Anzahl der Einwortterme mit Varianten in einem Korpus
FV	Anzahl der Einwortterme ohne Varianten in einem Korpus
BV	Anzahl von automatisch erkanntem branchenspezifischem Vokabular

Tabelle 9.5: **Übersicht des grundsätzlichen Aufbaus der Korpora**

B	BS	BSS	URL	Web	OG	KG	F	FV	BV
P1	14	60	3664	2682	197 M	52 M	442491	362153	5069
P2	35	165	1684	1509	114 M	18 M	241215	198021	4052
P3	18	80	2757	2449	160 M	33 M	411234	337409	3616
P4	23	105	9958	8455	620 M	126 M	940058	753098	2887
P5	21	95	953	788	66 M	14 M	195474	155835	5010
P6	26	120	4468	3593	470 M	90 M	878788	708601	1952
P7	26	120	4606	4020	233 M	56 M	643837	536730	5805
P8	23	105	11110	10079	529 M	99.6 M	664111	528772	5028
P9	35	165	17446	6330	481 M	135 M	1145331	929442	2820
P10	22	100	7361	6477	346 M	47 M	418794	335210	4924

Tabelle 9.5: **Übersicht des grundsätzlichen Aufbaus der Korpora**

B	BS	BSS	URL	Web	OG	KG	F	FV	BV
D1	27	125	11537	10537	527 M	98.4 M	675152	534057	5983
D2	36	170	17592	14466	677 M	149 M	982488	787125	3173
D3	35	165	6128	4708	258 M	65 M	505532	416847	3887
D4	32	150	19657	5633	521 M	145 M	1080936	873649	1188
D5	36	170	11125	9317	602 M	140 M	1046519	835810	1459
D6	27	125	12956	10821	584 M	132 M	1091391	858839	2773
D7	31	145	21092	14148	1322 M	333 M	1955733	1562570	1330
D8	22	100	19142	16884	1131 M	266 M	1512645	1190338	2401
D9	38	180	24910	19446	1650 M	499 M	2386242	1911542	3979
D10	37	175	10639	1477	271 M	76 M	824740	670969	1195

9.8 Erweiterung des Korpus

Die Korpora in den E-Commerce-Branchen werden durch das vorgestellte Masterprogramm für die Methode[b] "Extraktion mit Suchmaschinen (z.B. Google, Yahoo)" automatisch erstellt.
Für die Erweiterung des Korpus wird die Datei für "Secondary Level URLs" zur Methode [a] "Extraktion aus Startseiten" erstellt. Zur Ausführung des Masterprogramms für die Methode [a] braucht man nur eine Datei, die eine oder mehrere Start-URLs beinhaltet.
Die Korpusgröße im Bereich "Büroartikel" nach Ausführung der Methode[b] beträgt ca. 14 Megabyte für die 788 verschiedenen Webseiten, die lokal gespeichert wurden. Zur Erweiterung des Korpus ist das Masterprogramm für die Methode [a] "Extraktion mit Startseiten" nützlich.
Die Methode [a] wird wie folgt durchgeführt:

```
perl BVmasterMethodeA.pl ./startUrls.datei
(Dateiname: './startUrls.datei')
```

Dateien als Resultat pro URL-Adressen lokal speichern:
Durch die Methode [a] "Extraktion mit Startseiten" werden die schon vorgestellten Dateien als Resultat für ein Korpus und Frequenzlisten in den jeweiligen Startseiten pro URL-Adresse automatisch erstellt, indem Webseiten mit Hilfe der internen URL-Adressen erfasst werden. Start-URL-Adressen müssen eindeutig sein. Jede Start-URL-Adresse wird als Verzeichnisname für leichte Markierung wie folgt umgeschrieben, um Startseiten nur jeweils einmal zu

verarbeiten:

Startseite	Verzeichnisname
`http://www.allago.de/?MerchantID=A06`	httpwwwallagodeMerchantIDA06
http://www.officexl.de/	httpwwwofficexlde

9.8.1 Extraktion der internen Links

Die Extraktion der internen Links ist die wichtigste Aufgabe dieser Methode [a] "Extraktion mit Startseiten". Man unterscheidet zwei Arten der Verfolgung von Links:

- **Verfolgung interner Links**:
  Interne Links innerhalb des gesuchten Bereichs werden automatisch weiterverfolgt, um URL-Adressen zu sammeln.
  Z.B. Startseite: http://www.otto-office.com/de
  Interne Links, die als Basis-Adresse (z.B. http://www.otto-office.com/de) wie folgt beinhalten:

  http://www.otto-office.com/de/homepage.obtshop?wkid=OO-1-DE-7-464da3127240a
  http://www.otto-office.com/de/mz/merkzettel.obtshop?wkid=OO-1-DE-7-464da3127240a
  http://www.otto-office.com/de/kontakt/index.obtshop?wkid=OO-1-DE-7-464da3127240a
  http://www.otto-office.com/de/agbs.obtshop?wkid=OO-1-DE-7-464da3127240a
  http://www.otto-office.com/de/wkcard.obtshop?wkid=OO-1-DE-7-464da3127240a

- **Verfolgung externer Links**:
  Externe Links außerhalb des gesuchten Bereichs werden nicht weiterverfolgt. Sie beinhalten keine Basis-Adresse, z.B. http://www.otto-office.com/de. Die URL-Adresse 'http://www.otto-office.com' gehört im Beispielfall zu "externen Links". International bekannte Firmenseiten im Web bieten Kunden die gewünschte Sprachauswahl. Die Firma 'http://www.otto-office.com' bietet vier verschiedene Sprachen wie folgt zur Auswahl:

Links	Sprache
http://www.otto-office.com/de/	Deutsch
http://www.otto-office.fr/	Französisch
http://www.otto-office.com/cz/	Tschechisch
http://www.otto-office.com/sk/	Slowakisch

Für den Aufbau des deutschen Korpus wird die Basis-Adresse (http://www.otto-office.com/de/) für interne Links ausgewählt und eingeschränkt. Dies sollte man auch für die Struktur von Webseiten berücksichtigen, um bestimmte URL-Adressen zu sammeln. Andere URL-Adressen außerhalb des gesuchten Bereichs sind externe Links (z.B. www.bmw.de).

9.9 Vergleich der Frequenzlisten aus Korpora

Die Grundannahme für domainspezifische Terme (DST) aus Kapitel 4.3. gilt auch für branchenspezifisches Vokabular in den jeweiligen E-Commerce-Branchen. Die unterschiedlichen Frequenzlisten für Einwortterme aus den erstellten Korpora werden untereinander verglichen, um automatisch branchenspezifisches Vokabular der jeweiligen Branche zu gewinnen.

9.9.1 Normalisierung der Frequenzen

Mit Hilfe des Linux-Befehls "wc -l ./suchbegriffeFor*/freVariation.list[4]" werden die 20 Frequenzlisten (freVariation.list) aus den Korpora für E-Commerce wie folgt angezeigt. Die folgende Zeilennummer entspricht der Anzahl der verschiedenen Wörter, weil eine Zeile nur ein Wort beinhaltet:

[4] wc - print the number of newlines, words, and bytes in files (man wc)

1911542	./suchbegriffeForAerzte/freVariation.list
670969	./suchbegriffeForAltenpflege/freVariation.list
362153	./suchbegriffeForAutobranche/freVariation.list
787125	./suchbegriffeForBank/freVariation.list
155835	./suchbegriffeForBueroartikel/freVariation.list
528772	./suchbegriffeForComputer/freVariation.list
1562570	./suchbegriffeForFinanzen/freVariation.list
873649	./suchbegriffeForGesundheit/freVariation.list
198021	./suchbegriffeForHaushaltsgeraete/freVariation.list
835810	./suchbegriffeForHotel/freVariation.list
858839	./suchbegriffeForImmobilien/freVariation.list
753098	./suchbegriffeForKleidung/freVariation.list
708601	./suchbegriffeForKosmetik/freVariation.list
929442	./suchbegriffeForLebensmittel/freVariation.list
337409	./suchbegriffeForMoebel/freVariation.list
534057	./suchbegriffeForReisen/freVariation.list
1190338	./suchbegriffeForRestaurant/freVariation.list
335210	./suchbegriffeForSchmuck/freVariation.list
416847	./suchbegriffeForVersicherung/freVariation.list
536730	./suchbegriffeForWein/freVariation.list
14487017	total

Die jeweils häufigste Wortvariante und die zugehörige Häufigkeit aus den o.g. 20 Frequenzlisten (freVariation.list) werden in die jeweiligen neuen Frequenzlisten aufgenommen. Die jeweiligen Worthäufigkeiten werden zum Vergleich mit den neuen Frequenzlisten wie folgt normalisiert:

Normalisierung der Frequenzen
$NF = n_i * 100 / \sum_k n_k$
NF (Normalisierung der Frequenzen) n_i (Häufigkeit eines Wortes in einer Branche) $\sum_k n_k$ (Anzahl aller Wörter der Branche)

Tabelle 9.6: Normalisierung der Frequenzen

Die neue Frequenzliste im Computer-Bereich wird durch die o.g. Normalisierung der Frequenzen wie folgt automatisch erstellt und nach Worthäufigkeiten sortiert:

Solche domainneutralen Wörter (z.B. free, site, Internet) können nach dem Vergleich mit den anderen Frequenzlisten als "nicht branchenspezifisch" identifiziert werden. Das bedeutet, dass z.B. die Worthäufigkeit von "Internet" oder "free" im Computer-Bereich niedriger als oder fast gleich wie in anderen Bereichen sein kann.

$17.8449312747271 = 94359 * 100/528772$ (Computer)
$8.8028488649172 = 46547 * 100/528772$ (laptop)
$7.69689015303382 = 40699 * 100/528772$ (Dell)
17.8449312747271 Computer
8.8028488649172 laptop
7.69689015303382 Dell
5.98083862231737 Software
5.23855272215624 Notebook
4.69673129439532 free
4.57872202007671 site
4.38941547585727 Internet
4.09666169918226 Web
4.02687736869577 online
3.92607778021529 Memory
...

Tabelle 9.7: Normalisierung der Frequenzen im Computer-Bereich

9.9.2 Berechnung der Worthäufigkeit in branchenspezifischem Vokabular

Solche Wörter (z.B. Computer, Dell) sind spezifisch im Computer-Bereich und können mit den unterschiedlichen Worthäufigkeiten in allen o.g. Bereichen vorkommen. Die unterschiedlichen Häufigkeiten eines Wortes aus den Korpora werden verglichen, um branchenspezifisches Vokabular der jeweiligen Branche zu extrahieren.
Unter der oben erwähnten Grundannahme werden für "Branchenspezifisches Vokabular" mindestens die folgenden zwei Bedingungen [a.] und [b.] schrittweise ausgeführt:

a. **Die Worthäufigkeit des Terms T in einem Bereich** ist höher als in anderen Bereichen.

b. **Die Worthäufigkeit des Terms T in anderen Bereichen** ist niedriger.

Die Berechnung von [a.] ist einfach, weil alle getroffenen Worthäufigkeiten eines Terms nach dem Vergleich der normalisierten Frequenzen in allen Branchen ermittelt werden. Nach dem Vergleich der Frequenzlisten aus den erstell-

17.8449312747271 Computer [D9:0.767286306029373 D2:1.05040495474035 D5:0.67742668788361 D6:0.689419087861636 P4:0.883019208655447 D1:0.7375617209399 D8:0.678042707197452 P10:1.09334447063035 D3:0.443088231413444 P5:1.29816793403279 P2:1.11149827543543 P1:0.582350553495346 P6:1.14831901168641 P7:0.169917835783355 D7:1.28000665570182 P3:0.474498309173733 P9:0.450377753533841 D4:0.528244180443176 D10:0.524316324599199]
7.69689015303382 Dell [D9:0.0201408077876395 D2:0.0886771478481817 D5:0.0624543855660976 D6:0.0102463907670704 P4:0.056831912978125 D1:0.0307083326311611 D8:0.0435170514593334 P10:0.0184958682616867 D3:0.0218305517372081 P5:0.0500529406102609 P2:0.0393897616919418 P1:0.054120772159833 P6:0.142816620354755 P7:0.020680789223632 D7:0.0495977780195447 P3:0.0151151866132794 P9:0.0280813649480011 D4:0.0177416788664555 D10:0.0296585982362822]
5.98083862231737 Software [D9:1.0693461090575 D2:1.6997300301731 D5:1.17562603941087 D6:0.929976398370358 P4:0.876512751328512 D1:0.898780467253495 D8:0.866392570849624 P10:0.86483100146177 D3:0.792616955381711 P5:1.13774184233324 P2:0.977674085071785 P1:0.472176124455686 P6:1.18063621135166 P7:0.188176550593408 D7:2.28751351939433 P3:0.29400519843869 P9:0.442846352973074 D4:0.501002118699844 D10:0.914796361679899]
4.09666169918226 Web [D9:0.832050773668588 D2:1.96245831348261 D5:1.43334011318362 D6:1.05596043030184 P4:0.908513898589559 D1:1.81497480606003 D8:1.56442959898785 P10:0.933444706303511 D3:0.8420355670066 P5:0.479353161998267 P2:0.257043444887158 P1:0.703293911689258 P6:1.33742402282808 P7:0.403927486818326 D7:1.26778320331249 P3:0.32512470028956 P9:0.360646495424136 D4:0.580667979932444 D10:0.558893182844513]

Tabelle 9.8: Vergleich der normalisierten Frequenzen in allen Branchen

ten Korpora wurde eine entsprechende Datei (z.B. 'freVariation_ComputerAbgleich.list' im Computer-Bereich) in der jeweiligen Branche wie in Tabelle 9.8 erstellt. Vor dem Term steht die jeweilige Worthäufigkeit in einem Beispiel-Bereich "Computer". Nach dem Term stehen alle Branchen mit Treffern (z.B. D9: Lebensmittel, D2: Haushaltsgeräte) und Worthäufigkeiten des Terms in eckigen Klammern.

Die unterschiedlichen Worthäufigkeiten von "Computer" und "Web" in den jeweiligen Branchen (aus der oben erstellten Tabelle 9.8) werden in der grafischen Abbildung 9.2 dargestellt. Nach dem Vergleich der normalisierten Frequenzen in allen Branchen wird das Wort "Computer" bereichsspezifischer als

"Web" im Computer-Bereich (P8) identifiziert.

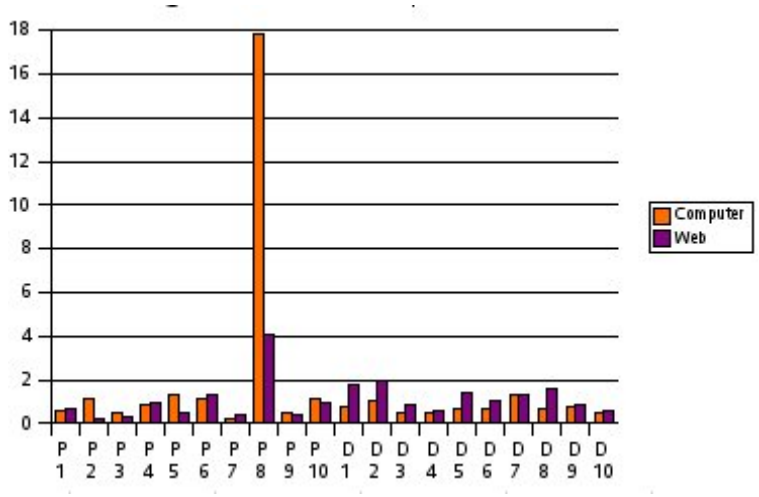


Abbildung 9.2: Vergleich der Worthäufigkeiten von Computer und Web

Im zweiten Schritt wird berechnet, dass die Worthäufigkeit eines Terms in anderen Bereichen niedriger ist. Die Worthäufigkeiten der oben genannten vier Terme - Computer, Dell, Software, Web - ist höher als in anderen Bereichen. Die beiden Terme "Computer" und "Dell" sind seltener als "Software" und "Web" in anderen Bereichen vorgekommen. Mit Hilfe des Vergleichs der Frequenzlisten aus erstellten Korpora können "Computer" und "Dell" als branchenspezifisch und "Software" und "Web" als branchenneutral erkannt werden. "Software" und "Web" können als Schlüsselwörter (keywords) im Computer-Bereich betrachtet werden. Aber sie können als branchenneutral nach der Berechnung von [a.] und [b.] im Experiment maschinell identifiziert werden. Die Berechnung von [b.] gibt den Abstand zwischen höheren und niedrigeren Frequenzen eines Terms an. Die Formel zur Berechnung von [b.] ist die folgende:

Die Datei ('freVariation_ComputerAbgleich.list') für den Beispiel-Bereich "Computer" beinhaltet 528772 Einträge von allen getroffenen normalisierten Frequenzen in allen Branchen und ist nach Frequenzen sortiert. Vom 1. bis zum 23. Term werden die jeweiligen Abstandswerte nach der Berechnung von [b.] in der folgenden Tabelle 9.11 angezeigt.

AVERAGE DEVIATION - Mittelwert der Abweichung (Abstandswert)

HF (High Frequency), LF (Low Frequency), Dev (Deviation)

$$Normalisierungsfaktor = 1/HF$$
$$NormalisierteLF_i = Normalisierungsfaktor * LF_i$$
$$NormalisierteDev_i = absolute\_value(1 - NormalisierteLF_i)$$

$$\text{AVERAGE DEVIATION} = \frac{\sum_{i=1}^{n} NormalisierteDev_i}{n}$$

n = Anzahl der getroffenen Branchen

Tabelle 9.9: AVERAGE DEVIATION - Mittelwert der Abweichung (Abstandswert)

In Perl wird es wie folgt berechnet

$\$normalisierungsfaktor = 1/\$highFrequence;$
$\$normalisierteLF = \$normalisierungsfaktor * \$lowFrequence;$
$\$normalisierteDev = abs(1 - \$normalisierteLF);$
$\$sumOfnormalisierteDev+ = \$normalisierteDev;$

Nach dem Addieren von '$*normalisierteDev*'

$\$averageDeviation = \$sumOfnormalisierteDev/\$cBranchen;$

Tabelle 9.10: AVERAGE DEVIATION - Mittelwert der Abweichung in Perl

Es folgt:

Je höher der Abstandswert eines Terms ist, desto branchenspezifischer ist dieser.

Für die Gewinnung von branchenspezifischem Vokabular kann man einen bestimmten Abstandswert festlegen. Im Experiment wurde der Abstandswert '0.90' für qualifiziertes branchenspezifisches Vokabular eingesetzt.

Die Terme, die klein geschrieben sind und deren Abstandswert höher als '0.90' ist (z.B. laptop, computers, battery, use), werden in einer anderen Datei

Worthäufigkeit	Term	Abstandswert
17.8449312747271	Computer	0.956976459900244
8.8028488649172	laptop	0.988949030632629
7.69689015303382	Dell	0.994528494633606
5.98083862231737	Software	0.845380065116664
5.23855272215624	Notebook	0.981119497343
4.38941547585727	Internet	0.626627945300166
4.09666169918226	Web	0.773610716910695
3.92607778021529	Memory	0.984949859779125
3.83530141535482	computers	0.983599906468475
3.82319790004009	Windows	0.916453016100691
3.80863585817706	battery	0.992539600546648
3.56051379422511	information	0.746053746764369
3.48146270982578	system	0.87773727633439
3.15258750463338	Business	0.72548324734255
3.13518870136845	News	0.720857074139313
2.98067976367886	use	0.908511628643349
2.9695218354981	Samsung	0.971394511519274
2.96895448321772	Email	0.477948211351884
2.91221925517993	Service	0.585042277295797
2.77435265104809	Hardware	0.941881939678878
2.5684037732709	http	0.508988260497845
2.53795586755728	data	0.922582271427747
2.49937591249158	Microsoft	0.925795615734004

Tabelle 9.11: Berechnung von [b.] im Beispiel-Bereich “Computer”

in den jeweiligen Branchen getrennt gespeichert, um daraus wieder branchenspezifisches Vokabular zu erkennen. Die Terme (z.B. Computer, Dell), die groß geschrieben sind und deren Abstandswert höher als ’0.90’ ist, werden im Experiment als branchenspezifisch betrachtet.

In der vorherigen Tabelle 9.5 (Übersicht des grundsätzlichen Aufbaus der Korpora) wird die Anzahl von erkanntem branchenspezifischem Vokabular mit dem Abstandswert “höher als 0.90 (branchenspezifisch)” in der letzten Spalte “BV” angezeigt.

Die oben genannten Terme aus Tabelle 9.11 werden durch den Abstandswert ’0.90’ wie folgt maschinell erkannt:

- **höher als '0.90' (branchenspezifisch)**:
  Computer, Dell, Notebook, Memory, Windows, Samsung, Hardware, Microsoft
- **niedriger als '0.90'**:
  Software, Internet, Web, information, system, Business, News, Email, Service, http, data

9.9.3 Ergebnis des Vergleichs der Frequenzlisten für Test 1

Groß- und Kleinschreibung eines Wortes hat bei Suchmaschinen im WWW i.a. keine große Bedeutung. Aber sie spielt eine große Rolle bei deutschen Texten im WWW, weil Substantive (Nomen) im Deutschen groß geschrieben werden. Heute tendiert man dazu (z.B. im Deutschen und Englischen), Schlüsselwörter im Web zur Verdeutlichung groß zu schreiben. Die erstellten Korpora aus dem Web werden danach "case sensitive" lokal gespeichert.

Nach dem Vergleich der Frequenzlisten (Berechnung von [a.] nach [b.]) werden die folgenden acht Dateien in den jeweiligen Bereichen automatisch erstellt. Im Beispiel-Bereich "Computer" sehen sie wie folgt aus:

Datei	**Bemerkung**
`Computer_BNeutral_*`	"Branchenneutrale Stoppwörter"
`Computer_BVGapKleinschreibung_*`	Kleingeschriebene Wörter, deren Abstandswert höher als '0.90' ist.
`Computer_BVGap_*`	Großgeschriebene Wörter, deren Abstandswert höher als '0.90' ist, werden als "branchenspezifisch" optimal erkannt.
`Computer_BV_*`	Der Abstandswert ist höher als '0.78'. Für mehr branchenspezifisches Vokabular mit Großschreibung kann man den Abstandswert niedriger einstellen.
`Computer_Birrelevant_*`	weitere Stoppwörter und branchenneutrale Wörter
`Computer_kleinschreibung_*`	Kleingeschriebene Wörter, die als branchenneutral betrachtet werden können.
`Computer_onlyB_*`	Wörter, die nur in einem bestimmten Bereich vorkommen.
`freVariation_ComputerAbgleich.list`	Input-Datei für den Vergleich der Frequenzen

Für die zweite Berechnung [b.] zur Gewinnung von branchenspezifischem Vokabular ist die Einstellung von 'Abstandswert' sehr wichtig. Der Abstandswert höher als '0.90' für optimales branchenspezifisches Vokabular ist eindeutig genug zur Unterscheidung in den erstellten Frequenzlisten (z.B. die o.g. Datei: Computer_BVGap_*). Der Abstandswert höher als '0.78' ist auch

nützlich für weiteres branchenspezifisches Vokabular. Die Einstellung des Abstandswerts kann im Programm beliebig geändert werden. Die folgende Tabelle 9.12 zeigt den Unterschied zwischen den Abstandswerten '0.90' und '0.78'. Die Anzahl steht für das erkannte branchenspezifische Vokabular.

Branche	**höher als '0.90'**	**höher als '0.78'**
P1 / Autobranche	5069	15346
P2 / Haushaltsgeräte	4052	16249
P3 / Möbel	3616	10687
P4 / Kleidung	2887	10202
P5 / Büroartikel	5010	21998
P6 / Kosmetik	1952	6260
P7 / Wein	5805	14935
P8 / Computer	5028	14616
P9 / Lebensmittel	2820	10751
P10 / Schmuck	4924	14196
D1 / Reisen	5983	18107
D2 / Bank	3173	10458
D3 / Versicherung	3887	16046
D4 / Gesundheit	1188	6300
D5 / Hotel	1459	7367
D6 / Immobilien	2773	10673
D7 / Finanzen	1330	9861
D8 / Restaurant	2401	13766
D9 / Ärzte	3979	21805
D10 / Altenpflege	1195	6967

Tabelle 9.12: Unterschied zwischen den Abstandswerten '0.90' und '0.78'

Schlußbemerkung

Durch Schlüsselwortextraktion (engl. keyword extraction) können Schlüsselwörter automatisch identifiziert werden. Dabei handelt es sich um die Berechnung der Worthäufigkeit. Aber nicht alle Schlüsselwörter sind branchenspezifisch.

Branchenspezifische Wörter der jeweiligen Branche werden durch den Vergleich der verschiedenen Frequenzlisten aus Korpora durch die Berechnung des Abstandswerts automatisch gewonnen.

9.10 Neue Startwörter für Test 2

9.10.1 Erstellung der Startwörter

Um ein neues Korpus aufzubauen, wird das vorgestellte Masterprogramm mit den zwei Parametern, nämlich "ein Dateiname" und "ein Basis-Wort" ausgeführt. Durch die branchenspezifischen Startwörter für "Test 1" wurden die besten branchenspezifischen Wörter, deren Abstandswert höher als '0.90' ist, in den jeweiligen E-Commerce-Branchen automatisch erkannt. In der Tabelle 9.12 wird die Anzahl der besten erkannten branchenspezifischen Wörter in der zweiten Spalte "höher als 0.90" angezeigt.

Um neue branchenspezifische Startwörter für "Test 2" zu erstellen, werden diese besten branchenspezifischen Wörter aus "Test 1" benutzt. Die neuen Startwörter für "Test 2" müssen als Startwörter für "Test 1" nicht verwendet werden. Dafür wird ein Programm mit den drei folgenden Parametern, nämlich Datei 1, Datei 2, Anzahl benötigt:

- **Datei 1**, die Startwörter für "Test 1" enthält, einlesen.
- **Datei 2**, die die jeweiligen branchenspezifischen Wörter, deren Abstandwert höher als '0.90' ist, einlesen, um neue Startwörter für Test 2 nach dem Vergleich mit den Startwörtern für "Test 1" automatisch zu erstellen.
- **Anzahl** der gewünschten Kandidaten der neuen branchenspezifischen Startwörter für Test 2

Dafür wurde das von mir in Perl erstellte Programm mit der Anzahl '100' in den jeweiligen Branchen wie folgt ausgeführt. Aus den besten 100 automatisch erstellten Kandidaten werden neue Startwörter für "Test 2" manuell ausgewählt:

```
test2forSuchbegriffe.pl ./Altenpflege ../Altenpflege_BVGap_* 100
```
Datei 1: Altenpflege, **Datei 2**: Altenpflege_BVGap_*, **Anzahl**: 100

Die neuen Startwörter für "Test 2" werden zum Vergleich der Ergebnisse von "Test 1" und "Test 2" verwendet. Es sind die folgenden:

Tabelle 9.13: **Branchenspezifische Wörter als Startwörter für "Test 2"**

Branche	Startwörter für "Test 2"
P1 (14)	Audi, Ford, Opel, Renault, Toyota, Alfa, Volvo, Fiat, Peugeot, Seat, Kia, Nissan, Rover, Mazda
P2 (35)	Staubsauger, AEG, Philips, Waschmaschine, Küchengeräte, Geschirrspüler, Klimaanlagen, Kenwood, Krups, Electrolux, Dampfbügelstation, Klimagerät, Toaster, Tefal, Bauknecht, Kleingeräte, Klimatechnik, Kühlschränke, Kärcher, CCD, Wasserkocher, Küchenmaschine, Mikrowellen, Speichertakt, Einbauherd, Kaffeemaschinen, Herde, Entsafter, Dampfdruck, Espressomaschinen, Nähmaschine, Ventilator, Backöfen, Spülmaschine, Brotbackautomat
P3 (18)	Bett, Tisch, Stühle, Sofa, Tische, Holzmöbel, Schränke, Gartenmöbel, IKEA, Gartenwölfe, Kinderzimmer, Schreibtisch, Esszimmer, Polstermöbel, Esstisch, Couchtisch, Massivholzmöbel, Kindermöbel
P4 (23)	Bekleidung, Shirts, Damenmode, Dessous, Unterwäsche, Hose, Jacke, Kleid, Wäsche, Herrenmode, Baumwolle, Blazer, Hemd, Socken, Klamotten, Shorts, Anzüge, Sportswear, Kindermode, Bluse, Anzug, Handschuhe, Berufsbekleidung
P5 (21)	Bürobedarf, Papier, Büromöbel, Schreibwaren, Werbeartikel, Druckerpatronen, Büromaterial, Schulbedarf, Kopierer, Etiketten, Bürotechnik, Schreibgeräte, Visitenkarten, Laserdrucker, Büromaschinen, Leitz, Folien, Faxgeräte, Patronen, Briefpapier, Kopierpapier
P6 (26)	Aloe, Aveda, Nail, Gel, Spray, Naturkosmetik, Parfum, Lotion, Shampoo, Toilette, Gesichtspflege, Fußpflege, Dior, Logona, Fecha, Pinsel, Kelompok, Nägel, Hauschka, Shave, Shadow, Duschgel, Fingernägel, Frisur, Maniküre, Blush
P7 (26)	Weingut, Riesling, Sauvignon, Pinot, Blanc, Cabernet, Bordeaux, Merlot, Weinbau, Champagne, Kabinett, Rebsorten, Auslese, Weinhandel, Weinshop, Gris, Qualitätswein, Bianco, Rioja, Loire, Reben, Rheingau, Chianti, Weinberg, Weinkeller, Eiswein
P8 (23)	Memory, Windows, Microsoft, Sony, Apple, ThinkPad, USB, Intel, Desktop, Toshiba, LCD, Compaq, Monitor, Mac, RAM, Adapter, Mouse, Pentium, AMD, CPU, Monitors, Printers, WLAN
P9 (35)	Gemüse, Fleisch, Obst, Salz, Eier, Nahrung, Salat, Olivenöl, Pfeffer, Kalorien, Früchte, Nudeln, Mehl, Eiweiß, Wurst, Gentechnik, Naturkost, Joghurt, Teig, Sahne, Backwaren, Brötchen, Süßwaren, Bohnen, Paprika, Fettsäuren, Suppe, Geflügel, Zimt, Zwiebeln, Süßigkeiten, Weizen, Speichererweiterung, Soja, Salate
P10 (22)	Silber, Gold, Ring, Armband, Kette, Jewelry, Swarovski, QVC, Juwelier, Diamanten, Edelstahl, Ohrschmuck, Tattoo, Silver, Goldschmiede, Mineralien, Titan, Schmuckstücke, Weißgold, Edelstein, Perle, Kristall
D1 (29)	Urlaub, Flüge, Mallorca, Yachtcharter, Ferien, Türkei, Kroatien, Spanien, Ferienwohnung, Ferienhaus, Reisebüro, Griechenland, Ostsee, Reiseveranstalter, Tourismus, Unterkunft, Ägypten, Buchung, Karibik, Australien, Tunesien, Inseln, Kanaren, Billigflüge, Asien, Familienurlaub, Reisebüros, Frühbucher, Rundreisen
D2 (38)	Visa, Finance, Zinsen, Sparkasse, Volksbank, Geldanlage, PayPal, Citibank, Zins, Dresdner, Bargeld, Fund, HSBC, IBAN, Rewards, Überweisung, Transaktionen, Volksbanken, Direktbank, DKB, Guthaben, Landesbank, Geldautomaten, Sparbuch, Raiffeisenbank, TAN, Hypo, Bankleitzahl, Banker, Bausparkasse, Internetbanking, Dispo, HypoVereinsbank, Direktbanken, Eypo, Treasury, Tilgung, Kreissparkasse
D3 (34)	Krankenversicherung, Versicherungsvergleich, Lebensversicherung, Krankenkassen, Rente, Altersvorsorge, Autoversicherung, Haftpflicht, Krankenkasse, Vorsorge, Rechtsschutzversicherung, Unfall, Versicherer, Risikolebensversicherung, Versicherten, Versicherungsschutz, Kranken, PKV, Allianz, Absicherung, Gebäudeversicherung, Krankenzusatzversicherung, Versicherungsnehmer, Versicherungsunternehmen, Privathaftpflicht, Sozialversicherung, Altersversorgung, Versicherungswirtschaft, Zusatzversicherung, Direktversicherung, Versicherungssumme, Kapitallebensversicherung, Rechtschutz, Versicherungsgesellschaften

Tabelle 9.13: **Branchenspezifische Wörter als Startwörter für "Test 2"**

Branche	Startwörter für "Test 2"
D4 (34)	DKK, Tomaten, Übergewicht, Diäten, Immunsystem, Gesundheitsförderung, Kursort, Vitalität, Strahlenbelastung, Schwankschwindel, Lahnstein, Gewaltmusik, Magnetfeldtherapie, Muskulatur, Trampolin, Raucherentwöhnung, Glukosamin, Lebensenergie, Gewichtskontrolle, Krafttraining, Gewichtsreduzierung, Morna, Beweglichkeit, Gewichtsabnahme, Fußreflexzonenmassage, Bindegewebe, Dehnübungen, Fatburner, Halsschmerzen, Körperarbeit, Dehnung, Rückenschule, Körperfett, Gesundheitspolitk
D5 (36)	Airport, Inn, Dusseldorf, Resorts, Stadthotel, Wellnesshotels, Kempinski, Brussels, InterContinental, Hotelkritiken, Sheraton, Israelis, Americans, Hyatt, Chaska, Alster, Italie, Lisbon, Sporthotel, Hotelführer, Stadthotels, Charleston, Ferienhotel, Romantikhotel, Hotelverzeichnis, Businesshotels, Feuerwerkskörper, CNA, Starlight, Palestinian, Dolomites, Savannah, Arlington, Verdon, Tagungshotels, Flughafenhotel
D6 (29)	Häuser, Grundstück, Miete, Makler, Gewerbeimmobilien, Einfamilienhaus, Hausverwaltung, Bauträger, Immobilienmarkt, Fertighaus, Bauregister, Immobilienangebote, Immo, Wohnfläche, Einfamilienhäuser, Mietvertrag, Reihenhaus, Mehrfamilienhaus, Immobilienkauf, Immobiliensuche, Wohnimmobilien, Auslandsimmobilien, Wohnbau, GRUNDSTÜCKSWESEN, Hochbau, Baugrundstück, Immobilienbewertung, Immobilienportal, Grundstücks
D7 (29)	Bundesministerium, Steuerberatung, UStG, Anwaltskanzlei, Steuerkanzlei, WKN, Einkommensteuer, Wirtschaftsprüfung, Optionsscheine, Jahresabschluss, Vorsteuerabzug, Steuerpflichtigen, Betriebswirtschaftslehre, Finanzverwaltung, Vertragsparteien, BFH, Finanzministerium, Kapitalerhöhung, Steuerbefreiung, Auftragsbearbeitung, Marktberichte, Außenwirtschaft, WpHG, Finanzmärkte, EBIT, Steuereinnahmen, Finanzpolitik, Finanzplan, Fondsmanager
D8 (21)	Gasthof, Ibis, Parkhotel, Steigenberger, Vegetarian, Bistro, Cathedral, Tulip, Stralsund, Garmisch, Minibar, Biergarten, Oktoberfest, Nurnberg, Golfclub, Germans, Ostseebad, Minotel, Mövenpick, Kantine, Banquet
D9 (41)	Patienten, Behandlung, Klinik, Therapie, Chirurgie, Kliniken, Psychotherapie, Innere, Facharzt, Psychiatrie, Homöopathie, Erkrankungen, Ärzten, Patient, Klinikum, Dosieraerosol, Diagnostik, Klinische, Cor, Insulin, Erkrankung, Seitensprung, Symptome, Neurologie, Baierbrunn, HCT, Zahnärzte, Orthopädie, Operationen, Ärztin, Aspirin, Fachärzte, Plastische, Störungen, Brustkrebs, Kassenärztliche, Therapeuten, Medikament, Ärztekammer, Selbsthilfegruppen, INTERNIST
D10 (37)	Altenheim, Ammersee, ALP, Detektivbüro, Facharbeiter, Altenhilfe, Altenheime, Seniorenzentrum, Seniorenheim, Standardeinband, Volksschule, Religionsunterricht, Fachkraft, Lehrplan, Geschäftsgang, Jahrgangsstufe, Politikberatung, Dernier, Literaturangaben, Teilnehmerkreis, Compostelle, Föderalismus, Dittmann, Verfassungsrecht, Schulstraße, Berufsfachschule, Fachwirt, Nachdr, HWO, Reformpolitik, Sozialstation, Feststehender, Seniorenresidenz, Föderalismusreform, Gesetzgebungskompeten, Jahrgangsstufen, Caritasverband

Bei den automatisch erkannten 100 Kandidaten, deren Abstandswert höher als '0.90' ist, wurden in der Autobranche (P1) überwiegend Automarken verwendet. Deshalb wurden sie als Startwörter in der o.g. Tabelle erstellt. Das Wort "Tomaten" ist das dritthäufigste Wort im Bereich "Gesundheit" (D4) in der Datei ('`Gesundheit_BVGap_*`') - Abstandswert höher als '0.90' - siehe unten. Die Worthäufigkeit steht vor dem Term. Nach dem Term steht der

Abstandswert innerhalb der eckigen Klammer:

```
0.328049365362978 Akupunktur [0.901688062177247]
0.254450013678262 DKK [0.90062052566918]
0.208893960846976 Tomaten [0.913212528517816]
0.198935728192901 Übergewicht [0.90068248377644]
0.16745855601048 Diäten [0.934329131275699]
...
```

Das bedeutet, dass das Wort "Tomaten" im Bereich "Gesundheit" (D4) sehr häufig verwendet wird. In der o.g. Tabelle gibt es deshalb das Wort 'Tomaten', um interessante Webseiten für die Gesundheit mit dem Startwort "Tomaten" zu finden.

9.10.2 Wortgruppen der Startwörter für Test 1 und Test 2

Um mit Hilfe von Suchmaschinen (z.B. Google, Yahoo) verschiedene branchenspezifische Webseiten als Korpus zu sammeln, wurden die branchenspezifischen Startwörter für "Test 1" und "Test 2" verwendet. Sie sind in folgende Wortgruppen eingeteilt:

a. **Marken (bzw. Firmennamen)**: *BMW, Audi, Porsche, Mercedes*

b. **Generische Terme (EGT und KGT)**: *Hosen, Lederhosen*

c. **branchenspezifische Abkürzungen mit ähnlichen Eigenschaften wie EGT**: *KFZ, PKW, LKW, ADAC*

d. **Geographische Namen als Hilfsmittel** für Suchbegriffe: *Hamburg, München, Berlin, Stuttgart, Dresden*

e. **Branchenbezeichner**: *Damenmode, Herrenmode, Büromaterial, Schulbedarf*

Die drei Wortgruppen von [a], [b] und [c] sind sehr wichtig für die schon vorgestellten Zielsetzungen des Gesamtexperiments. Sie können automatisch erkannt, manuell verbessert und bei der semantischen Annotation unterschiedlich gekennzeichnet werden.

9.10.3 Startbedingungen für Test 2

Die Startwörter für "Test 2" werden under den folgenden Startbedingungen verwendet, um neue Korpora in den 20 ausgewählten E-Commerce-Branchen automatisch aufzubauen:

- **Die Anzahl** der Startwörter für Test 1 und Test 2 ist gleich.
- **Die verwendeten Suchmaschinen** (Google und Yahoo) sind gleich.
- **Die verwendeten Basiswörter** für alle Suchbegriffe in den jeweiligen E-Commerce-Branchen sind gleich.

Das vorgestellte Masterprogramm für den Aufbau der Korpora wird für die 20 ausgewählten Bereiche 20 mal verwendet. Dafür wird ein Shell-Skript mit Hilfe des Linux/Unix-Befehls "nohup[5]" wie folgt geschrieben, um in der Shell das Masterprogramm 20 mal parallel zu starten:

Tabelle 9.14: **paralleles Starten mit denselben Basiswörtern für Test 1 und Test 2**

```
#!/bin/sh
# The Bourne shell, or sh, was the default Unix shell of Unix Version 7

nohup BV_masterKorpus.pl ./suchbegriffe/Autobranche Autobranche &
nohup BV_masterKorpus.pl ./suchbegriffe/Haushaltsgeraete Haushaltsgeräte &
nohup BV_masterKorpus.pl ./suchbegriffe/Moebel Möbel &
nohup BV_masterKorpus.pl ./suchbegriffe/Kleidung Kleidung &
nohup BV_masterKorpus.pl ./suchbegriffe/Bueroartikel Büroartikel &
nohup BV_masterKorpus.pl ./suchbegriffe/Kosmetik Kosmetik  &
nohup BV_masterKorpus.pl ./suchbegriffe/Wein Wein &
nohup BV_masterKorpus.pl ./suchbegriffe/Computer Computer &
nohup BV_masterKorpus.pl ./suchbegriffe/Lebensmittel Lebensmittel &
nohup BV_masterKorpus.pl ./suchbegriffe/Schmuck Schmuck &
nohup BV_masterKorpus.pl ./suchbegriffe/Reisen Reisen &
nohup BV_masterKorpus.pl ./suchbegriffe/Bank Bank &
nohup BV_masterKorpus.pl ./suchbegriffe/Versicherung Versicherung &
nohup BV_masterKorpus.pl ./suchbegriffe/Gesundheit Gesundheit &
nohup BV_masterKorpus.pl ./suchbegriffe/Hotel Hotel &
nohup BV_masterKorpus.pl ./suchbegriffe/Immobilien Immobilien &
nohup BV_masterKorpus.pl ./suchbegriffe/Finanzen Finanzen &
nohup BV_masterKorpus.pl ./suchbegriffe/Restaurant Restaurant &
nohup BV_masterKorpus.pl ./suchbegriffe/Aerzte Ärzte &
nohup BV_masterKorpus.pl ./suchbegriffe/Altenpflege Altenpflege &
```

[5]nohup - run a command immune to hangups, with output to a non-tty [man nohup]

9.10.4 Ergebnisse von Test 1 und Test 2

In Tabelle 9.15 werden die Ergebnisse von "Test 1" und "Test 2" für den Aufbau der neuen Korpora in den 20 E-Commerce-Branchen zusammengelegt und verglichen. Die zusätzlichen Abkürzungen in der neuen zweiten Spalte "**T**" sind T1 für "Test 1" und T2 für "Test 2":

Tabelle 9.15: **grundsätzliche Übersicht für Test1 und Test 2**

B	T	BS	BSS	URL	Web	OG	KG	F	FV	BV
P1	T1	14	60	3664	2682	197 M	52 M	442491	362153	5069
	T2	14	60	1501	945	48 M	11 M	121122	100183	5503
P2	T1	35	165	1684	1509	114 M	18 M	241215	198021	4052
	T2	35	165	1101	851	39 M	7.6 M	127086	108572	3900
P3	T1	18	80	2757	2449	160 M	33 M	411234	337409	3616
	T2	18	80	3065	2814	72 M	12 M	160732	132312	9539
P4	T1	23	105	9958	8455	620 M	126 M	940058	753098	2887
	T2	23	105	18582	15110	1295 M	316 M	1874665	1498191	3839
P5	T1	21	95	953	788	66 M	14 M	195474	155835	5010
	T2	21	95	2400	2010	172 M	31 M	402090	325913	3464
P6	T1	26	120	4468	3593	470 M	90 M	878788	708601	1952
	T2	26	120	11909	9784	775 M	149.6 M	1173972	945253	4762
P7	T1	26	120	4606	4020	233 M	56 M	643837	536730	5805
	T2	26	120	15073	12649	697 M	154 M	1178791	949604	9043
P8	T1	23	105	11110	10079	529 M	99.6 M	664111	528772	5028
	T2	23	105	37629	33919	1732 M	288 M	663158	519403	12762
P9	T1	35	165	17446	6330	481 M	135 M	1145331	929442	2820
	T2	35	165	21551	16376	1129 M	326 M	1738178	1426618	6860
P10	T1	22	100	7361	6477	346 M	47 M	418794	335210	4924
	T2	22	100	11916	10212	827 M	172 M	1391082	1116506	3300
D1	T1	27	125	11537	10537	527 M	98.4 M	675152	534057	5983
	T2	27	125	27308	24472	1224 M	230 M	1132610	911345	10334
D2	T1	36	170	17592	14466	677 M	149 M	982488	787125	3173
	T2	36	170	42843	31668	1980 M	519 M	2387663	1905876	2943
D3	T1	35	165	6128	4708	258 M	65 M	505532	416847	3887
	T2	35	165	18058	14057	649 M	160 M	941428	763612	5684
D4	T1	32	150	19657	5633	521 M	145 M	1080936	873649	1188
	T2	32	150	19370	15110	963 M	261 M	1579143	1299417	2128
D5	T1	36	170	11125	9317	602 M	140 M	1046519	835810	1459
	T2	36	170	34416	5796	555 M	118 M	891868	700104	3358
D6	T1	27	125	12956	10821	584 M	132 M	1091391	858839	2773
	T2	27	125	16330	14567	571 M	104 M	727522	571201	7851
D7	T1	31	145	21092	14148	1322 M	333 M	1955733	1562570	1330
	T2	31	145	14056	9065	678 M	183 M	1070813	884434	5103

Tabelle 9.15: **grundsätzliche Übersicht für Test1 und Test 2**

B	T	BS	BSS	URL	Web	OG	KG	F	FV	BV
D8	T1	22	100	19142	16884	1131 M	266 M	1512645	1190338	2401
	T2	22	100	24173	7373	734 M	156 M	980505	781498	3030
D9	T1	38	180	24910	19446	1650 M	499 M	2386242	1911542	3979
	T2	38	180	31686	23298	1515 M	450 M	2047554	1652298	11884
D10	T1	37	175	10639	1477	271 M	76 M	824740	670969	1195
	T2	37	175	4475	3037	319 M	100 M	872561	705746	4462

Die lokale Speicherung von Webseiten der beiden Branchen "Hotel" und "Restaurant" wurde nach neun Tagen Laufzeit manuell abgebrochen. Die bis dahin lokal gespeicherten Webseiten für "Hotel" und "Restaurant" sind ausreichend für den Test. Für die Erstellung der Frequenzlisten werden die bereits vorgestellten Standalone-Programme weiter durchgeführt.
Alle Suchbegriffe von "Autobranche/P1" für "Test 2" wurden aus internationalen Automarken ausgewählt. Deshalb kann die Anzahl der lokal gespeicherten Webseiten wesentlich niedriger als die Anzahl für "Test 1" sein.
Als Suchbegriffe für ein Korpus sollen Firmennamen und branchenspezifische Wörter in den jeweiligen Bereichen angemessen gemischt werden, um mehrere bereichsspezifische URL-Adressen aus Suchmaschinen zu extrahieren.
Die 23 neuen Suchbegriffe von "Computer/P8" für "Test 2" liefern die höchste Anzahl (33919) der lokal gespeicherten Webseiten.
Die bereits vorgestellte Tabelle 9.12 zeigt den Unterschied zwischen den Abstandswerten '0.90' und '0.78'. Die Anzahl steht für das erkannte branchenspezifische Vokabular. Die zusätzliche Bezeichnung 'S' bedeutet die Summe der verschiedenen branchenspezifischen Wörter von "Test 1" und "Test 2" in der dritten Spalte von "höher als 0.90". Die folgende Tabelle zeigt den Unterschied zwischen Abstandswerten für "Test 1" (T1) und "Test 2" (T2):

Tabelle 9.16: **Abstandswerte - '0.90' und '0.78'- für Test 1 und Test 2**

Branche	T	höher als '0.90'	höher als '0.78'
P1 / Autobranche	T1	5069	15346
	T2	5503	9816
	S	9067	
P2 / Haushaltsgeräte	T1	4052	16249
	T2	3900	11280
	S	7461	

Tabelle 9.16: **Abstandswerte - '0.90' und '0.78'- für Test 1 und Test 2**

Branche	**T**	**höher als '0.90'**	**höher als '0.78'**
P3 / Möbel	T1	3616	10687
	T2	9539	17950
	S	11692	
P4 / Kleidung	T1	2887	10202
	T2	3839	13186
	S	5697	
P5 / Büroartikel	T1	5010	21998
	T2	3464	19011
	S	7444	
P6 / Kosmetik	T1	1952	6260
	T2	4762	11183
	S	5818	
P7 / Wein	T1	5805	14935
	T2	9043	19443
	S	11399	
P8 / Computer	T1	5028	14616
	T2	12762	23411
	S	15934	
P9 / Lebensmittel	T1	2820	10751
	T2	6860	22990
	S	8046	
P10 / Schmuck	T1	4924	14196
	T2	3300	8256
	S	6663	
D1 / Reisen	T1	5983	18107
	T2	10334	27406
	S	12900	
D2 / Bank	T1	3173	10458
	T2	2943	9980
	S	5256	
D3 / Versicherung	T1	3887	16046
	T2	5684	17256
	S	7713	
D4 / Gesundheit	T1	1188	6300
	T2	2128	10620
	S	3163	
D5 / Hotel	T1	1459	7367
	T2	3358	15743
	S	4633	
D6 / Immobilien	T1	2773	10673
	T2	7851	26614
	S	9216	
D7 / Finanzen	T1	1330	9861
	T2	5103	22662
	S	6076	
D8 / Restaurant	T1	2401	13766
	T2	3030	15136
	S	4915	

Tabelle 9.16: **Abstandswerte - '0.90' und '0.78'- für Test 1 und Test 2**

Branche	T	höher als '0.90'	höher als '0.78'
D9 / Ärzte	T1	3979	21805
	T2	11884	33824
	S	13689	
D10 / Altenpflege	T1	1195	6967
	T2	4462	18004
	S	5216	

Die folgende graphische Abbildung 9.3 wird aus der dritten Spalte ("höher als 0.90") der Tabelle 9.16 hergeleitet. "S" steht für die Summe der verschiedenen branchenspezifischen Wörter von "Test 1" und "Test 2":

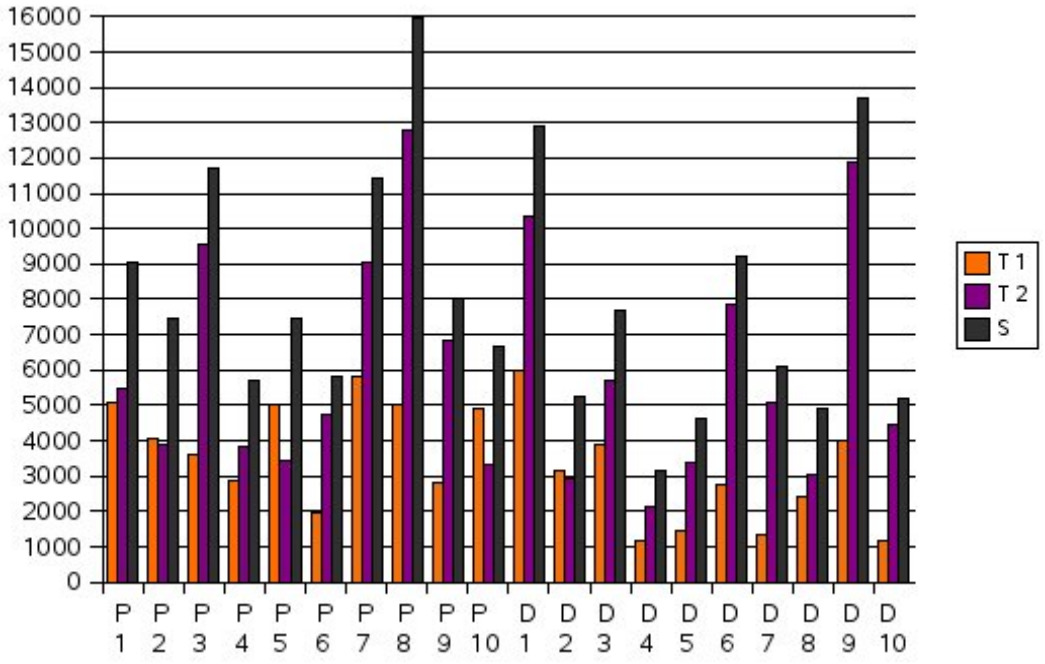


Abbildung 9.3: AGBV aus Test 1 (T1) und Test 2 (T2)

Das Ziel ist die "Automatische Gewinnung von branchenspezifischem Vokabular (AGBV)". Diese graphische Abbildung zeigt, dass die Gesamtzahl von AGBV bei Test 2 wesentlich höher als bei Test 1 ist. Die AGBV bei Test 2 ist effizienter als bei Test 1. Die verwendeten branchenspezifischen Wörter als Startwörter für "Test 2" sind branchenspezifischer als Startwörter für "Test 1", weil sie aus dem Teil von "Test 1" mit der höchsten Frequenz ausgewählt wurden. Startwörter für Suchmaschinen sollten bereichsspezifisch sein und sehr

oft im Web vorkommen, um die AGBV zu verbessern. Die Korpusgröße spielt keine große Rolle bei der AGBV. Im Bereich von "Restaurant" (D8) ist die Korpusgröße von "Test 1" 266 Megabyte. Die Anzahl der "Einwortterme ohne Varianten" ist 1190338. Davon wurden die 2401 branchenspezifischen Wörter, deren Abstandswert höher als '0.90' ist, maschinell identifiziert. Im Bereich "Finanzen" (D7) bei Test 1 ist das Korpus 333 Megabyte groß und enthält 1562570 Wörter. Davon wurden die 1330 branchenspezifischen Wörter nach dem Vergleich der Frequenzlisten automatisch erkannt.
Der Vergleich der verschiedenen Frequenzlisten in den jeweiligen E-Commerce-Branchen spielt eine wichtige Rolle für die AGBV.
Vor dem Testen wurde berücksichtigt, dass ca. 5000 - 10000 branchenspezifische Wörter in den jeweiligen 20 ausgewählten E-Commerce-Branchen automatisch erkannt werden. In den drei Bereichen "Gesundheit" (D4), "Hotel" (D5) und "Restaurant" (D8) ist die Gesamtzahl der branchenspezifischen Wörter niedriger als 5000. Die Anzahl kann je nach Bedürfnis ohne weiteres erhöht werden. In den restlichen 17 Bereichen ist sie viel höher als 5000. Die höchste Anzahl von automatisch erkanntem branchenspezifischem Vokabular ist 15934 im Bereich von "Computer" (P8).

Schlußbemerkung

Das automatisch erkannte branchenspezifische Vokabular in den jeweiligen Branchen ist eine qualifizierte Basis, um manuelle Arbeiten für die linguistische Analyse miteinander zu kombinieren. Im Kapitel 6.8.2 "Wortgruppen der branchenspezifischen Wörter" wurden die drei relevanten Wortgruppen - Marken, EGT und KGT, branchenspezifische Abkürzungen - vorgestellt. Die am häufigsten im Web vorgekommenen Begriffe in den Wortgruppen können durch die 'AGBV' gewonnen werden. Beispielsweise kann eine Liste der besten Firmennamen, die am häufigsten im Web aufgetaucht sind, nach der Worthäufigkeit sortiert und erstellt werden.
Das automatisch erkannte branchenspezifische Vokabular durch die **AGBV** kann als Basis für Grundwortschatz in den jeweiligen E-Commerce-Branchen sehr effizient benutzt werden.

9.11 AGBV aus einer Webseite

"Automatische Gewinnung von branchenspezifischem Vokabular (**AGBV**)" ist das im Rahmen dieser Arbeit erstellte Terminologie-Extraktionssystem. Bis jetzt wurde 'AGBV' aus den 20 erstellten Korpora für E-Commerce auf Korpusebene implementiert. Dafür wurden die automatisch erstellten 20 Datenbanken für den Vergleich der Frequenzlisten in den jeweiligen E-Commerce-Branchen verwendet, um branchenspezifische Wörter maschinell zu erkennen. Durch den Gebrauch von diesen unterschiedlichen Datenbanken für "Test 1" und "Test 2" wird automatische Gewinnung von branchenspezifischem Vokabular (AGBV) aus einer Webseite gleich verwendet. Ergebnisse von Datenbanken für "Test 1" und "Test 2" werden verglichen.

9.11.1 Erstellung des Masterprogramms

Das Masterprogramm für "AGBV aus einer Webseite" wird im Prinzip genauso wie die Programme für den Aufbau der Korpora und den Vergleich der Frequenzlisten konstruiert. Durch die HTML-Analyse (siehe Kapitel 6.3.1.) wurden lokal gespeicherte Webseiten für Korpora verarbeitet, um Wörter aus sechs Quellen zu finden. Der Prozeß für "HTML-Analyse" wird hier nicht benötigt, weil der Schwerpunkt des Masterprogramms der Vergleich von normalisierten Worthäufigkeiten mit Hilfe von Datenbanken von "Test 1" und "Test 2" ist. Deshalb wird der textbasierte Browser "lynx[6]" auf Shell-Ebene einfach verwendet, um nur den Inhalt der HTML-Seite in ASCII-Format umzuwandeln. Bei der Eliminierung von Javascripten und Style Sheets für die Verarbeitung des reinen Inhalts der HTML-Seite können manchmal Probleme auftauchen, wenn ein Programm nicht sorgfältig programmiert ist. Es wird HTML-Stripper zur Textbereinigung genannt. Beim Browser "lynx" gibt es dafür die folgenden drei Optionen (`-force_html, -nolist, -dump`)[7]. Mit Hilfe von backtik (‘) wird es im Programm verwendet:

```
lynx -force_html -nolist -dump $url
```

[6] lynx - a general purpose distributed information browser for the World Wide Web [man lynx]

[7] -force_html: forces the first document to be interpreted as HTML.
-nolist: disable the link list feature in dumps.
-dump: dumps the formatted output of the default document or one specified on the command line to standard output.

Der Input des Masterprogramms ist eine URL-Adresse.
Mit Hilfe von 'lynx' wird der reine Inhalt der HTML-Seite identifiziert. Die folgende dreistufige Untersuchung von a bis c wird schrittweise durchgeführt, um branchenspezifisches Vokabular aus einer Webseite automatisch zu erkennen:

a. **Tokenisierung**

b. **Schlüsselwortextraktion (engl. keyword extraction)**

c. **AGBV aus einer Webseite**

Tokenisierung ist die erste Stufe.
Durch das Verfahren "Tokenisierung" kann ein Text in Wörter und sonstige Texteinheiten zerlegt werden. Diese Aufgabe ist in der Praxis aber nicht einfach, z.B. bei der Behandlung von Kontraktionen, Abkürzungen und mit Bindestrich zusammengesetzten Wörtern. Letztere werden in dieser Arbeit als ein Token behandelt. Das bedeutet, dass sie nicht weiter zerlegt werden. Es folgt eine allgemeine Definition von "Tokenisierung":

> Tokenisierung bezeichnet in der Computerlinguistik die Segmentierung eines Textes in Einheiten der Wortebene (manchmal auch Sätze, Absätze o.ä.). Die Tokenisierung des Textes ist Voraussetzung für dessen Weiterverarbeitung, z.B. zur syntaktischen Analyse durch Parser oder im Textmining.[8]

Schlüsselwortextraktion ist die zweite Stufe.
Es gibt verschiedene Berechnungen zur Termgewichtung, um Schlüsselwörter aus einem Text automatisch zu erkennen. Die TF-IDF-Gewichtung, die im Kapitel 2 vorgestellt wurde, ist eines der häufigsten Verfahren zur Termgewichtung.
Für die Schlüsselwortextraktion werden folgende Wortgruppen im Masterprogramm verwendet:

- **1220 Stoppwörter**
- **13987 geographische Namen** für Deutschland

[8] http://de.wikipedia.org/wiki/Tokenisierung [Stand: 03.07.2007]

- **23729 branchenneutrale Stoppwörter**, die durch "Test 1" automatisch gesammelt wurden. Im Kapitel 9.4. ("Branchenneutrale Stoppwörter") wurden sie erklärt.

"AGBV aus einer Webseite" ist die dritte Stufe.
Im "Test 1" wurden bereits die 20 Datenbanken für den Vergleich der Frequenzlisten in den jeweiligen E-Commerce-Branchen maschinell erstellt. Diese 20 Datenbanken werden zum Vergleich der Frequenzlisten für "AGBV aus einer Webseite" auch experimentell wieder verwendet. Die AGBV aus einer Webseite hängt von dem Vergleich der Frequenzlisten ab. Die folgenden zwei Arten automatischer Gewinnung von branchenspezifischem Vokabular werden prinzipiell gleich behandelt:

I. AGBV aus **den erstellten Korpora aus dem Web**

II. AGBV aus **einer Webseite**

9.11.2 Ein Beispiel von "AGBV aus einer Webseite"

Die URL-Adresse "de.wikipedia.org/wiki/BMW" (Stand: 04.07.2007) zeigt eine Webseite zu "Bayerische Motoren Werke AG (BMW)". Die Webseite beinhaltet viel gutes branchenspezifisches Vokabular. Deswegen wird derjenige Teil der automatisch erkannten branchenspezifischen Wörter mit der höchsten Frequenz als Beispiel (siehe unten) gezeigt. Um das Masterprogramm für "AGBV aus einer Webseite" auszuführen, werden zwei Parameter, nämlich eine URL-Adresse und ein Branchenname benötigt. Jeder Abstandswert der branchenspezifischen Wörter ist höher als '0.90'. Die normalisierte Worthäufigkeit zum Vergleich der Frequenzlisten steht direkt vor dem Wort. Und nach dem Wort steht der jeweilige Abstandswert in eckiger Klammer (z.B.14.2061281337047 BMW [0.991524861070098]):

Programm	Eingabe (Parameter)
BVaus1webseiteMaster.pl	URL Branchenname
BVaus1webseiteMaster.pl	de.wikipedia.org/wiki/BMW P1

Tabelle 9.17: **Beispiel von "AGBV aus einer Webseite"**

0.90: 14.2061281337047 BMW [0.991524861070098]
0.90: 2.22841225626741 Motorrad [0.919268985149296]

Tabelle 9.17: **Beispiel von "AGBV aus einer Webseite"**

0.90: 1.48560817084494 Sechszylinder-Reihenmotor [0.99986059404798]
0.90: 1.39275766016713 Zweizylinder-Boxermotor [0.999962438701321]
0.90: 1.29990714948932 Rover [0.985260711170038]
0.90: 1.20705663881151 Personenwagen [0.997674752470457]
0.90: 1.1142061281337 Automobile [0.937348904960906]
0.90: 1.0213556174559 Rolls-Royce [0.997674602571698]
0.90: 1.0213556174559 Flugmotor [0.999399177356128]
0.90: 0.928505106778087 Motoren [0.979915793036673]
0.90: 0.928505106778087 Fahrzeuge [0.93372948266612]
0.90: 0.835654596100279 Motorsport [0.971787120718969]
0.90: 0.74280408542247 Umsatz [0.920244026037079]
0.90: 0.557103064066852 Vierzylinder-Reihenmotor [0.999718290259905]
0.90: 0.557103064066852 Vorstandsvorsitzender [0.986487450957429]
0.90: 0.557103064066852 Automobilhersteller [0.996222663465486]
0.90: 0.464252553389044 CSL [0.998118592018757]
0.90: 0.371402042711235 Weltkrieges [0.983164525852041]
0.90: 0.371402042711235 Automobilen [0.996351516203289]
0.90: 0.371402042711235 Jahresumsatz [0.984972452780236]
0.90: 0.371402042711235 Motorbuch [0.998905158069696]
0.90: 0.371402042711235 Coup [0.978636175903457]
0.90: 0.371402042711235 Olympiaturm [0.998384443625789]
0.90: 0.371402042711235 Kuenheim [0.999264974066327]
0.90: 0.278551532033426 Baubeginn [0.985498029380388]
0.90: 0.278551532033426 Eisenacher [0.996488512842116]
0.90: 0.278551532033426 Vierzylinder [0.999290126750807]
0.90: 0.278551532033426 Mille [0.986526102552025]
0.90: 0.278551532033426 Vorstandsvorsitz [0.998523284363705]
0.90: 0.278551532033426 Nockenwelle [0.997474405371503]
0.90: 0.278551532033426 FIZ [0.995432503511843]
0.90: 0.278551532033426 BFW [0.992421527193507]
0.90: 0.278551532033426 V-Form [0.998722812554121]
0.90: 0.278551532033426 Stammwerk [0.998432384942972]
0.90: 0.278551532033426 Flugzeugmotoren [0.998914742700096]
0.90: 0.278551532033426 Innovationszentrum [0.994698974748456]
0.90: 0.278551532033426 Kleinstwagen [0.999209227755816]
0.90: 0.278551532033426 Produktionsstandorte [0.998048539951172]
0.90: 0.278551532033426 Tourenwagen [0.997962774909133]
0.90: 0.278551532033426 BMW-Motorräder [0.99930839154763]

Ergebnis von [de.wikipedia.org/wiki/BMW] im Bereich von [P1 / Autobranche]

Die erste Stufe ist Tokenisierung - Anzahl der gesamten Tokens: [5403]

Die zweite Stufe ist Schlüsselwortextraktion - Anzahl der erkannten Schlüsselwörter: [1077]

Die dritte Stufe ist Automatische Gewinnung von branchenspezifischem Vokabular:
Anzahl der erkannten branchenspezifischen Wörter: [369]

9.11.3 CGI-Programm für "AGBV aus einer Webseite"

Das entsprechende CGI-Programm[9] wird im WWW präsentiert.

Die Abbildung 9.4 zeigt das CGI-Programm für "AGBV aus einer Webseite" mit einer URL-Adresse beispielsweise in den erstellten 20 E-Commerce-Branchen. Durch das CGI-Programm werden branchenspezifische Einwortterme aus einer Webseite erkannt. Als Zusatzergebnis wird der Inhalt der Original-Webseite phrasenweise zerlegt. Bei jedem phrasenweise zerlegten Teil wird überprüft, ob automatisch erkannte bereichsspezifische Einwortterme enthalten sind. Die Phrasen, die automatisch erkannte bereichsspezifische Einwortterme beinhalten, sind gute Kanditaten, um branchenspezifische Mehrwortterme maschinell zu identifizieren. Es wird angenommen:

> Ein branchenspezifischer Mehrwortterm beinhaltet einen branchenspezifischen Teil des Terms. Der Teil muss zuerst erkannt werden. Dann kann ein branchenspezifischer Mehrwortterm auch erkannt werden.

9.11.4 Branchenneutrale Stoppwörter aus Test 1 und Test 2

Im Kapitel 6.3. wurden branchenneutrale Stoppwörter vorgestellt.
Im Experiment werden die branchenneutralen Stoppwörter, die öfter vorgekommen und sogar in mehr als 15 Branchen aufgetaucht sind, in einer Datei (z.B. '`Computer_Birrelevant_*`' im Beispielbereich von 'Computer') in den jeweiligen Branchen automatisch gesammelt. Die Gesamtzahl der automatisch erkannten branchenneutralen Stoppwörter aus "Test 1" und "Test 2" ist die folgende:

Test 1	Test 2
23729	1259563

Tabelle 9.18: Branchenneutrale Stoppwörter aus Test 1 und Test 2

Die 23729 branchenneutralen Stoppwörter aus Test 1 wurden ohne manuelle Arbeit für das schon oben genannte CGI-Programm "AGBV aus einer

[9] http://www.cis.uni-muenchen.de/ kimda/cgi-bin/AGBV/agbvLocaldbm.pl

Webseite" direkt gebraucht. Diese sind nützlich für die Schlüsselwortextraktion und AGBV.
Beim CGI-Programm "AGBV aus einer Webseite" gibt es die folgenden drei Auswahlmöglichkeiten als Update. Die Anzahl steht für verschiedene Stoppwörter:

- die 23729 branchenneutralen Stoppwörter aus Test 1
- die 1259563 branchenneutralen Stoppwörter aus Test 2
- die 1277559 branchenneutralen Stoppwörter aus Test 1 und Test 2

Dadurch kann der direkte Nutzen der drei branchenneutralen Stoppwörter für **AGBV** ohne manuelle Arbeit überprüft werden.

Das entsprechende CGI-Programm[10] für den direkten Nutzen der drei branchenneutralen Stoppwörter wird zum Test im WWW präsentiert.
Nach dem Test wird festgestellt, dass die 1259563 (Test 2) und 1277559 (Test 1 und Test 2) branchenneutralen Stoppwörter nicht effizient sind.
Viele branchenspezifische Wörter können nicht identifiziert werden, weil sie in der riesigen Stoppwortliste vorhanden sein können.
Die beste Qualität haben die 23729 branchenneutralen Stoppwörter aus "Test 1". Sie können manuell verbessert werden.

9.12 Teile von Test1 und Test2 mit der höchsten Frequenz

Der Teil des automatisch erkannten branchenspezifischen Vokabulars von "Test2" mit der höchsten Frequenz in den jeweiligen ausgewählten 20 Branchen wird im Anhang beigefügt.
Der Teil zwischen "Test1" und "Test2" mit der höchsten Frequenz in den jeweiligen Bereichen beinhaltet viele gemeinsame Wörter. Die AGBV von Test1 und Test2 ist besonders optimal. Die Unterschiede sind gering.
Die Top40-Terme des automatisch erkannten branchenspezifischen Vokabulars

[10] http://www.cis.uni-muenchen.de/ kimda/cgi-bin/AGBV/agbvLocalsto.pl

mit Großschreibung - der Abstandswert ist höher als '0.90' - werden exemplarisch in den jeweils ausgewählten 20 Branchen angezeigt.
Die Worthäufigkeit steht vor dem Term. Nach dem Term steht der Abstandswert innerhalb der eckigen Klammer:

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

Altenpflege

3.94250703087624 ISBN [0.905048964380562]
1.34119460064474 Ill [0.915920012857051]
1.1513199566597 Anmeldeschluss [0.968877604903164]
1.01495002004564 Darst [0.90879080485069]
0.530575928247058 Univ [0.90684314912089]
0.464552013580359 Aufsatz [0.967673320077508]
0.448157813550253 Schlagworte [0.974872347645469]
0.434148224433618 Veranstaltungsort [0.964847705943688]
0.432657842612699 Altenpflege [0.9665116976762]
0.329523420605125 Bemerkung [0.941540413464156]
0.322369587864715 Nachdruck [0.980760539008441]
0.31640806058104 Bibliogr [0.994748655990049]
0.22713418950801 Pflegeheim [0.953948893114891]
0.221321700406427 Lehrerfortbildung [0.963469887787188]
0.213869791301834 Krankenpflege [0.904813378671471]
0.201201545824025 Schulamt [0.967059287906915]
0.196432323997085 Staatliches [0.959299123872437]
0.184807345793919 Dillingen [0.935049644313255]
0.183913116701368 Lehrkräfte [0.952384969691977]
0.169456413038456 Personalführung [0.962870796979451]
0.161557389387587 Verl [0.901963277601724]
0.160663160295036 Lehrgang [0.905826200462702]
0.158874702109934 Hauptschule [0.932869125284764]
0.15619201483228 Berufsausbildung [0.927959671355487]
0.147845876635135 Altenheim [0.956591529946224]
0.143970883900747 Ammersee [0.95519915040891]
0.143076654808195 ALP [0.950717276409317]
0.137562242070796 Angebotsraum [0.995816790694795]
0.131302638422938 Detektive [0.912273813237567]
0.116249782031659 Zeitschriftenaufsatz [0.90369671446163]
0.10283634564339 Pflegedienst [0.932534737705203]
0.100153658365737 Wirtschaftsdetektei [0.915913394728278]
0.0892738710730302 Literaturverz [0.904757189116209]
0.0892738710730302 SYNONYME [0.925345645664879]
0.0891248328909383 Detektivbüro [0.938890411964876]
0.0886777183446627 Facharbeiter [0.975345627922626]
0.0856969547028253 Privatdetektive [0.935384249360327]
0.0809277328758855 Häusliche [0.917244836830069]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

0.0774998546877725 Jobbörsen [0.903281972823875]
0.0766056255952212 Ausbildungsberuf [0.983699573853658]

Autobranche

1.88097295894277 Nederland [0.976239506028967]
1.78736611321734 Reacties [0.992482692754573]
1.78267196461164 BMW [0.932461545377272]
1.31270485126452 Audi [0.952333303109762]
1.29365213045315 Volkswagen [0.975019838937566]
1.18402995419063 AUTOHAUS [0.957956687128599]
1.16663399171069 Ford [0.911273468328706]
1.14509613340218 Reageer [0.985076188085771]
1.11831187371083 Permanente [0.987166725486534]
1.01642123632829 Geplaatst [0.99292540082601]
0.996816262739781 Opel [0.948306741286677]
0.950427029459924 Autobedrijf [0.999424006231778]
0.846051254580246 Renault [0.966956776043295]
0.838871968477412 Toyota [0.947900528308462]
0.825894028214594 Porsche [0.949778129012194]
0.812639961563207 Aangeboden [0.988834753793764]
0.802423285186095 Alfa [0.971812447514781]
0.643650611758014 Volvo [0.952569112885312]
0.630396545106626 Romeo [0.97208428866764]
0.62763528122092 Nederlandse [0.986618103202246]
0.624045638169503 Fiat [0.946568549764727]
0.623217259003791 Peugeot [0.96815453444238]
0.623217259003791 Editie [0.999004649672739]
0.556394672969712 Seat [0.925423553252968]
0.495370741095614 Kia [0.980526688470192]
0.490400466101344 Chrysler [0.932127394515862]
0.489019834158491 Nissan [0.935086253491818]
0.48846758138135 Rover [0.960775888393097]
0.476594146672815 Automotive [0.919767766742419]
0.470795492512833 Europese [0.990654344411818]
0.45478016197574 Mazda [0.93682511087497]
0.453951782810028 Lees [0.977630295374554]
0.441802221712922 Mitsubishi [0.928443984857086]
0.421092742570129 Skoda [0.973507678652325]
0.418883731461565 Rotterdam [0.936361219061343]
0.404249032867324 GTI [0.976468007711905]
0.392375598158789 Saab [0.978139827932626]
0.381606669004537 Bron [0.984966268609675]
0.381606669004537 Krant [0.996507992223872]
0.379673784284543 Dura [0.973881273834959]

Bank

15.8502143878037 Bank [0.976892821961235]
10.232936318882 Kredit [0.980260032099562]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

3.87371764332222 Banking [0.981266525695223]
3.80384309988884 Schufa [0.982073176489352]
2.24030490709862 Konto [0.962463717950239]
2.05126250595522 Banken [0.935736096037818]
1.45212005717008 Visa [0.942807138972903]
1.27743369858663 Kreditkarte [0.951001322738172]
1.16830236620613 Sofortkredit [0.985357000180921]
1.14263935207242 MasterCard [0.982310520330761]
1.08546927108147 Finance [0.906409470033986]
1.08534222645704 Zinsen [0.949625504773555]
1.06349055105606 Commerzbank [0.990669596779802]
1.061076703192 Sparkasse [0.956748036995449]
1.03846276004447 Girokonto [0.96974748145962]
0.673590598697793 Privatkredit [0.98438611507429]
0.673336509448944 Kreditkarten [0.947219511722107]
0.651230744799111 Postbank [0.982471820349981]
0.644116245831348 Volksbank [0.960020836551624]
0.579704621248213 Geldanlage [0.912326688027233]
0.571954899158329 PayPal [0.915015031755722]
0.55378751786565 Savings [0.949038789087187]
0.549849134508496 Citibank [0.984145361417507]
0.541337144672066 Zins [0.957079159012628]
0.494711767508337 Online-Banking [0.981485018643852]
0.469938065745593 Dresdner [0.932621645900882]
0.457106558678736 Bargeld [0.938023732411609]
0.446307765602668 Konditionen [0.910201183023581]
0.433222169286962 Fund [0.907402217572568]
0.422042242337621 Arbeitslos [0.948803814436895]
0.397395585199301 Zus [0.956879022232473]
0.384309988883595 Sparkassen [0.934416081644722]
0.383547721137049 Bsp [0.953817449269457]
0.377576623789106 BLZ [0.955380043244197]
0.368937589328252 Zahlungsverkehr [0.959023198666224]
0.362966491980308 Bankkonto [0.963115167800758]
0.354835636017151 PIN [0.906137735688561]
0.354327457519454 Tagesgeld [0.974726524710082]
0.344163887565507 Asset [0.923498703874287]
0.324344926155312 Bausparen [0.920391622174323]

Bueroartikel

7.60676356402605 Hong [0.979332076175518]
7.60419674655886 Kong [0.979176743591636]
2.12789168030288 Biete [0.903061764123204]
1.62928738730067 Bürobedarf [0.980241619088686]
1.16982706067315 Papier [0.921658429225928]
1.02672698687715 Büroartikel [0.990114700732875]
0.99656688163763 Toner [0.92507285274358]
0.680848333172907 Büromöbel [0.940200997695145]
0.607052330991112 Tinte [0.946711663582336]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

0.589726313087561 Schreibwaren [0.950891070878053]
0.505663041036994 Werbeartikel [0.940487871448391]
0.492187249334232 Amway [0.989192710029145]
0.486411910033048 Druckerpatronen [0.964278748687715]
0.458818622260725 Tintenpatronen [0.953590830001685]
0.441492604357173 Büromaterial [0.980489107767557]
0.390797959380114 Lexmark [0.933368115474733]
0.355504219206212 Ordner [0.944788177266595]
0.354862514839414 Schulbedarf [0.986890527845091]
0.334969679468669 Kugelschreiber [0.959836016175583]
0.324060705233099 Kopierer [0.944235105608894]
0.308659800429942 Etiketten [0.943876282360445]
0.295825713093978 Bürotechnik [0.966303346266774]
0.284916738858408 Druckerei [0.912640811604631]
0.272724355889242 Schreibgeräte [0.973272481111111]
0.236147206981744 Visitenkarten [0.910590966137663]
0.231655276414156 Shopverzeichnis [0.922231551977584]
0.224596528379376 ROTTWEIL [0.958750362194673]
0.223954824012577 Laserdrucker [0.911107721326532]
0.213045849777008 WUNSTORF [0.979590571248423]
0.211762441043411 Büromaschinen [0.950512731504545]
0.207912214842622 Viking [0.925014036623914]
0.197003240607052 Leitz [0.977323952825486]
0.193153014406263 Folien [0.911415374762841]
0.184169153271088 Shopsuchmaschine [0.974816644892072]
0.184169153271088 Hauptfach [0.927095156899779]
0.181602335803895 EDV-Zubehör [0.976013708025118]
0.177110405236308 Pirna [0.936338438931189]
0.175826996502711 Farbbänder [0.981081763785306]
0.167484839734334 Locher [0.971010069599357]
0.162992909166747 Faxgeräte [0.938567231744411]

Computer

17.8449312747271 Computer [0.956976459900244]
7.69689015303382 Dell [0.994528494633606]
5.23855272215624 Notebook [0.981119497343]
3.92607778021529 Memory [0.984949859779125]
3.82319790004009 Windows [0.916453016100691]
2.9695218354981 Samsung [0.971394511519274]
2.77435265104809 Hardware [0.941881939678878]
2.49937591249158 Microsoft [0.925795615734004]
2.43242834340699 Sony [0.943065580756084]
2.35205343702011 Apple [0.965136468601142]
2.31744494791706 Ringtones [0.934749632878545]
2.16955512016521 ThinkPad [0.997408068353361]
1.98327445477446 USB [0.941132590763211]
1.78810527032445 Intel [0.975437424987834]
1.78564674377615 Desktop [0.974220238211614]
1.73950209163874 Linux [0.933359831528582]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

1.68333421588132 Notebooks [0.969663335775764]
1.6196016430522 Wireless [0.923990457228963]
1.60068990037294 Toshiba [0.976695924565312]
1.54187438064043 Series [0.944862678770448]
1.44050743987957 LCD [0.959630494175345]
1.42197393205389 Compaq [0.979176935952131]
1.3858525035365 Monitor [0.945782224836766]
1.34330108250815 Mac [0.927449268291941]
1.22434622105558 RAM [0.967829840566109]
1.2005174252797 Vista [0.958892843335759]
1.17706686435742 Acer [0.981345107021644]
1.1517251291672 Adapter [0.966169931761881]
1.07702374558411 Core [0.967682371420161]
1.07059375307316 Accessories [0.943096547437725]
0.979817388212689 Inspiron [0.996723646027537]
0.957879766704742 DISS [0.942510586478636]
0.943128607414916 Electronics [0.946579422823934]
0.934429205782454 Mouse [0.972381165226943]
0.931025092100187 Rechner [0.917555129707202]
0.921190985906969 PCs [0.950502508942864]
0.911167762286959 Logitech [0.984108183470697]
0.908709235738655 Shipping [0.937054968356149]
0.852163125127654 Maus [0.949026272237818]
0.829090799058952 Networking [0.966575634209183]

Finanzen

2.29948098325195 Finanzen [0.918717726080232]
0.355312082018726 Bundesministerium [0.909312656085727]
0.346928457605099 Buchhaltung [0.927033418080388]
0.299890564902692 Steuerberatung [0.918829595407058]
0.217590251956712 UStG [0.957312977306558]
0.199607057603819 Steuererklärung [0.934621396318209]
0.15103323371113 Anwaltskanzlei [0.913142336455532]
0.147449394267137 Steuerkanzlei [0.957626011113978]
0.13868178705594 Anleihen [0.90627858359822]
0.130106171243528 Finanzplanung [0.905487031152406]
0.128058262989818 Betriebswirtschaft [0.903520832420037]
0.111291014162566 Beifall [0.913390257973828]
0.108411143180785 Finanzbuchhaltung [0.904993097109621]
0.108347146047857 BMF-Schreiben [0.988740938722707]
0.100603492963515 Gemeinderat [0.938621659558542]
0.100219510165945 E-Government [0.910994268004328]
0.0981716019122343 WKN [0.930145349912238]
0.0967636649878085 BMF [0.965364220336107]
0.0954197251963112 Jiangsu [0.9900103998268]
0.0920918742840321 Einkommensteuer [0.901418158448863]
0.0895319889668943 Parlaments [0.900570848423188]
0.083452261338692 Scheer [0.949325593812212]
0.0796124333629853 Businessplan [0.947415446243098]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

0.0761565881848493	Aktienkurse [0.906054916083763]
0.0757086082543502	ISIN [0.916664833120417]
0.0720607716774288	Wirtschaftsprüfung [0.912470454266923]
0.0698848691578617	Optionsscheine [0.914755415534844]
0.0684769322334359	Jahresabschluss [0.914160508345965]
0.0680929494358653	Vorsteuerabzug [0.903421044234588]
0.0671329924419386	Steuerpflichtigen [0.91518199779323]
0.0652770755870137	Betriebswirtschaftslehre [0.900484893626288]
0.0641891243272301	Finanzverwaltung [0.926845306790166]
0.0622052132064484	Vertragsparteien [0.909836469494767]
0.0611172619466648	BFH [0.900040841604292]
0.0597733221551674	Schultz [0.928218902321135]
0.0584293823636701	Kostenrechnung [0.914595590679869]
0.057981402433171	Haufe [0.921323800847192]
0.0548455429196772	Sachbearbeiter [0.900421500473605]
0.0540135801916074	Erwägung [0.941068938257727]
0.0529896260647523	Lexware [0.927170268508117]

Gesundheit

0.328049365362978	Akupunktur [0.901688062177247]
0.254450013678262	DKK [0.90062052566918]
0.208893960846976	Tomaten [0.913212528517816]
0.198935728192901	Übergewicht [0.90068248377644]
0.16745855601048	Diäten [0.934329131275699]
0.128770249837177	ISSN [0.90770966010325]
0.125794226285385	Gewichtsreduktion [0.969444043007967]
0.118125242517304	Anreisetag [0.907436040887517]
0.116637230741408	Aella [0.936399737795842]
0.115378143854111	Immunsystem [0.902952234363874]
0.112745507635217	Ernährungsberatung [0.906308141471562]
0.10770916008603	Syndrom [0.904034581480371]
0.104847598978537	Gesundheitsförderung [0.915251812560343]
0.0999257138736495	Kursort [0.993575941125168]
0.0994678640964506	Bildungsinstitution [0.997499294506444]
0.0980943147648541	Gesundheitsberatung [0.95646372839184]
0.0960339907674592	Triathlon [0.906384059519373]
0.0858468332247848	Berufsbegleitend [0.940357342455181]
0.0796658612326003	Möblierung [0.917921542175193]
0.0774910747909057	Schwindel [0.920251295995659]
0.0758886005707097	Farb-TV [0.918626258005779]
0.0753162883492112	Springer-Verlag [0.96235859955431]
0.0738282765733149	HWS [0.979374495511034]
0.0733704267961161	Fritsch [0.944075554998519]
0.0708522530215224	Duschkabine [0.921906203680552]
0.0695931661342255	Kaffeem [0.911251652250068]
0.0694787036899258	Holzofen [0.917193108768062]
0.0694787036899258	Vitalität [0.910717061548231]
0.0621531072547442	Dehnen [0.968195437226541]
0.0615807950332456	Strahlenbelastung [0.943520973048316]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

0.0600927832573493 Schwankschwindel [0.997978302677937]
0.0597493959244502 Lahnstein [0.909632962914579]
0.0591770837029516 Gewaltmusik [0.990059355681145]
0.0582613841485539 Waschm [0.902185368130474]
0.0565444474840582 Verseuchung [0.925457573000351]
0.0529961117107671 Atemwege [0.933558001595468]
0.052652724377868 Parabol [0.923901672698611]
0.0497911632703752 Tennisarm [0.980830724516084]
0.0491043886045769 Mietze [0.997126124971748]
0.0481886890501792 Trennkost [0.915664495819017]

Haushaltsgeraete

1.9644381151494 Haushaltsgeräte [0.980770239355009]
1.35389680892431 Bosch [0.962538038252303]
1.19482277132223 Miele [0.990753296300752]
0.965049161452573 Staubsauger [0.968136745324623]
0.894854586129754 Ver [0.959038113637404]
0.827184995530777 AEG [0.965685748108132]
0.680735881547917 Elektro [0.924577181097572]
0.665080976260094 Philips [0.917378457150231]
0.629226193181531 Watt [0.95331103268143]
0.622661232899541 Waschmaschine [0.94082731280936]
0.52317683478015 Bodenstaubsauger [0.995406149748485]
0.505501941713253 Bügeleisen [0.984874704043537]
0.484797066977745 Vorkasse [0.942140520262974]
0.484292070032976 Klimageräte [0.992190946937336]
0.480252094474828 Küchengeräte [0.977341826419014]
0.47974709753006 Energieeffizienzklasse [0.997395080225047]
0.478737103640523 Fassungsvermögen [0.993761481392691]
0.461567207518395 Geschirrspüler [0.939436598972065]
0.458032228905015 Klimaanlagen [0.978401578291396]
0.446922296120108 Waschmaschinen [0.968227083389812]
0.446922296120108 Luftbefeuchter [0.994125507332985]
0.433792375556128 Nachnahme [0.930060886855975]
0.409047525262472 Delonghi [0.994511032871203]
0.397432595532797 Trockner [0.965562820806203]
0.370667757460067 HERNE [0.946683617078642]
0.351477873558865 Stabmixer [0.995887219094014]
0.349962882724559 Elektrogeräte [0.925321970041064]
0.337337959105347 Luftentfeuchter [0.996258088786645]
0.33632796521581 Mixer [0.947674153749059]
0.332287989657663 Privileg [0.975984564509124]
0.321178056872756 Dampfreiniger [0.989193368940687]
0.318148075204145 Kenwood [0.975817731311949]
0.318148075204145 Leopoldshöhe [0.97469798523608]
0.314108099645997 Kondenstrockner [0.996262815310972]
0.312088111866923 Luftreiniger [0.996507036626968]
0.304513157695396 Krups [0.986855892489766]
0.300978179082017 Standmixer [0.997373459767419]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

0.298958191302943 Electrolux [0.990218537858805]
0.296938203523869 Dampfbügelstation [0.996754805853457]
0.274213341009287 Hausgeräte [0.981080005993483]

Hotel

17.4290807719458 Hotels [0.937577120370212]
2.27755111807708 Airport [0.92287315732385]
1.39385745564183 Inn [0.921242973977603]
0.598222083966452 Suites [0.947395515935593]
0.59307737404434 Dusseldorf [0.941163186088356]
0.476663356504469 Resorts [0.921267965010042]
0.470322202414424 Kaiserslautern [0.904315570995019]
0.441248609133655 Prague [0.917195660665345]
0.390639020830093 Downtown [0.944171809043048]
0.357736806211938 Hostel [0.906768081444067]
0.354386762541726 Check-in [0.953913616023759]
0.30688792907479 Check-out [0.941469758190776]
0.303537885404578 Marriott [0.919677590622719]
0.298991397566433 Hostels [0.917108142393054]
0.277694691377227 Israeli [0.958336584765138]
0.267405271533004 Iraq [0.917244004962675]
0.261423050693339 Radisson [0.930335476537663]
0.25891051794068 Floors [0.956671906499634]
0.252330075017049 Typical [0.931179446482929]
0.248381809262871 Hospitality [0.929052769857332]
0.246347854177385 Stadthotel [0.941368925903982]
0.238212033835441 Dorint [0.942864165845559]
0.224572570321006 Sights [0.947510065615016]
0.2032758641318 Wellnesshotels [0.950658123339078]
0.202797286464627 Kempinski [0.926770374538472]
0.201241909046314 Brussels [0.900773004976841]
0.198490087460069 InterContinental [0.942210533639233]
0.194063244038717 Emirates [0.936255132375673]
0.183773824194494 Hotelkritiken [0.975505408770683]
0.183295246527321 Hoteles [0.934255896441109]
0.179346980773142 Naples [0.945038458240672]
0.17527907060217 FBI [0.941905817910291]
0.171330804847992 Crisco [0.998449360695045]
0.166784317009847 Wellnesshotel [0.905002799852025]
0.165587872841914 Sheraton [0.915492631230233]
0.161639607087735 Israelis [0.979307508554018]
0.150512676325959 Americans [0.901008250581571]
0.148957298907646 Hyatt [0.928179740502121]
0.141539345066462 Chaska [0.99799979948668]
0.140103612064943 Alster [0.918500521638741]

Immobilien

10.6812801933773 Immobilien [0.962817178218222]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

2.05381916750404 Wohnung [0.908499971520891]
1.81139887685585 Wohnungen [0.95318280556687]
1.29395614311879 Immobilie [0.920615176945446]
1.27916873826177 Häuser [0.945726983114484]
1.16890360125705 Graz [0.919265552253509]
0.929161344559341 Immobilienmakler [0.966236699391693]
0.884799129988275 Grundstücke [0.977453295291045]
0.829375470839121 Grundstück [0.946616857954011]
0.756020627847594 Miete [0.947112169573108]
0.735877155089603 Makler [0.927607383113922]
0.485888507624828 Eigentumswohnung [0.953307858018007]
0.465512162349404 Alanya [0.961525794071612]
0.464813544797104 Gewerbeimmobilien [0.973048614645748]
0.427204633231607 Einfamilienhaus [0.952324765724817]
0.419054095121437 Mietwohnung [0.948696198227094]
0.41893765886272 Eigentumswohnungen [0.976833746830597]
0.411136429528701 Hausverwaltung [0.964813902532129]
0.401472220055214 Bauträger [0.964317764871235]
0.394835353308362 Mietwohnungen [0.950292851211601]
0.378184968311872 IMMOBILIENVERWALTUNG [0.971126220319592]
0.357226441742864 Objekte [0.90592825225512]
0.329747484685721 Villen [0.944959990612467]
0.278748403367802 Objekt [0.905305104691454]
0.275953933158601 Annonces [0.957073174609734]
0.266406159943831 Immobilienmarkt [0.947738851256004]
0.257673440540078 Büros [0.902018138028502]
0.256741950470344 Gesendet [0.901633363107918]
0.253598171484993 Fertighaus [0.951538019683646]
0.24451614330509 Bauregister [0.99951068908917]
0.24451614330509 Immobilienwirtschaft [0.9826097716327]
0.242536726906906 Immobilienangebote [0.974274619020171]
0.23531767886647 Wohnungssuche [0.966691689501796]
0.234619061314169 Immo [0.929091738308793]
0.214009843521312 Wohnfläche [0.938843430595662]
0.205044251600125 Blanca [0.90507445136517]
0.198989566146856 Bauunternehmen [0.916941160175073]
0.194099243280755 Grundstücken [0.952099716249122]
0.189441792932086 Einfamilienhäuser [0.975060303200966]
0.184318597548551 Mietvertrag [0.931274183672146]

Kleidung

4.71465865000305 Remote [0.900171744184154]
4.03254290942215 Kleidung [0.966862386935684]
2.15124724803412 Jeans [0.963902874112471]
1.64998446417332 Hosen [0.97996570597319]
1.60072128726939 Schuhe [0.916771314825978]
1.43221732098611 Damen [0.913580262686075]
1.3736592050437 T-Shirts [0.961001778311554]
1.24260056460115 Bekleidung [0.930047610645463]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

1.01115658254304 Jacken [0.973875224543551]
0.899882883768115 Pullover [0.964046287273894]
0.878106169449394 Shirts [0.940216466717597]
0.869873509158171 Damenmode [0.955037025749888]
0.825523371460288 T-Shirt [0.9356550179714]
0.820743117097642 Kleider [0.963194373052061]
0.807597417600365 Hemden [0.977056754840234]
0.722482332976585 Dessous [0.90472780796991]
0.663924217034171 Taschen [0.910777593463753]
0.662729153443509 Unterwäsche [0.95742372359725 7]
0.645732693487435 Hose [0.943887243074585]
0.51547076210533 Röcke [0.982564329891369]
0.511752786489939 Jacke [0.9530854973351]
0.478025436264603 Kleid [0.954328313687564]
0.477627081734383 Wäsche [0.909903877224904]
0.468863282069531 Herrenmode [0.946473740883683]
0.461825685368969 Gürtel [0.944317221647138]
0.458904418814019 Sweatshirts [0.978520006814113]
0.452663531173898 Blusen [0.977602353165002]
0.448679985871693 Konfektionsgröße [0.997500207381154]
0.441111249797503 XXL [0.951068362107645]
0.439119477146401 Baumwolle [0.925473108039203]
0.437924413555739 Blazer [0.971477646525854]
0.432878589506279 Restzeit [0.949784509634179]
0.432214665289245 Damenbekleidung [0.962574029429905]
0.430488462324956 Gebote [0.909854722254306]
0.428496689673854 Entfernung [0.900647984714182]
0.410039596440304 Bademode [0.942910755659846]
0.396495542412807 Hemd [0.956961267376789]
0.361705913440216 Kostüme [0.964613737004139]
0.339265274904461 Socken [0.944337683190351]
0.335680084132477 Klamotten [0.957044756114208]

Kosmetik

2.40092802578602 Kosmetik [0.945852033434301]
2.36465937812676 Pte [0.993605181200146]
2.28788838852895 Det [0.975814494014241]
2.07380458113946 Thanks [0.947035662223718]
2.04205187404477 Thank [0.962146390875576]
2.01142815209123 CENSURERET [0.953822743186096]
1.82500448065978 Beauty [0.904898300778693]
1.53203283653283 Make-up [0.983589218521907]
0.692209014664106 Aloe [0.939814411472959]
0.670617173839721 Aveda [0.998632665477123]
0.598079878521199 Nail [0.985083500067173]
0.537114680899406 Vera [0.904786929863553]
0.497882447244641 Gel [0.943466006646358]
0.48574585697734 Spray [0.954696139202181]
0.460767060729522 Körperpflege [0.912689272276229]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

0.419982472505684 Eau [0.964708367711733]
0.387665272840428 Bye [0.907243595342651]
0.382443716562635 Naturkosmetik [0.95012148724401]
0.343493729193157 Parfum [0.923604803441713]
0.342646990337298 Lip [0.977371272298277]
0.342223620909369 Lotion [0.968529171955388]
0.334179601778716 Exporter [0.968329527193677]
0.319926164371769 Lippen [0.945595520129454]
0.307507327819182 Greeting [0.963342956089943]
0.304261495538392 Haarpflege [0.954405040806894]
0.301156786400245 Yves [0.950334250624388]
0.289443565560873 Shampoo [0.952096674770879]
0.288596826705015 Bendrath [0.999026970894727]
0.284786501853652 Powder [0.965626974148466]
0.274202266155425 Lippenstift [0.971179300966503]
0.270533064446706 VICHY [0.983789714702582]
0.267851724736488 YOOX [0.969169002066856]
0.263900276742483 Rouge [0.92156263741442]
0.253174917901612 Toilette [0.901206706285765]
0.24287292848867 Mascara [0.989553977126747]
0.24103832763431 Kiev [0.921828458246554]
0.226220397656791 Lipstick [0.990651661506965]
0.223962427374503 Gesichtspflege [0.969229500431293]
0.221704457092214 Rocher [0.962192904424637]
0.1996892468399 Gloss [0.986344465562072]

Lebensmittel

2.20056765241941 Lebensmittel [0.961928115235074]
1.04568117214415 Kaffee [0.900746455263179]
0.851586220549534 Gemüse [0.947507172473301]
0.787999681529348 Synonym [0.946479418360344]
0.776594989251616 Fleisch [0.936224653929445]
0.701926532263444 Milch [0.942599261245356]
0.629947861189832 Obst [0.930870420559227]
0.590246621090934 Zucker [0.9582536919749]
0.497287619883758 Käse [0.940866948533409]
0.492983962420463 Brot [0.914372190494924]
0.452852356575235 Schokolade [0.931284624274633]
0.44628927894371 Fett [0.909789734737355]
0.400347735523034 Gewürze [0.96380964307377]
0.400240144086452 Gramm [0.941984266420315]
0.388943043245302 Salz [0.920252847049043]
0.376139662291999 Lebensmitteln [0.964328958343876]
0.353653052046282 Reis [0.93222679152511]
0.339020616671078 Zutaten [0.959877092845734]
0.338267476615001 Nahrungsmittel [0.936673216741823]
0.315027726313207 Kräuter [0.923050095917763]
0.309648154484088 Eier [0.941701253711249]
0.303623034035475 Honig [0.93981204046805]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

0.297275139277115 Nahrung [0.905392579459365]
0.292003158884578 Kartoffeln [0.960398745773502]
0.263276245317083 Kakao [0.940118201891827]
0.257573899178217 Getreide [0.947366639601451]
0.25047286436378 Salat [0.943716093961771]
0.228416619864392 Olivenöl [0.917709590324502]
0.223037048035273 Pfeffer [0.933341789658006]
0.215720830347671 Kalorien [0.948547118893547]
0.215720830347671 Früchte [0.904435820036733]
0.208512204096652 Nudeln [0.960716758270381]
0.206145192491839 Mehl [0.963709223654109]
0.194955683087272 Eiweiß [0.946370202026895]
0.194202543031195 Essig [0.93663989059971]
0.188715379765494 Wurst [0.927057214359039]
0.188069831145999 Gentechnik [0.923591080480882]
0.176019590248773 Zubereitung [0.942455893589826]
0.170209652673324 Fette [0.913683706882518]
0.162247886366228 Naturkost [0.913863503185774]

Moebel

4.97171089093652 Möbel [0.955187186246413]
1.00590085030334 Bett [0.920224110647577]
0.947218361098844 Sessel [0.974053095988472]
0.907207573004869 Tisch [0.920039752978336]
0.833706273395197 Betten [0.950079675973283]
0.820072967822435 Schrank [0.969885229037002]
0.762575983450352 Stühle [0.97595730734397]
0.745682539588452 Stuhl [0.968898946455073]
0.709524642199823 Sofa [0.953447202896168]
0.682554407262403 Tische [0.973524653336313]
0.618537146312043 Holzmöbel [0.997570379682188]
0.607571226612213 Regal [0.966694925692015]
0.548592361199612 Sofas [0.985892712193525]
0.526067769383745 Kommode [0.981857749675539]
0.508285196897534 Schränke [0.981970436463607]
0.50709969206512 Haba [0.979309393832556]
0.505025058608395 Gartenmöbel [0.945079834074095]
0.496726524781497 IKEA [0.962459238151218]
0.478054823670975 Regale [0.972877160705013]
0.476572942630457 Inserat [0.963913854021748]
0.466199775346834 Matratzen [0.952937911672618]
0.461161379809074 Kontaktdaten [0.965247870162125]
0.451084588733555 Hocker [0.988924641433214]
0.44812082665252 Gartenwölfe [0.999766519280739]
0.439525916617518 Sideboard [0.991536190543511]
0.416408572385443 Buche [0.957894214719902]
0.405739028893717 Schubladen [0.989722327258013]
0.330755848243526 Kinderzimmer [0.934588662266667]
0.314455156797833 Kiefer [0.929807832428521]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

0.313269651965419 Vitrinen [0.979372689172013]
0.300525475016968 Schreibtisch [0.917794188127725]
0.292523317398173 Esszimmer [0.932859427700723]
0.29133812565758 Tischlerei [0.96170617887587]
0.290152307733344 Polstermöbel [0.969984717544702]
0.289559555317137 Esstisch [0.966697482558148]
0.286003040819895 Rattan [0.979429666849913]
0.284521159779378 Rollcontainer [0.992976744653497]
0.282446526322653 Kid [0.909929191671291]
0.279186388033514 Matratze [0.953790805367766]
0.260514686922993 Kleiderschrank [0.931525655331717]

Reisen

18.3566548139993 Reisen [0.975616156635433]
10.8535231258087 Urlaub [0.958249413885657]
6.83110604298792 Last [0.948139798782664]
6.75508419513273 Reise [0.955243966316914]
6.70246808861226 Lastminute [0.985229651478389]
6.35381616569018 Minute [0.972717557527768]
3.59474737715263 Pauschalreisen [0.991555209698161]
3.54156953284013 Flüge [0.974417518170897]
3.45974306113392 Flug [0.963379057629834]
3.18748747792839 Mietwagen [0.971512391373007]
2.8903281859427 Mallorca [0.950117210896379]
2.84613814630274 Yachtcharter [0.930756910657917]
2.62069404576665 Ferien [0.958294365277689]
2.61769811087581 Türkei [0.942900296711432]
2.40011833942819 Kroatien [0.945964034315667]
2.39543719116124 Spanien [0.905096387770549]
2.38139374636041 Ferienwohnungen [0.952854165518953]
2.28589832171472 Ferienwohnung [0.937079734740257]
2.23665264194646 Ferienhaus [0.939763155264293]
2.17336351737736 Reisebüro [0.970858750334134]
2.05670930256508 Ferienhäuser [0.965767891091829]
1.73633151517535 Griechenland [0.938184941544367]
1.66461632372574 Flughafen [0.929817450244068]
1.52811404026162 Ostsee [0.954399458272977]
1.46913157209811 Kreuzfahrten [0.986445683780927]
1.40191028298477 Reiseveranstalter [0.979505212093725]
1.38992654342139 Teneriffa [0.956911328667629]
1.33618696131686 Canaria [0.981977092294067]
1.2135408767229 Tours [0.924707003495901]
1.18077283885428 Tourismus [0.901366852238914]
1.17290850976581 Reiseangebote [0.979238827049124]
1.16953808301361 Costa [0.910875677087409]
1.16485693474667 Flugreisen [0.982275612545204]
1.10606171251383 Städtereisen [0.98137083085491]
1.0970739078413 Reiseführer [0.968368338490899]
1.06730180486353 Unterkunft [0.934091848185269]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

1.02648219197576 Ägypten [0.954773034017187]
0.950460344120571 Buchung [0.966580212259864]
0.925182143479067 Badeferien [0.998533924535127]
0.895410040501295 Sprachreisen [0.987248312947913]

Restaurant

6.42918229948132 Germany [0.911624795859225]
6.35920217618861 Restaurant [0.959592860126482]
1.94448971636628 Restaurants [0.917956687702469]
1.82032330312903 Munich [0.945143041331177]
1.40077860238016 Western [0.918911066416801]
0.684595467841907 River [0.903170127180747]
0.67501835613078 Mercure [0.962355262722707]
0.67081786853818 Cafe [0.91434003656311]
0.590504545767673 Cologne [0.912776246295264]
0.574206653908386 Rothenburg [0.980315881232853]
0.518760217686069 Church [0.912048133468006]
0.500026043023074 Palace [0.903763659984234]
0.416352330178487 Bavaria [0.920349211959783]
0.409547540278476 Nuremberg [0.973369428765031]
0.348388440930223 Garmisch-Partenkirchen [0.965536849002579]
0.3305783735376 Gasthof [0.922712408560348]
0.304787379719038 Pub [0.928201242059476]
0.288993546370863 Ibis [0.949711246365894]
0.280928610193071 Parkhotel [0.954070241193095]
0.256229743148585 Steigenberger [0.936895301665226]
0.255473655381917 Hesse [0.949757066864147]
0.255389645630065 IAN [0.959955068015899]
0.254213509104137 Novotel [0.939708520744826]
0.248332826474497 Rhine [0.958703731044023]
0.23917576352263 Allemagne [0.914179860192598]
0.235647353944846 Sushi [0.941287548583899]
0.232118944367062 Directions [0.906343831280473]
0.226406281241126 TRAVELNOW [0.994898356061113]
0.220105549852227 Tauber [0.97756781311076]
0.219769510844819 Garni [0.921078939981139]
0.215905062259627 Maritim [0.926349453969543]
0.212796701441103 Landhotel [0.945576010516829]
0.208176165089244 Icons [0.934072091874092]
0.207924135833688 Icon [0.901733666072533]
0.206411960300352 Bavarian [0.964428263888493]
0.200111228911452 Hansestadt [0.908301151111136]
0.19960717040034 Danny [0.961424293917371]
0.196750838837372 Ramada [0.930261221354615]
0.195994751070704 Vegetarian [0.922723338451829]
0.195322673055888 Bistro [0.93481442104965]

Schmuck

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

15.9538796575281 Schmuck [0.988878924824688]
6.46281435518033 Uhren [0.983176950322226]
4.7066018316876 Silber [0.980775152381178]
4.47719340115152 Gold [0.955806142750029]
3.61206407923391 Anhänger [0.985046972131311]
2.56108111333194 Silberschmuck [0.99628420527209]
2.56108111333194 Perlen [0.993098557487211]
2.41788729453179 Ringe [0.988144322656493]
2.33107604188419 Piercing [0.988818904002456]
2.12493660690314 Ring [0.930811972054622]
1.86360788759285 Edelsteine [0.992509413563458]
1.7577041257719 Modeschmuck [0.99283670404979]
1.74905283255273 Armband [0.992882752381063]
1.73533009158438 Ketten [0.989967320031717]
1.73443513021688 Ohrringe [0.995177883079818]
1.68252737090182 Armbänder [0.996736493930725]
1.68014080725515 Kette [0.98732776601085]
1.61600190925092 Jewelry [0.964706975933592]
1.43820291757406 Swarovski [0.986716272593454]
1.28755108737806 Goldschmuck [0.996090184720053]
1.1965633483488 Collier [0.99644112082038]
1.16225649592793 Ohrstecker [0.995204260001173]
1.13421437307956 Trauringe [0.994824266348856]
1.11810506846454 QVC [0.994932934979693]
1.05008800453447 Juwelier [0.988953566408115]
0.985352465618567 Diamanten [0.989075449055293]
0.801288744369201 Edelstahl [0.912408885915719]
0.797410578443364 Ohrschmuck [0.995424295738477]
0.764297007845828 Tattoo [0.941677348471776]
0.699561468929925 Silver [0.902031466933168]
0.697473225739089 Sterling [0.969366020680088]
0.679275677933236 Colliers [0.995295074873793]
0.658989886936547 Halsketten [0.995627041471886]
0.646758748247367 Goldschmiede [0.994095069606081]
0.619313266310671 Fossil [0.974430162618884]
0.615733420840667 Gesamtpreis [0.931790786792094]
0.614241818561499 Esprit [0.921889781602501]
0.598132513946481 Mineralien [0.964181806385435]
0.594850988932311 Armbanduhren [0.976708993971147]
0.594254348020644 Broschen [0.99365688843407]

Versicherung

7.1889686143837 Versicherung [0.971287969425955]
4.72907325709433 Versicherungen [0.958475800360905]
3.58045038107507 Krankenversicherung [0.957695615552343]
3.0183736478852 Vergleich [0.90837005764114]
1.59219089977858 Versicherungsvergleich [0.975885905250883]
1.59027173039509 Lebensversicherung [0.966077821023977]
1.43769776440756 Krankenkassen [0.955318585294914]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

1.22658913222357 Rentenversicherung [0.966748052846018]
1.06561880018328 Unfallversicherung [0.979244491236687]
1.05458357622821 Haftpflichtversicherung [0.984471987829735]
1.03947011733322 Versicherungsmakler [0.960083776303853]
1.02987427041576 Rente [0.924914515828808]
0.99460953299412 Altersvorsorge [0.936128813906095]
0.943271751985741 Hausratversicherung [0.982384569810234]
0.87754020060118 Kfz-Versicherung [0.985471985027741]
0.862186845533253 Autoversicherung [0.977013049928404]
0.834838681818509 Haftpflicht [0.980046938688834]
0.80845010279551 Berufsunfähigkeitsversicherung [0.978034958724004]
0.757592114133003 Krankenversicherungen [0.954186777140281]
0.739599901162777 Berufsunfähigkeit [0.977969538696563]
0.700976617320024 Eventid [0.999714184361191]
0.677946584718134 Krankenkasse [0.926422413899343]
0.663792710514889 Vorsorge [0.90447422753397]
0.662353333477271 Rechtsschutzversicherung [0.985482037646996]
0.6302072463038 Rechtsschutz [0.975649109590697]
0.627088596055627 Unfall [0.933467140883377]
0.621810880251027 Versicherer [0.976877111606487]
0.597101574438583 Versicherungsvergleiche [0.979705666342614]
0.447646258699235 Risikolebensversicherung [0.982859521256184]
0.446206881661617 Versicherten [0.94526019822735]
0.439729684992335 Versicherungsschutz [0.972406840145119]
0.42557581078909 Lebensversicherungen [0.969297349722396]
0.4157400676987 Kranken [0.916054218313472]
0.404225051397755 PKV [0.948847713710626]
0.399427127939028 Hausrat [0.955782139250231]
0.388871696329828 Allianz [0.902675772494898]
0.383114188179356 Wohngebäudeversicherung [0.992430578864348]
0.339213188532003 Riester [0.945401430812372]
0.331776407170976 VALUES [0.938514823355577]
0.326018899020504 INSERT [0.966851492460825]

Wein

6.18690961936169 Wein [0.974278999739971]
2.82525664673113 Weine [0.972926572778829]
2.44759935162931 Weingut [0.989224150049415]
2.13925064743912 Riesling [0.997394916145963]
1.25295772548581 Sauvignon [0.997158081877752]
1.12030257298828 Pinot [0.996904724019849]
1.08415777020103 Blanc [0.984881444895656]
1.08005887503959 Cabernet [0.995376936244812]
1.02845005868873 Rotwein [0.985792209859071]
1.02267434277942 Vin [0.98579587618577]
0.998826225476497 Chardonnay [0.996063533110638]
0.879771952378291 Winzer [0.985641564694958]
0.818288524956682 Weißwein [0.988813785008605]
0.807854973636652 Bordeaux [0.962755047147834]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

0.744694725467181 Sekt [0.963991733459525]
0.704451027518492 Merlot [0.995555214194946]
0.675013507722691 Noir [0.985468628053098]
0.613716393717512 SysK [0.999475989760156]
0.551674026046616 Rheinhessen [0.985804741843106]
0.528012222160118 Domaine [0.974365783548944]
0.516460790341512 Weinbau [0.978075723019615]
0.497643135282172 Chateau [0.964948361606197]
0.490563225457865 Spätlese [0.998168985736069]
0.459448884914203 Winery [0.990551398802627]
0.457585750749911 Champagner [0.95000943543321]
0.432433439531981 Cru [0.992213837091885]
0.419577813798372 Spätburgunder [0.996469986734968]
0.411007396642632 Champagne [0.967139246416769]
0.39461181599687 Vineyards [0.981012834787794]
0.384923518342556 Veltliner [0.993610996358128]
0.382687757345406 Vino [0.970226525884738]
0.3756078475211 Weingüter [0.987143653823992]
0.370577385277514 Weinen [0.958670535008437]
0.36927319136251 Mosel [0.930429949950176]
0.368714251113223 Kabinett [0.967252969763393]
0.366478490116073 Rebsorten [0.991751568847465]
0.357162819294617 Jahrgang [0.950089658665265]
0.346729267974587 Ros [0.964531952633741]
0.342816686229575 Vineyard [0.983671247545746]
0.326607419000242 Syrah [0.995407382769141]

Aerzte

3.71464503526472 Ärzte [0.956689568702376]
2.56876385661419 Medizin [0.908608397777594]
2.44770975474251 Arzt [0.949856622859123]
1.9115457572996 Patienten [0.952798733078944]
1.50517226406744 Presseportal [0.966112005133612]
1.0720664259535 Behandlung [0.911735468005048]
0.935684384648624 Klinik [0.945797639763853]
0.914392673558834 Therapie [0.908982873572615]
0.702260269457851 Krankenhaus [0.913926769483418]
0.604381174988569 Chirurgie [0.926127193173872]
0.60255019246242 Kliniken [0.946100873082758]
0.533966818411523 Psychotherapie [0.91623696815751 1]
0.489343158559948 Innere [0.912283795049037]
0.482490052533504 Facharzt [0.966233632515358]
0.470143998928614 Psychiatrie [0.94144677060426]
0.434256741416092 Homöopathie [0.909493990804972]
0.433576662192094 Erkrankungen [0.92366101987347]
0.419451939847516 Sexkontakte [0.976670287150052]
0.402502273034022 Ärzten [0.954098987513377]
0.402136076528792 Patient [0.920476639305416]
0.40093285944018 Vereinigung [0.91056447315757]

Tabelle 9.19: **Top40-Terme aller 20 Branchen**

0.337528550248961 Pressetext [0.952869657846208]
0.331564778592362 Chiffre [0.969728139869686]
0.318852528482241 Diagnose [0.911088550278041]
0.317283114888399 Mediziner [0.951917675502207]
0.30650647487735 Allgemeinmedizin [0.957856674227551]
0.295782148652763 Klinikum [0.949823878520241]
0.295625207293379 Dosieraerosol [0.910020483070039]
0.28856284612109 Diagnostik [0.93121888010267]
0.285528646506328 Klinische [0.943734371980844]
0.281029660870648 Cor [0.920661634231184]
0.274281182417127 Insulin [0.905698479217068]
0.270619217364829 Erkrankung [0.931217437843327]
0.269834510567908 Seitensprung [0.947779315779834]
0.261778187452852 Kneipp [0.917559691827296]
0.255605160650407 Symptome [0.90695857038028]
0.249432133847961 Neurologie [0.927840878926587]
0.247967347827042 Baierbrunn [0.954459228198086]
0.232744035966774 HCT [0.916650269179757]
0.230442229362473 Fertiginhalat [0.902636592958197]

Automatische Gewinnung von Brachenspezifischem Vokabular (AGBV)

Multiword | Single-Word

URL:

www.webauto.de

Autobranche_P1 ***DBM***: 20_db_Test1

Submit Reset

Ergebnis von [http://www.webauto.de]
Im Bereich von [P1 / Autobranche]

Die erste Ebene ist Tokenisierung:
Anzahl der gesamten Tokens: [257]

Die zweite Ebene ist Schlüsselwortextraktion (engl. keyword extraction):
Anzahl der erkannten Schlüsselwörter: [66]

Die dritte Ebene ist "Automatische Gewinnung von branchenspezifischem Vokabular":
Anzahl der erkannten branchenspezifischen Wörter: [36]

Nr	AGBV aus einer Webseite
1	0.90: 10.6060606060606 Gebrauchtwagen [0.987544609294403]
2	0.90: 6.06060606060606 OPEL [0.99149776764169]
3	0.90: 6.06060606060606 Corsa [0.998516384015364]
4	0.90: 4.54545454545455 Autos [0.971637720276813]
5	0.90: 4.54545454545455 Neuwagen [0.990612867942417]
6	0.90: 3.03030303030303 Xenon [0.998485773738853]
7	0.90: 3.03030303030303 SKODA [0.996318610996397]

Abbildung 9.4: CGI-Programm für "AGBV aus einer Webseite" (Stand: 29.08.2007)

1989. Automatische Gewinnung von branchenspezifischem Vokabular (AGBV)

Kapitel 10

Vergleich mit allgemeinen Korpora für AGBV

Im vorherigen Kapitel 9 wurden die 20 erstellten Korpora aus dem E-Commerce-Bereich miteinander verglichen, um branchenspezifisches Vokabular automatisch zu erkennen. In diesem Kapitel werden allgemeine Korpora zur "Automatischen Gewinnung von branchenspezifischem Vokabular (AGBV)" verglichen, um nicht benötige Wörter, die als "nicht branchenspezifisch" betrachtet werden, zu entfernen. Die Nutzung von allgemeinen Korpora für die AGBV wird überprüft.

10.1 Erstellung von allgemeinen Korpora

Vier Bereiche, nämlich Bibel, Politik, Gedicht und Zeitung, wurden zum Aufbau der Korpora ausgewählt. Für den Aufbau der allgemeinen Korpora werden nur die folgenden zwei Schritte a und b in dieser Arbeit benötigt. Die Anzahl der verwendeten Suchbegriffe pro Korpus ist 30 für das Experiment:

a. **Suchbegriffe als Startwörter für Korpora erstellen**

b. **Masterprogramm für den Aufbau der Korpora aufrufen**

Die jeweiligen folgenden 30 Suchbegriffe für Bibel (A1), Politik (A2), Gedicht (A3) und Zeitung (A4) werden verwendet. 'A' steht für allemeine Korpora. Zwischen der Klammerung steht die Anzahl der verwendeten Suchbegriffe:

Tabelle 10.1: **Suchbegriffe als Startwörter**

Bereich	Suchbegriffe als Startwörter
A1 (30)	Jesus, Liebe, Predigt, Ehebrechen, Fasten, Immanuel, Heilung, Petrus, Gottesfurcht, Täufer, Sabbat, Pharisäer, Senfkorn, Luther, Auferstehung, Vergebung, Nazareth, Segnung, Jerusalem, Dreieinigkeit, Gott, Sohn, Geist, Abendmahl, Gethsemane, Lukas, Lobgesang, Seligpreisung, Tempel, Römer
A2 (30)	Lyrik, Abecedarium, Cento, Haiku, Ode, Terzine, Allegorie, Strophe, Hexameter, Pentameter, Trochäus, Daktylus, Hymne, Reim, Versmaß, Alliteration, Distichon, Jambus, Rhythmus, Anagramm, Elegie, Knittelvers, Schüttelreim, Anapäst, Limerick , Epigramm, Metapher, Sonett, Figurengedicht, Ballade
A3 (30)	Kommunalpolitik, Landespolitik, Bundespolitik, Europapolitik, Weltpolitik, Merkel, Arbeitsmarktpolitik, Außenpolitik, Behindertenpolitik, Bildungspolitik, Drogenpolitik, Energiepolitik, Entwicklungspolitik, Familienpolitik, Finanzpolitik, Forschungspolitik, Frauenpolitik, Gleichstellungspolitik, Gesundheitspolitik, Innenpolitik, Internationale, Landwirtschaftspolitik, Stoiber, Kulturpolitik, Lohnpolitik, Medienpolitik, Minderheitenpolitik, Schulpolitik, Sozialpolitik, Sprachpolitik
A4 (30)	Tageszeitung, Sonntagszeitung, Wochenzeitung, Sonderausgabe, Abonnementzeitung, Berlin, Boulevardzeitung, Anzeigenblatt, Offertenblatt, Mitgliederzeitung, Firmenzeitung, Kommunikation, Betriebszeitung, Pressevertrieb, Zeitungsantiquariat, Archive, Pendlerzeitung, Straßenzeitung, Zeitungsprojekte, Bild, Sport, München, Schülerzeitung, Abiturzeitung, Studentenzeitung, Parteizeitung, Kirchenzeitung, Hochzeitszeitung, Amtsblatt, Gefangenenzeitungen

10.2 Korpuserstellung aus einer Startseite

Im Kapitel 6 “Domainspezifische Korpora aus dem Web” wurde die Methode “Extraktion aus Startseiten” vorgestellt, um domainspezifische Korpora zu erstellen. Dafür braucht man keine Suchbegriffe für Suchmaschinen sondern nur eine Startseite oder mehrere Startseiten. Aus der Startseite - www.vodafone.de (Stand: 07.08.2007) - wird ein Korpus für die Extraktion der domainspezifischen Terme in einem Bereich experimentell aufgebaut, indem die zwei internen Links nämlich “www.vodafone.de und shop.vodafone.de” verfolgt werden. Die Top10-Wörter mit orthographischen Varianten sind:

Die ersten Wörter (z.B. Vodafone, BlackBerry) mit den höchsten Frequenzen sind Kanditaten für Grundformen der orthographischen Varianten. “BlackBerry” bedeutet in diesem Kontext nicht “Brombeere”, sondern ein Handy-Hersteller, weil ’Vodafone’ ein international tätiges Mobilfunkunternehmen ist. ’BlackBerry’ muss als Firmenname domainspezifisch berücksichtigt werden. Im automatischen Terminologie-Extraktionssystem “AGBV” wurde “Black-

138286	Vodafone/ vodafone/ VODAFONE/ Vo-dafone/ Voda-fone
123956	Nokia/ NOKIA
57147	Ericsson
56646	Sony
51149	Samsung
47775	Zubehör/ Zubehoer
45893	Handy/ HANDY/ handy
43785	Motorola
26081	BlackBerry/ Blackberry
21717	Business/ business

Tabelle 10.2: Orthographische Varianten zur Berechnung der Worthäufigkeiten

Berry" als branchenspezifisch erkannt. Solche Wörter (z.B. Zubehör) müssen als domainneutral z.B. durch den Vergleich der Korpora oder die Nutzung der domainneutralen Stoppwortliste erkannt und entfernt werden.

10.3 Erweiterung der normalisierten Datenbanken

Durch "Test 2" wurden 20 Datenbanken für die 20 ausgewählten E-Commerce-Branchen erstellt.
Die fünf Datenbanken, vier allgemeine Korpora und 'www.vodafone.de', wurden als Erweiterung den normalisierten Datenbanken hinzugefügt, um öfter vorgekommene unnötige Wörter (z.B. Zubehör) zu entfernen. Zum Vergleich der Korpora werden die Datenbanken normalisiert, um domainspezifische Terme in den jeweiligen Branchen zu erkennen.
Der Bereich "Bibel" wird z.B. mit den 24 verschiedenen anderen Datenbanken im Experiment zur "AGBV" verglichen. Jeder Bereich hat spezifisches Vokabular, das durch die AGBV erkannt wird.

10.4 Ergebnis der allgemeinen Korpora

Die Tabelle 10.3 fasst den grundsätzlichen Aufbau der allgemeinen Korpora übersichtlich zusammen. In der Tabelle werden die folgenden Abkürzungen verwendet:

Abkürzung	Bemerkung
B	Kodierung der Branchen
BS	Anzahl der verwendeten branchenspezifischen Startwörter
BSS	Anzahl der automatisch erstellten Suchbegriffe für Suchmaschinen
URL	Anzahl der verschiedenen URL-Adressen
Web	Anzahl der verschiedenen lokal gespeicherten Webseiten
OG	Original-Größe der Gesamt-Webseiten
KG	Größe des automatisch erstellten Korpus
F	Anzahl der Einwortterme mit Varianten in einem Korpus
FV	Anzahl der Einwortterme ohne Varianten in einem Korpus
BV	Anzahl von automatisch erkanntem branchenspezifischem Vokabular
Vo	www.vodafone.de

B	BS	BSS	URL	Web	OG	KG	F	FV	BV
A1	30	140	28429	21511	1675 M	678 M	2445477	1986708	12961
A2	30	140	31026	20551	1355 M	460 M	1812490	1499504	10722
A3	30	140	14163	11215	1027 M	342 M	2327805	1923599	4677
A4	30	140	16912	13296	898 M	250 M	1589052	1294069	3807
Vo			13364	12316	257 M	30.2 M	37220	34099	8474

Tabelle 10.3: Übersicht des grundsätzlichen Aufbaus der allgemeinen Korpora

10.4.1 Laufzeit von einem Korpusaufbau

Das Korpus 'Vodafone' wurde durch die Methode "Extraktion aus Startseiten" aufgebaut. Die Laufzeit des Korpusbaus war ca. 90 Minuten. In der in Kapitel 9.6.1. vorgestellten Datei "runtime_doubles.info" steht die Laufzeit pro Branche. Die andere Methode "Extraktion mit Suchmaschinen" dauert wesentlich länger, für das Korpus "Bibel" ca 20 Stunden wie folgt:

Start: 2007-08-07_13:28:48
End: 2007-08-08_09:15:08

10.4.2 Top20-Terme der vier allgemeinen Korpora und Vodafone

Die automatisch erkannten Top20-Terme der vier allgemeinen Korpora und des Vodafone-Korpus werden in der folgenden Tabelle angezeigt, um die Qualität der AGBV zu demonstrieren. Sie enthalten wenige Fehltreffer (z.B. Hilfe, Kontakt). Im Test wurden insgesammt 25 normalisierte Datenbanken aus

den 20 ausgewählten E-Commerce-Branchen, den vier zusätzlichen allgemeinen Korpora (Bibel, Politik, Gedicht, Zeitung) und "www.vodafone.de" für die Entfernung der branchenneutralen Wörter verwendet. Trotzdem bleiben unnötige und domainneutrale Wörter (z.B. Hilfe, Kontakt). Bei der manuellen Auswahl können solche branchenneutralen Stoppwörter zur Entfernung gesammelt werden.
Bei "AGBV" wird die Lemmatisierung nicht angewendet. Deshalb stehen z.B. "Jesus" und "Jesu[1]" oder "Gott" und "Gottes" getrennt in der Liste:

Tabelle 10.4: **Top20-Terme der vier allgemeinen Korpora und 'Vodafone'**

Bibel / A1

15.7110154084043 Gott [0.991187295778645]
12.4580965094015 Jesus [0.994919003469138]
8.94771652401863 Menschen [0.904518214192277]
8.88288565808362 Gottes [0.994974644762667]
7.5517388564399 Leben [0.907628415291328]
6.35669660564109 Kirche [0.974018942351471]
5.70692824511705 Bibel [0.994420658270697]
4.60737058490729 Christus [0.995318491722003]
3.88683188470575 Glauben [0.970159574444795]
3.7665827086819 Geist [0.981172586642356]
3.27622378326357 Wort [0.953768318141223]
3.09345912937382 Jesu [0.995921680086699]
3.02218544446391 Heiligen [0.988963127326986]
2.89977188394067 Bible [0.988670445341098]
2.79225734229691 Mensch [0.941137641187012]
2.74801329636766 God [0.971594309882208]
2.73306394296494 Christen [0.987642662795306]
2.70230954926441 Herrn [0.966381570969196]
2.65398840695261 Vater [0.965711990116926]
2.52875611312785 Sohn [0.969801097235398]

Politik / A2

7.37270457431257 Politik [0.959698154397043]
3.20839424236281 SPD [0.973185578179552]
2.9979913358017 Europäischen [0.927793017754509]
2.41159743488514 CDU [0.972257756081051]
2.13410567761073 Europäische [0.935844180116754]
2.12430243600551 Bildung [0.923310674026565]
1.65214630971308 Grünen [0.948774814632448]
1.59312679392653 Regierung [0.924126122789121]

[1] die Genitiv-Form von "Jesus"

Tabelle 10.4: **Top20-Terme der vier allgemeinen Korpora und 'Vodafone'**

1.44607816984816 FDP [0.976633323977063]
1.40679851470886 Partei [0.941266709597076]
1.39752878285086 Bundesregierung [0.946851342995252]
1.21433487339814 Bürger [0.921257835628342]
1.18172409009913 Beifall [0.9802478093573]
1.17918991879982 Demokratie [0.959706208365997]
1.17718925724773 Präsident [0.929891696019982]
1.12677258613515 Kommission [0.900718767734756]
1.07535558424652 Parteien [0.944985810435577]
1.05014724869023 Integration [0.905696661356442]
1.02433871466832 Merkel [0.96354044657745]
0.986859654925895 Staaten [0.907737010139458]

Gedicht / A3

2.03602725931964 Gedicht [0.983982953388375]
1.52209478170866 Gedichte [0.975223561587404]
1.43439459055656 Treffer [0.962647537464337]
0.976970771974824 Lyrik [0.983102763047103]
0.851528826954059 Dichter [0.964973984159028]
0.722603827512907 Reacties [0.936927125079443]
0.666718999126117 Poetry [0.971342114140138]
0.597369826039627 Geplaatst [0.954769893395016]
0.550946429063438 Goethe [0.947791129747053]
0.444323375090131 Reageer [0.912296454860434]
0.428155764273115 Lied [0.911301310571615]
0.399615512380699 Permanente [0.929158401888119]
0.363537306891925 Poesie [0.965205674663955]
0.349968990418481 Lieder [0.910650764177202]
0.329486550991137 PKM [0.998451378907642]
0.30708063374955 Dichtung [0.956985229848835]
0.285402518924162 Schiller [0.936862893327605]
0.285038617716062 Heidegger [0.983932769179404]
0.280723789105734 Haiku [0.987391905078498]
0.258993688393475 Strophe [0.987307922694195]

Zeitung / A4

5.61616111660197 Zeitung [0.958991898023493]
2.34446540331312 Magazine [0.906154284352079]
0.847713684509868 Tageszeitung [0.967818938361396]
0.790916094891385 Zeitungen [0.916316870517524]
0.500436993699718 Rundschau [0.941484167045158]
0.452757928673046 Süddeutsche [0.949730519826686]
0.366981977004317 Gunning [0.998073357063592]
0.355236080919951 VTEC [0.994354385011893]
0.344803870581862 Wochenzeitung [0.969386220935714]
0.344340216789058 Zürcher [0.915477379188777]
0.333057974497496 Anzeiger [0.96516395555853]
0.322548488527273 NZZ [0.92510186935911]

Tabelle 10.4: **Top20-Terme der vier allgemeinen Korpora und 'Vodafone'**

0.302843202333106 Tagblatt [0.963922244648157]
0.260573431555813 Kurier [0.914153264361715]
0.254159554088692 Schülerzeitung [0.991443798500254]
0.245659234553953 Sonntagszeitung [0.980612386409678]
0.236772536858545 Journalismus [0.926782517071671]
0.220621929742541 Tagesspiegel [0.909878990366098]
0.207175969751227 Morgenpost [0.949176729376613]
0.199448406537828 Wochenblatt [0.980467920959996]
Vodafone (www.vodafone.de)
405.542684536203 Vodafone [0.999951443732198]
363.517991729963 Nokia [0.999422023848293]
167.591424968474 Ericsson [0.999593354594162]
166.122173670782 Sony [0.996102373546427]
150.00146631866 Samsung [0.998609000239595]
140.106747998475 Zubehör [0.998492518261707]
134.587524560838 Handy [0.998286385255052]
128.405525088712 Motorola [0.998976239130752]
76.4861139622863 BlackBerry [0.99968420752221]
63.6880846945658 Business [0.981236693437343]
57.0309979764803 VPA [0.999991073331352]
53.9898530748702 Privat [0.995176362580965]
53.8138948356257 Sagem [0.999824255459028]
51.0161588316373 Zuhause [0.998697301964058]
50.6877034517141 Handys [0.99873964274049]
45.2652570456612 Siemens [0.997002608916628]
41.8692630282413 Presse [0.994655371812477]
41.3120619373002 GmbH [0.957341942545295]
39.1243144960263 Hilfe [0.987936402725397]
38.5788439543682 Kontakt [0.982193123756373]

Kapitel 11

Zusammenfassung und Ausblick

Um Webseiten für E-Commerce inhaltlich zu erfassen, wird in dieser Arbeit branchenspezifisches Vokabular für die jeweiligen Bereiche (z.B. Auto, Computer, Lebensmittel) automatisch gewonnen und semantisch analysiert. E-Commerce-relevante Webseiten können jeweiligen Branchen mit Hilfe von EGT (Elementare Generische Terme) maschinell zugeordnet werden. Die Qualität der EGT spielt eine entscheidende Rolle dafür.

Die folgende Grundannahme für domainspezifische Terme (DST) ist die Grundlage zur Überprüfung der erkannten Wörter in einer Branche:

> Ein Term wird als domainspezifisch betrachtet, wenn er in einem Bereich öfter als andere Terme vorkommt und seltener in anderen Bereichen. Ein domainspezifischer Term beinhaltet mindestens einen domainspezifischen Teil als "Elementaren Generischen Term (EGT)".

Danach wurden die folgenden zwei Zielsetzungen dieser Arbeit realisiert:

I. **Erkennung der domainspezifischen Terme im jeweiligen Bereich durch EGT und domainspezifische Listen (z.B. Firmennamen)**

II. **Erstellung des Terminologie-Extraktionssystems AGBV**: **A**utomatische **G**ewinnung von **b**ranchenspezifischem **V**okabular aus den erstellten Korpora

Domainspezifische Korpora aus dem Web werden erstellt und mit den normalisierten Worthäufigkeiten verglichen, um domainspezifische Terme in den

jeweiligen Bereichen zu erkennen. Dabei handelt es sich um die vorgestellte Berechnung für den Abstand zwischen höheren und niedrigeren Frequenzen eines Terms in allen Bereichen (Abstandswert / Average Deviation). Der Vergleich von Korpora ist sehr nützlich, um branchenneutrale Wörter zu entfernen. Durch die AGBV wird branchenspezifisches Vokabular als eine Basis für die semantische Arbeit in den jeweiligen E-Commerce-Branchen gewonnen. Bei der semantischen Annotation können EGT, Marken und branchenneutrale Stoppwörter im jeweiligen Bereich erkannt und abschließend manuell überprüft werden. Die automatisch und manuell erstellten branchenneutralen Stoppwortlisten werden für die Eliminierung der branchenneutralen Wörter dringend benötigt. Die Affixanwendung von EGT und die Überprüfung mit Hilfe domainspezifischer Listen (z.B. Marken) sind signifikant zur Erkennung der DST im jeweiligen Bereich. Webseiten können dadurch bewertet und klassifiziert werden.

11.1 EGT und KGT

Komplexe generische Terme (KGT) müssen mindestens einen EGT beinhalten. Die EGT sind eine Teilmenge von KGT. Ein EGT kann einem Bereich oder mehreren Bereichen zugeordnet werden. In den jeweiligen Bereichen gibt es eine endliche Menge (z.B. 10000) von EGT, um domainspezifische Terme bzw. KGT für E-Commerce zu erkennen und zu klassifizieren.

11.2 Bootstrapping-Verfahren mit EGT und Marken

Verschiedene automatische Annotationsverfahren (bzw. POS-Tagger) werden bei vielen NLP-Techniken (z.B. POS-Muster) eingesetzt, um Mehrwortterme bzw. Phrasen zu erkennen.
Die wichtigsten POS-Muster für Mehrwortterme sind Adjektiv-Nomen, Nomen-Nomen und Nomen-Präposition-Nomen. Automatisch annotierte Nomen und Adjektive, die falsch oder unerwartet annotiert sind, bringen jedoch unvermeidliche Fehltreffer. Die NLP-Techniken beschäftigen sich damit, solche Fehltreffer von POS-Muster zu verhindern. Hierfür sind "lokale Grammatiken" eine Lösung. Durch sie können bestimmte Phrasen für Mehrwortterme z.B. mit-

tels "Bootstrapping-Verfahren mit EGT und Marken" erkannt werden, indem solche Fehltreffer von POS-Mustern verhindert werden. Die korrekt erkannten Mehrwortterme (z.B. VW Golf) müssen bei der Bewertung domainspezifischer als Einwortterme (z.B. VW) sein.
Neue EGT können durch das im Rahmen dieser Arbeit erstellte Terminologie-Extraktionssystem "AGBV" erkannt werden. Durch Kontextbetrachtung der erkannten EGT und Marken können branchenspezifische Mehrwortterme mit relativ wenig Fehltreffern gewonnen werden.

11.3 AGBV und semantische Kodierung

Durch die "AGBV" und "semantische Kodierung" können EGT, Marken und branchenneutrale Wörter pro Branche unterschiedlich gekennzeichnet werden. Die korrekt semantisch annotierten EGT und Marken sind die Basis für die Erkennung der Einwort- und Mehrwortterme.

11.4 Im Rahmen der Dissertation erstellte Webdemonstrationen und Informationen

Im Rahmen der Dissertation erstellte Webdemonstrationen und Informationen kann man auf der Homepage (`www.cis.uni-muenchen.de/~kimda`) finden.

Anhang

Anhang A

Semantische Annotation im Automobilbereich

Tabelle A.1: **Semantische Annotation im Automobilbereich**

Frequenz	**Varianten [semantische Annotation]**
368318	EUR/ Eur/ eur/ EuR [NO]
137192	Auto/ auto/ AUTO/ AUto [Auto;EDST]
131216	GmbH/ GMBH/ gmbh/ Gmbh/ gmbH/ GmBH/ GMbH [NO]
112108	war/ War/ WAR/ wAr/ WAr [NO]
98822	Audi/ AUDI/ audi/ AUdi/ AuDi [Audi;Automarke]
91348	BMW/ bmw/ Bmw/ BMw/ B-M-W/ bMW/ BmW [BMW;Automarke]
91158	dann/ Dann/ DANN/ DAnn [NO]
79255	www/ Www/ WWW/ wWW/ WwW [NO]
75382	dass/ Dass/ DASS/ daSS [NO]
71071	eBay/ ebay/ Ebay/ EBAY/ e-bay/ E-Bay/ EBay/ E-bay/ ebaY/ e-Bay/ E-BAY/ eBAY [NO]
67945	Renault/ RENAULT/ renault/ REnault [Renault;Automarke]
60709	Artikel/ artikel/ ARTIKEL [NO]
60040	Ford/ FORD/ ford/ FOrd [Ford;Automarke]
56591	Preis/ preis/ PREIS [NO]
55456	Fiat/ fiat/ FIAT [Fiat;Automarke]
53758	Alfa/ ALFA/ alfa/ ALfa [Alfa;Automarke]
53700	Opel/ OPEL/ opel/ OPel/ O-P-E-L [Opel;Automarke]
52495	Deutschland/ deutschland/ DEUTSCHLAND/ Deutsch-land/ DEutschland [NO]
51886	Uhr/ uhr/ UHR/ UHr [NO]
51764	finden/ Finden/ FINDEN [NO]
49113	wurde/ Wurde/ WURDE/ wur-de [NO]
46459	Mercedes/ mercedes/ MERCEDES [Mercedes;Automarke]
45343	Golf/ golf/ GOLF/ GOlf/ gOLF [vw;Automodell]
44947	Euro/ EURO/ euro/ EURo/ EUro/ EU-RO [NO]
43915	OLDTIMER/ Oldtimer/ oldtimer/ OLDTiMER/ OldTimer/ old-timer/ Old-Timer [OLDTIMER;EDST]
42835	Suche/ suche/ SUCHE/ SUche [NO]
42576	Fahrzeuge/ fahrzeuge/ FAHRZEUGE/ Fahr-zeuge [ET]
39174	Motor/ motor/ MOTOR/ MOtor [ET]
38708	Autos/ autos/ AUTOS/ aut-os/ auto-s/ Auto-S/ Au-tos/ AUTOs [auto;=A]
38601	online/ Online/ ONLINE/ on-line/ ON-Line/ On-Line/ ON-LINE/ OnLine/ On-line/ ONline [NO]

Tabelle A.1: **Semantische Annotation im Automobilbereich**

Frequenz	Varianten [semantische Annotation]
38589	neue/ Neue/ NEUE/ neü/ n-e-u-e/ Neü [NO]
38105	Rover/ ROVER/ rover [Rover;Automarke]
37007	links/ Links/ LINKS [NO]
36675	Gebrauchtwagen/ gebrauchtwagen/ GEBRAUCHTWAGEN/ gebraucht-wagen/ GEBRAUCHT-WAGEN/ Ge
36474	Diesel/ diesel/ DIESEL/ DiESEL/ DIESEl/ DIesel [Diesel;EDST]
36253	KFZ/ kfz/ Kfz/ KfZ/ KFz/ k-f-z/ K-F-Z [Kraftfahrzeug;Abk.]
35304	Mazda/ MAZDA/ mazda/ MAzda [Mazda;Automarke]
35031	bitte/ Bitte/ BITTE/ BIT-TE [NO]
34748	Motorrad/ motorrad/ MOTORRAD/ Motor-rad [ET]
34192	gut/ Gut/ GUT [NO]
34158	LKW/ lkw/ Lkw/ LkW/ LKw/ lKW [Lastkraftwagen;Abk.]
33926	Beschreibung/ beschreibung/ BESCHREIBUNG/ BEschreibung [NO]
33491	kostenlos/ Kostenlos/ KOSTENLOS/ k-o-s-t-e-n-l-o-s [NO]
33334	Peugeot/ PEUGEOT/ peugeot [Peugeot;Automarke]
32645	kaufen/ Kaufen/ KAUFEN/ KAUFEn [NO]
32559	free/ Free/ FREE [NO]
32184	Forum/ forum/ FORUM/ FOrum [NO]
32128	Mini/ MINI/ mini/ mi-ni/ MiNi/ MI-NI [Mini;Automarke]
32112	Citroen/ CITROEN/ citroen/ Citrön/ citrön/ CITRÖN/ CItroen [Citroen;Automarke]
31815	rechts/ Rechts/ RECHTS [NO]
31678	gibt/ Gibt/ GIBT/ GiBT [NO]
31394	gebraucht/ Gebraucht/ GEBRAUCHT [NO]
31200	PKW/ Pkw/ pkw/ PkW/ PKw [Persoanlkraftwagen;Abk.]
30579	Berlin/ BERLIN/ berlin [NO]
30478	bin/ Bin/ BIN/ BIn/ biN [NO]
30362	http/ HTTP/ Http [NO]
30119	heute/ Heute/ HEUTE [NO]
29950	Stunden/ stunden/ STUNDEN/ STunden [NO]
29936	Porsche/ PORSCHE/ porsche/ POrsche [Porsche;Automarke]
29784	suchen/ Suchen/ SUCHEN [NO]
29748	car/ Car/ CAR [car;EDST]

Anhang B

Top40-Terme aller 20 Branchen im "Test2"

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

Altenpflege
4.8456526852437 ISBN [0.949553948223692]
2.3898116319469 Anmeldeschluss [0.990704063383647]
1.71704834317162 Verfügbarkeit [0.92884267789577]
1.64308405573677 Ill [0.952347539529034]
1.16132432914958 Darst [0.946140227352307]
1.00616935838106 Altenpflege [0.984807681986697]
0.961960818764825 Merken [0.901650970806495]
0.944674146222579 Veranstaltungsort [0.986073169846711]
0.840670722894639 Aufl [0.927779161780293]
0.716546746279823 Bemerkung [0.974591976763324]
0.654768146046878 Genehmigung [0.924503596711502]
0.556290790170968 Univ [0.921281365144664]
0.541412916261658 Pflegeheim [0.978148735077627]
0.524976407942801 Fortbildung [0.906560587956648]
0.468440487087422 Schulamt [0.986253073638118]
0.454979553550428 Staatliches [0.978621046057088]
0.40779543915233 Altenheim [0.977971619169755]
0.398018550583354 Grundschule [0.906369760920156]
0.396034834062113 Krankenpflege [0.946061994172703]
0.37435564636569 Lehrkräfte [0.975780660445135]
0.372088541769985 Ammersee [0.944717846676176]
0.357210667860675 Lehrerfortbildung [0.992402088561188]
0.336523338424872 Hauptschule [0.945158413669484]
0.305350650233937 Dillingen [0.96788422714287]
0.290047694212932 Einemillioneurohomepage [0.942054087387274]
0.286221955207681 Zugl [0.928040974374873]
0.279704029495031 Lehrgang [0.954471097032077]
0.260008558319849 Personalführung [0.989217993372249]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

0.25915839409646 Referat [0.923906744760122]
0.243146967889297 Kardinal-von-Waldburg-Str [0.98429441331284]
0.23691243025111 Altenheime [0.977693012416175]
0.231953138948007 Verl [0.941711514719408]
0.224159966900273 ALP [0.977570881027427]
0.222743026527958 Facharbeiter [0.963808635702427]
0.221326086155642 Seniorenheim [0.980241267065527]
0.221042698081179 Pflegewissenschaft [0.990175525716856]
0.213816302182372 Berufsausbildung [0.941985595944832]
0.207865152618647 Altenhilfe [0.983409844925071]
0.199363510384756 Assistent [0.93514538727273]
0.197096405789052 Detektive [0.959344081774199]

Autobranche

4.08552349200962 Editie [0.999589856486011]
2.90767894752603 Nederland [0.983303114768634]
2.02229919247777 Opel [0.982306715818979]
1.97538504536698 Autobedrijf [0.999754949656652]
1.9624087919108 BMW [0.952390086791542]
1.73682161644191 Audi [0.968819940787289]
1.53618877454259 Ford [0.935537633508786]
1.53419242785702 Volvo [0.981064802281897]
1.51722348102972 Renault [0.983247056443589]
1.41740614675144 Kia [0.991948054734816]
1.38247007975405 Mercedes [0.960330886976078]
1.3565175728417 Volkswagen [0.976473504901618]
1.33655410598605 Peugeot [0.98483579016833]
1.30261621233143 Fiat [0.976389531920508]
1.23873311839334 Toyota [0.949598550297755]
1.2177714781949 Alfa [0.979996137586984]
1.19082079793977 Jaguar [0.985265131561371]
1.12693770400168 Lees [0.98917871782985]
1.01214776958167 Rover [0.985481473600848]
1.00915324955332 Porsche [0.957899063107301]
0.891368795104958 Nissan [0.954346248215655]
0.822494834452951 Romeo [0.976655869160669]
0.819500314424603 Mazda [0.977221865391433]
0.816505794396255 Saab [0.990506284018457]
0.762604433885989 Chrysler [0.963249792464866]
0.754619047143727 Mitsubishi [0.95213529493372]
0.726670193545811 Seat [0.931643134679649]
0.691734126548416 Reacties [0.982149195930181]
0.684746913148937 Suzuki [0.954954104810533]
0.679756046435024 Hyundai [0.962817258867705]
0.665781619636066 Nederlandse [0.98327577547574 5]
0.620863819210844 Lancia [0.989789908922753]
0.599902179012407 Citroen [0.967469841133474]
0.597905832326842 Chevrolet [0.944264085353608]
0.569956978728926 Subaru [0.981832313865924]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

0.554984378587185 Daewoo [0.981844009486626]
0.546000818502141 AutoRAI [0.999737331133468]
0.543006298473793 Europese [0.984427453594863]
0.530030045017618 Ferrari [0.940353706126021]
0.522044658275356 Lexus [0.966696320639281]

Bank

17.6760187966059 Bank [0.979147377048655]
4.14539036117775 Banking [0.986434103030361]
2.17516774438631 Visa [0.966354389164016]
1.70236678566706 Finance [0.934538108688503]
1.55492802259958 PayPal [0.964151143874053]
1.43613750317439 Banken [0.914620493960335]
1.3765848355297 HSBC [0.995586658221456]
1.18675086941648 Fund [0.961179333318541]
1.17053785241013 Zinsen [0.935446847370244]
1.04634299398282 Konto [0.93045546968494]
0.771823560399522 Sparkasse [0.956255772850014]
0.76788836209701 Citibank [0.987107209032128]
0.728588848382581 Trust [0.927744277747207]
0.668091733145283 Volksbank [0.971172964734691]
0.661218253443561 Kreditkarte [0.93242103901659]
0.617353909698217 Treasury [0.984933809833141]
0.586711832249317 Securities [0.970138680042931]
0.582514287393304 Girokonto [0.940843046052307]
0.513202327958377 Dresdner [0.937008227520504]
0.48660038743339 MasterCard [0.959642345440266]
0.485708409151487 SWIFT [0.972471634645876]
0.457532389305495 Rewards [0.974505449254913]
0.41015260174324 Investments [0.929839682448203]
0.407424197586831 IBAN [0.991844858310262]
0.404538385498322 Bargeld [0.943881989250586]
0.396300703718395 Tan [0.957121110680269]
0.395618602679293 Asset [0.901644360021393]
0.395251317504392 Chase [0.950937394913255]
0.365133933162493 Raiffeisenbank [0.985295929953736]
0.353905500672657 ATM [0.961892314773286]
0.3366430974523 Kreditkarten [0.906958223400374]
0.325099849098262 Postbank [0.962966253706257]
0.304164594128894 Sparkassen [0.921174986778874]
0.300963966176184 Hypo [0.97280887213059]
0.288371331608142 Citigroup [0.988148262949183]
0.283544155023727 BIC [0.979547181197588]
0.277300307050406 DKB [0.990609716363589]
0.249071817893714 Banco [0.963262250427316]
0.249071817893714 Landesbank [0.970817484355283]
0.248442186165312 Online-Banking [0.967158666691551]

Bueroartikel

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

3.63992844716228 Kong [0.971030039276587]
3.63900795611099 Hong [0.971385634430895]
1.67744152580597 Bürobedarf [0.987146147186767]
1.17331926004793 Papier [0.94377181382949]
1.11072586856001 Büroartikel [0.993364609134401]
0.907297346224299 Drucker [0.925664915068351]
0.896865114309647 Toner [0.920835098789568]
0.710925921948495 Schreibwaren [0.962744899092555]
0.644650566255412 BELIEBT [0.951255549119951]
0.501360792604161 Tintenpatronen [0.983988378516566]
0.496144676646835 Tinte [0.971034730439315]
0.478655346672271 Büromaterial [0.989716421703836]
0.450120124082194 Bürotechnik [0.98893712704766]
0.427414678150304 Werbeartikel [0.948648331967115]
0.417289276586083 Druckerpatronen [0.973943131810824]
0.405936553620138 Kopierer [0.972211066733397]
0.404095571517552 Verzeichnisdienst [0.951063334247054]
0.370037402619718 Etiketten [0.963926709245255]
0.330763117764557 Visitenkarten [0.95290191143894]
0.330149457063695 Stempel [0.947171042932434]
0.314194278841286 Büromaschinen [0.981945813859711]
0.298545930969308 Werbemittel [0.91323112692374]
0.273079011883539 Druckerei [0.926114523961715]
0.266942404874921 Viking [0.925035095376466]
0.24730526244734 Informationsbeschaffung [0.950070013706182]
0.24730526244734 Kontaktaufnahmen [0.957570740543786]
0.235645709130964 Folien [0.947084592367938]
0.227054459318898 Erzeuger [0.913748511004333]
0.225213477216312 Tinten [0.98446100399841]
0.222452004062434 Laserdrucker [0.960807849623528]
0.220917852310279 Faxgeräte [0.974736024810894]
0.189314326215892 Kugelschreiber [0.950381970268507]
0.18870066551503 Briefpapier [0.958878652565208]
0.181336737104687 Bausuchmaschine [0.917751775729116]
0.180723076403826 Schreibgeräte [0.979116647497123]
0.176427451497792 Büroeinrichtungen [0.924335703326488]
0.172131826591759 Ordner [0.9336215476154]
0.168143032036157 EDV-Zubehör [0.990257058603843]
0.154028835916334 Patronen [0.968079365313676]
0.153108344865041 Lipstick [0.935221861288794]

Computer

28.5077290658699 Windows [0.990736816880187]
26.9180193414362 Memory [0.997539536341547]
20.1933758565122 Apple [0.995241815788886]
17.7807983396322 Mac [0.994648840535204]
17.5578500701767 Intel [0.998379294825267]
14.5613329149042 RAM [0.997677645332549]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

14.5572898115721 USB [0.994521162735822]
13.9021145430427 Desktop [0.99678187286489]
13.7775484546681 Microsoft [0.988196961165996]
13.6834019056494 Web [0.9305127586952]
12.8890283652578 Sony [0.989790339760437]
12.4508329755508 LCD [0.995790210720205]
11.8701663255699 Price [0.976541336681546]
11.4400571425271 Digital [0.98139018024406]
11.2232697924348 Notebook [0.992702878580836]
11.210177838788 Wireless [0.980962642360191]
10.5051376291627 Search [0.938786854707063]
10.3892353336427 Dell [0.994353128500771]
10.2544652225728 News [0.903770801159901]
9.85073247555366 Business [0.906392087745741]
9.84495661365067 DVD [0.961196798091211]
9.16667789750926 Pentium [0.99867288255696]
8.85478135474766 CPU [0.998070501074379]
8.40811470091624 Toshiba [0.996013722550408]
8.37808021902068 AMD [0.997551121791659]
8.27315206111632 Security [0.973110149992309]
7.9781980466035 Compaq [0.997351248943719]
7.73926989254972 Click [0.968311683219996]
7.53730725467508 Media [0.964546034697776]
7.53287909388278 Server [0.975314041344513]
7.51227851976211 Vista [0.992539494441655]
7.03211186689334 Reviews [0.962039708122894]
7.00072968388708 Core [0.995805962875749]
6.89137336519042 View [0.940184845492924]
6.73966072587182 Services [0.919718942340174]
6.599114752899 Contact [0.945782567957839]
6.51382452546481 Laptops [0.997246387353274]
6.32822682964865 Monitors [0.998759238631016]
6.22676418888609 Mobile [0.962649339646269]
6.22175844190349 Read [0.97062537151411]

Finanzen

2.53167562531517 Finanzen [0.935436849405783]
2.20706123916539 Abs [0.920284287473113]
1.21535354814492 Steuern [0.921485190995534]
0.941732226486092 UStG [0.927807656342386]
0.927146627108411 Aktien [0.90724864901253]
0.866316763037151 Steuerberater [0.950326948569824]
0.85489703019106 Steuer [0.925870467224891]
0.825047431464643 Unternehmer [0.908635931814038]
0.73041063550248 EStG [0.988137639619823]
0.72995836885511 Bundesministerium [0.945533723744491]
0.605132774181002 Umsatzsteuer [0.965692304294659]
0.549843176540024 Steuerberatung [0.955797797707966]
0.518184511224128 Steuerrecht [0.955638366210545]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

0.495910378841157 Finanzamt [0.950217356209707]
0.427618115088294 CDU [0.90133823343097]
0.402065049511891 Mrd [0.906550945282745]
0.392115183269752 Kanton [0.923713743472872]
0.370971717505207 Anwaltskanzlei [0.960502282098223]
0.368484250944672 Besteuerung [0.940428232544498]
0.354803184861731 Buchhaltung [0.941531370803707]
0.343044252030112 IHK [0.92130732397634]
0.334677319053768 Einnahmen [0.924997397641386]
0.333320519111658 BFH [0.983588012071939]
0.332868252464288 Signatur [0.928406316537259]
0.329250119285328 Ministry [0.907775478951954]
0.316360519835284 Einkommensteuer [0.966844307871329]
0.314890653231332 Jiangsu [0.979890246499219]
0.312629319994482 Steuerkanzlei [0.978409692024412]
0.302905587076028 DAX [0.909062008152738]
0.300644253839178 Rechnungen [0.933174786278425]
0.298835187249699 ORF [0.908423030308022]
0.291259720906252 Einkünfte [0.951647163844354]
0.290694387597039 Aufwendungen [0.918247090493177]
0.290694387597039 LSA [0.994382462432844]
0.287867721050977 Bundesrat [0.902066710971435]
0.277239454837783 Vorsteuerabzug [0.948868453193152]
0.271925321731186 Umsätze [0.920622137141362]
0.271699188407501 BGBl [0.950039541087525]
0.271246921760131 Hartmann [0.920678313822447]
0.269663988494336 Finanzverwaltung [0.982364345149832]

Gesundheit

0.536702228768748 Yoga [0.924445554093452]
0.493913809038977 Übergewicht [0.922845332228433]
0.365471592260221 Immunsystem [0.929270879494615]
0.357698875726576 Reiki [0.954379083319989]
0.348617880172416 Gesundheitsförderung [0.964258811383275]
0.310831703756377 Übungen [0.909179287608885]
0.302905072043847 Muskeln [0.926692217653354]
0.266735005006091 Diäten [0.912933081283023]
0.265041938038366 Meditation [0.908293797319192]
0.2511895719388 Bodybuilding [0.916108048915714]
0.232411920114944 Stoffwechsel [0.918764168371441]
0.218944341962588 Muskulatur [0.958197894625505]
0.213634268291087 BMI [0.921361996186507]
0.209632473640102 Hypnose [0.903402360003284]
0.192470931194528 Fasten [0.905425483405459]
0.190777864226803 Ausdauer [0.952487688465251]
0.185006045018651 Gong [0.926616560879302]
0.182774274924832 Shiatsu [0.941933926013405]
0.177695074021657 Trampolin [0.976462405416108]
0.173539364191788 CIS [0.926477474032882]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

0.171846297224063 Nahrungsergänzungsmittel [0.908176476490451]
0.169922357488012 Trainingsplan [0.986032088688665]
0.154376924420721 Gelenke [0.924305826797841]
0.154069094062953 Dissertation [0.921299283731051]
0.151914281558576 Vitalität [0.929242694882147]
0.150836875306387 Aerobic [0.900007419624757]
0.149682511464757 Qigong [0.955747055686795]
0.148143359675916 Muskelaufbau [0.915403262623512]
0.145757674403213 Rückenschule [0.982056009332616]
0.138292788227336 Kursort [0.995335876183321]
0.135291442239096 Magnetfeldtherapie [0.958499934456131]
0.134060120808024 Kreatin [0.96589090807905]
0.132367053840299 Krafttraining [0.964571472362746]
0.129134835083734 LMU [0.961295941006173]
0.128519174368197 Strahlung [0.914671792071634]
0.122670397570603 Raucherentwöhnung [0.958165931865448]
0.122131694444509 Beweglichkeit [0.956429679632865]
0.121285160960646 Mnchen [0.985708547765342]
0.119053390866827 Bildungsinstitution [0.996881463947012]
0.117899027025197 Fakultt [0.99635679352702]

Haushaltsgeraete

2.17090962679144 Restzeit [0.958830173897516]
2.15064657554434 Entfernung [0.941842668471773]
2.11748885532181 PLZ [0.922300388044883]
2.11748885532181 Standort [0.933153424498976]
1.35946652912353 Gebote [0.937620364823743]
1.22315145709759 Haushaltsgeräte [0.971808308019598]
1.06749438160852 Coffee [0.900002084284791]
1.03341561360203 Tefal [0.997416658402158]
0.936705596286335 Belgium [0.925825874022475]
0.906311019415687 OVP [0.963431814162641]
0.88512692038463 Bosch [0.934434958134465]
0.827100909995211 Sofortkauf [0.978940252820022]
0.806837858748112 AEG [0.972931870087815]
0.802232619828317 Miele [0.981250943292615]
0.739601370519103 Vacuum [0.955845595287108]
0.655786022178831 Devil [0.975442210485032]
0.63736506649965 Dirt [0.968433710512996]
0.629996684227978 Staubsauger [0.967746245804435]
0.627233540876101 Whirlpool [0.949800309477362]
0.594075820653576 Tassimo [0.992722319996685]
0.550786574807501 Appliances [0.938236052577227]
0.527760380208525 Waschmaschine [0.915758161878397]
0.507497328961427 Espresso [0.940878399572061]
0.474339608738901 Cleaner [0.918024041901318]
0.468813322035147 Gewählte [0.980395659313269]
0.459602844195557 Normalbetrieb [0.998579343183212]
0.454997605275762 Krups [0.990538168943796]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

0.444866079652212 DKK [0.909604323200892]
0.442102936300335 Geschirrspüler [0.949536332967301]
0.422760932837196 Bauknecht [0.986944686185692]
0.417234646133441 Kenwood [0.969618139542743]
0.387761117046752 Guinea [0.924379541796128]
0.376708543639244 Swirl [0.98526492065286]
0.344471871200678 Wasserkocher [0.982282793018606]
0.33341929779317 Jura [0.904419737077351]
0.324208819953579 Gaggia [0.99526400876741]
0.319603581033784 Household [0.909109722257058]
0.292893195298972 Düse [0.988236850006624]
0.272630144051873 Rowenta [0.98977531298878]
0.268945952916037 Vorwerk [0.988241478302501]

Hotel

18.4529727011987 Hotels [0.931932986289389]
5.88241175596769 Inn [0.955073576399204]
4.5093300423937 Airport [0.949847585120679]
3.98569355410053 Holiday [0.928654353103781]
3.67673945585227 School [0.906362078867962]
2.31579879560751 Suites [0.951220695679367]
2.16082182075806 Resort [0.905170057726445]
1.78030692582816 Israel [0.937869330316867]
1.28595180144664 Budapest [0.941786874538128]
1.23910161918801 Lake [0.905952381624015]
1.20839189606116 Plaza [0.934036430120789]
1.19196576508633 Reservations [0.94328373237879]
1.02984699416087 Hilton [0.930128340692073]
0.992995326408648 Comfort [0.902405981073135]
0.97599785174774 Downtown [0.956519814838619]
0.954143955755145 Dusseldorf [0.977507564341341]
0.90186600847874 Resorts [0.932976200915737]
0.891010478443203 Check-in [0.934045047180655]
0.878869425113983 Jerusalem [0.975805217058927]
0.851302092260578 Iraq [0.956507476514983]
0.840303726303521 Check-out [0.905065255479678]
0.80930833133363 Rome [0.922813642236972]
0.791596677065122 Virginia [0.915460988712405]
0.770885468444688 Israeli [0.984130188874722]
0.768457257778844 Floors [0.926520671629945]
0.744460822963445 Brussels [0.954805272148026]
0.718893193011324 Hyatt [0.974873671759172]
0.668472112714682 Sheraton [0.960973347854741]
0.641190451704318 Jewish [0.95512472866758]
0.601339229600174 Replies [0.946165296966156]
0.594483105367203 Marriott [0.928177392203225]
0.586198621919029 Academy [0.935442513360733]
0.571629357923966 Wisconsin [0.938894386616624]
0.533206495035023 Committee [0.919055573539465]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

0.529492761075497 Iowa [0.951148277200868]
0.526350370802052 Arab [0.926060216568536]
0.520779769862763 Palestinian [0.987441341423794]
0.509495732062665 Cologne [0.900957751668288]
0.499211545713208 Kansas [0.924728659928985]
0.494640796224561 Minneapolis [0.969999742149166]

Immobilien

21.6792337548429 Immobilien [0.984389119872215]
4.55006206221628 Wohnung [0.961391980021444]
3.98231095533796 Wohnungen [0.978373981862977]
3.15107991757718 Häuser [0.976190006411321]
3.02975660056618 Immobilie [0.973931519659812]
2.54498854168673 Verkauf [0.905201033363109]
2.29954079212046 Kauf [0.950961899950043]
2.29428870047496 Immobilienmakler [0.988080508231683]
2.25087140953885 Grundstück [0.977191276130854]
2.23091346128596 Makler [0.977692821784754]
2.04306365009865 Vermietung [0.956776377973994]
1.94047279328993 Grundstücke [0.989578085098474]
1.87464657799969 Miete [0.979233693434673]
1.82615226513959 Gewerbeimmobilien [0.9942051722871]
1.71848438640689 Bauen [0.928556580204863]
1.56092163704195 Einfamilienhaus [0.987903554118167]
1.37552280195588 Bau [0.917188322599947]
1.29866719421009 Wohnfläche [0.988467962212462]
1.29569100894431 Eigentumswohnung [0.986866530700985]
1.11344342884554 Eigentumswohnungen [0.992240021220835]
1.09313534114961 Hausverwaltung [0.987344073782777]
1.00507527122677 Immo [0.992546016844661]
0.947652402569323 Bauträger [0.987441956324827]
0.942575380645342 Mietwohnung [0.986311588705689]
0.941875101759276 Gewerbe [0.919239255004887]
0.865544703178041 Einfamilienhäuser [0.994993000012232]
0.857666565709794 Mietwohnungen [0.98445173805753]
0.849263219076997 Villen [0.972706739118985]
0.811973368393963 Alanya [0.967285993099765]
0.807071416191498 Objekte [0.963858120859395]
0.756301196951686 Mieter [0.948133028812458]
0.747897850318889 Objekt [0.957855894638844]
0.730040738724197 Immobilienmarkt [0.987913810492405]
0.718486137104102 Fertighaus [0.980721697023546]
0.706581396040973 Neubau [0.970208140822901]
0.667540848142773 Immobilienangebote [0.993665097013511]
0.663339174826375 Architektur [0.902841212374334]
0.631126346067321 Wohn [0.942809329934678]
0.606091375890448 Reihenhaus [0.992123764844298]
0.603290260346183 Hausbau [0.954366165563823]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

Kleidung

3.01904096340186 Kleidung [0.969371421117528]
1.55847952630873 Schuhe [0.941380564333162]
1.41937843706176 Jeans [0.961041089798222]
1.24356640775442 Damen [0.931931200571938]
1.17368212731221 Bekleidung [0.948461382674139]
0.980248846775878 Hose [0.97372098189292]
0.964630010459281 Herren [0.916940078666386]
0.902221412356635 Hosen [0.977464393957756]
0.818119986036493 Mädchen [0.916889912117755]
0.773532880654069 Shirt [0.911957728068981]
0.753308490038987 Jacke [0.981679272438059]
0.738490619687343 T-Shirt [0.942503981795832]
0.728545292289167 Unterwäsche [0.970237531680079]
0.718866953545976 Dessous [0.94230002561943]
0.708788131820309 Damenmode [0.973209336375855]
0.680287092900705 Teens [0.90137103444598]
0.644043382986548 Kleider [0.963612513576405]
0.630827444564812 Kleid [0.981436098829236]
0.628758282488681 Pullover [0.968415841604456]
0.60886762769233 Herrenmode [0.974851916279271]
0.604996292195054 Jacken [0.976548217426018]
0.586106844854895 Hemd [0.980596242529827]
0.573157895088143 Baumwolle [0.962650984233088]
0.568619087953405 T-Shirts [0.923734482379504]
0.523431258097265 Hemden [0.975471356468082]
0.519493175436243 Wäsche [0.935563570835408]
0.514019908009059 Shirts [0.924591798474035]
0.480379337481002 Kindermode [0.970796150250866]
0.476174266164995 Latex [0.946553781619115]
0.455215656748706 Socken [0.963877238093255]
0.446071295315484 Anzug [0.979078257782043]
0.42691485932034 Größen [0.903671210916532]
0.412630966278665 Klamotten [0.974136421480157]
0.396611646979591 Taschen [0.901701748804331]
0.394942967885937 Robin [0.939115314601882]
0.35109008130472 Bluse [0.982023995715683]
0.322922778203847 Handschuhe [0.964100227762269]
0.309573345454618 Berufsbekleidung [0.970064862224603]
0.290283415131983 Stiefel [0.943806559682201]
0.288614736038329 Blazer [0.968883408813661]

Kosmetik

4.67366937740478 Kosmetik [0.975886212360932]
2.71906039970251 Aloe [0.983188159444469]
2.69356458006481 Det [0.959587762320123]
2.65198840945228 Beauty [0.927729037949565]
2.16587516781222 Vera [0.974655514186448]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

1.4452215438618 Parfum [0.985677369276658]
1.2417839456738 Gel [0.976503558070402]
1.16402698536794 Make-up [0.979687383719516]
1.0798167263156 Cosmetics [0.984744533327737]
1.00364664275067 Naturkosmetik [0.980504527554296]
0.9185900494365 Eau [0.980167875439261]
0.894469522974273 Spray [0.969157591753786]
0.84591109470163 Lotion [0.98596324571917]
0.829090201247708 Nail [0.986157205457413]
0.777146435927736 Creme [0.945217608586345]
0.71197869776663 Shampoo [0.98422052478917 8]
0.699601059187329 Körperpflege [0.955643268987326]
0.622584641360567 Aveda [0.997862440649391]
0.575401506263402 Indonesia [0.909419563333237]
0.575295714480673 Shave [0.990942562701725]
0.556782152503086 Cream [0.919778892045223]
0.553608399021214 Fußpflege [0.967181304502992]
0.50864689136136 Toilette [0.947276307886847]
0.471090808492541 Duft [0.911389932217398]
0.441469109328402 Shiseido [0.997992923172011]
0.436496895540136 Gesichtspflege [0.983816178193694]
0.410260533423327 Islam [0.901069367824809]
0.408462073116933 Friseur [0.931545497178493]
0.407086779941455 Yves [0.972536898272379]
0.383072045261956 Nagelstudio [0.984732992594972]
0.365510609328931 Lavera [0.993122342976439]
0.363606357239808 Haarpflege [0.971487002707254]
0.361702105150685 Mlm [0.93649936927089]
0.340755332170329 Rocher [0.980663515883854]
0.333349907379294 Duschgel [0.986990771870949]
0.311027841223461 Maniküre [0.983133686063865]
0.302987665736052 Jakarta [0.948867851367937]
0.29865020264416 Logona [0.996307718435952]
0.296428575206849 Vichy [0.980852044346805]
0.294735906683184 Nails [0.972266618482707]

Lebensmittel

3.37125986073357 Lebensmittel [0.976980527304031]
2.38115599270442 Ernährung [0.911209011457825]
1.98294147417178 Fleisch [0.971256163898928]
1.96002013152785 Gemüse [0.970774967159187]
1.31920387938467 Obst [0.959996681993793]
1.22450438729919 Milch [0.963422677783461]
1.21139646352422 Rezepte [0.93420334283575]
1.18763397069152 Fett [0.952208399598636]
1.13422093370475 Zucker [0.973755615056055]
1.08059760917078 Salz [0.967578908849982]
1.04386738426124 Kochen [0.935291016625047]
0.966341375196444 Fisch [0.947983331046629]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

0.853136578958067 Brot [0.957980776012594]
0.790400794045778 Zutaten [0.977576032086752]
0.769652422722831 Käse [0.960841682470695]
0.747992805362052 Eier [0.970986259013102]
0.746871271777028 Rezept [0.939633746622409]
0.729207117812897 Nahrung [0.928408070670079]
0.710982197056255 Lebensmitteln [0.973601190309972]
0.670116317051937 Butter [0.95753399585215]
0.669064879315977 Getränke [0.905467777121922]
0.668083537429081 Eiweiß [0.956958051784317]
0.665560086862776 Reis [0.948288791585597]
0.616422896668905 Kartoffeln [0.974369557544548]
0.581164684589708 Mehl [0.98644175025798]
0.570299827984786 Pfeffer [0.974425337470868]
0.551444044586568 Gramm [0.958599676205852]
0.543172734397015 Nahrungsmittel [0.95190671616878]
0.519199954017123 Gentechnik [0.964068374070986]
0.519199954017123 Nudeln [0.978410148627623]
0.48260992080571 Schokolade [0.93853564743619]
0.46985247627606 Salat [0.9636113527435]
0.466838354766307 Früchte [0.94292740241806]
0.464945766841579 Gewürze [0.962651722507564]
0.45779599023705 Tomaten [0.952501029426906]
0.453029472500697 Wurst [0.970138558216153]
0.449594775896561 Kalorien [0.961432467325648]
0.422257394761597 Olivenöl [0.932972243953528]
0.418402123063076 Stichworte [0.930592688050697]
0.413285126081404 Kuchen [0.93640617738965]

Moebel

10.0391498881432 Möbel [0.98546200292432]
2.58102061793337 Stühle [0.993994183637347]
2.50090694721567 Tisch [0.970732674039253]
2.25603119898422 Holz [0.947345637060766]
2.21143962754701 Tische [0.993488145912055]
1.81616179938328 Bett [0.951278417456632]
1.78819759356672 Sofa [0.983371560064824]
1.73907128605115 Gartenmöbel [0.985569890121091]
1.66273656206542 Betten [0.969904636113706]
1.61814499062821 Sessel [0.991214278331035]
1.43070923272266 Stuhl [0.978329774995291]
1.41559344579479 Kinderzimmer [0.986311664032742]
1.11554507527662 Kindermöbel [0.99500553201135 7]
1.10496402442711 Schränke [0.992538772421514]
1.08758086946006 Massivholzmöbel [0.99754347743879]
1.03921035129089 Büromöbel [0.953331374003054]
1.01275772416712 Schrank [0.983316612407918]
1.00217667331761 Polstermöbel [0.988763743751532]
0.999153515932039 Couchtisch [0.997481016858841]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

0.962875627305158 Teak [0.99433563631074]
0.897121954168934 Matratzen [0.981875890758948]
0.834391438418284 Antiquitäten [0.949075296100593]
0.828345123647137 Küchen [0.958155799634971]
0.814740915412056 Wasserbetten [0.98515008707523]
0.795846181752222 Wohnzimmer [0.920879627179533]
0.754277767700587 Buche [0.979786076016692]
0.753521978354193 Sofas [0.992167021070474]
0.700616724106657 Esszimmer [0.96799816333296]
0.69457040933551 Massivholz [0.995313988273851]
0.656780942015841 Regale [0.983874959370119]
0.622014632081746 Schreibtisch [0.964514089378323]
0.607654634500272 Designermöbel [0.990444250353819]
0.606898845153879 Regal [0.968112428389373]
0.585736743454864 Eiche [0.977617336876634]
0.572132535219784 Esstisch [0.979367590432349]
0.563063063063063 Lampen [0.931515997319759]
0.529808331821755 Kommode [0.991307193356588]
0.528296753128968 Sideboard [0.995434687158761]
0.527540963782575 Möbelhaus [0.991377391139171]
0.51091359816192 Rattan [0.991884863459279]

Reisen

22.5306552403316 Reisen [0.985538998075386]
16.0728373996675 Urlaub [0.974042190775105]
9.19388376520418 Lastminute [0.993789207282407]
9.05178609637404 Last [0.942920326868428]
8.98814389720688 Reise [0.972160234537843]
8.45760935759784 Minute [0.981045452401361]
6.8522897475709 Yachtcharter [0.989873713648325]
6.356429233715 Kroatien [0.988779225384447]
5.88010029132765 Ferienhaus [0.976899749981437]
5.73043139535522 Türkei [0.976445037253619]
5.40651454718027 Spanien [0.957984469671972]
4.93819574365361 Mallorca [0.972655011961654]
4.41929236458202 Ferienwohnung [0.967958163437721]
4.40129698412786 Flug [0.984520577887857]
4.3824237802369 Angebote [0.902106368204125]
4.34358009315901 Flüge [0.988838547636646]
4.10920123553649 Italien [0.944110651486983]
3.64845365915213 Griechenland [0.977314178742275]
3.41769582320636 Ferienwohnungen [0.96391089116109]
3.3918000318211 Ferien [0.970910844089368]
3.19900805951643 Pauschalreisen [0.993388890045584]
3.07699060180283 Reisebüro [0.987014210469164]
3.03551344441457 Mietwagen [0.98218158031824 8]
2.95233967377887 Ferienhäuser [0.972431062775966]
2.68372570212159 Ägypten [0.987164641924719]
2.60483132074022 Ostsee [0.976074033842573]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

2.39711635001015 Costa [0.934179010435195]
2.33325469498379 Charter [0.983368381761618]
2.22264894194844 Tourismus [0.962152998813731]
2.21398043551015 Yacht [0.988620045148346]
2.20311737047989 Insel [0.949856006630479]
2.18632899725132 Inseln [0.981540167713582]
2.16822388886755 Teneriffa [0.978632897464642]
2.06705473777768 Australien [0.95429672586095]
1.99375648080584 Karibik [0.988181721738444]
1.94953612517762 Kanaren [0.99360773953785]
1.94876802967043 Reiseveranstalter [0.988770274169359]
1.89906127756228 Tunesien [0.992861671549259]
1.89423324865995 Thailand [0.932171151686761]
1.8751405889098 Unterkunft [0.958326817798295]

Restaurant

49.5950085604826 Hotel [0.921494403256611]
7.32439494406895 Restaurant [0.957084157923898]
2.26385736111929 Restaurants [0.931949344506381]
2.23903324128789 Centre [0.920478333716529]
2.12361388000993 Town [0.932250153785503]
1.82226953875762 Munich [0.958043332634523]
1.12284356453887 Hof [0.904231933397672]
1.02380300397442 Pensionen [0.90609904653879]
0.994116427681197 Cafe [0.930081554420986]
0.839669455328101 Mercure [0.953106679862819]
0.83391128320226 Ibis [0.957979698683112]
0.712989668559612 Lodge [0.933412385130338]
0.674090016865046 Nearby [0.929302255446957]
0.634550568267609 Bistro [0.972067309959621]
0.623801980299374 Castle [0.938955535484711]
0.608702773391615 Tulip [0.964529162906274]
0.569291284174752 Vegetarian [0.980592576444573]
0.55956637125111 Thai [0.92852483937221]
0.548945742663449 Cathedral [0.97508966653769]
0.5486898239023 Gasthof [0.935899894273138]
0.548433905141152 Ramada [0.958164658274184]
0.541268179828995 Steigenberger [0.960890780965102]
0.539220829739807 Parkhotel [0.951983297386659]
0.52962387619674 Novotel [0.94911918352066]
0.52962387619674 Prague [0.907474311196963]
0.495714640344569 Catering [0.903315774461374]
0.48394237733174 Theatre [0.908211672680115]
0.46308499829814 Avenue [0.900876633841041]
0.429687599968266 Gardens [0.93392028387608 5]
0.425081062267594 Hostel [0.931843695430289]
0.413820636777061 Ave [0.914586953809994]
0.399745104913896 Rhön [0.921919690290279]
0.390532029512552 Trail [0.929184706855022]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

0.379655482163742 Pub [0.911883558332552]
0.371977919329288 Garni [0.937010224462291]
0.350736662153966 Ostseebad [0.943301954051308]
0.34664196197559 Minotel [0.944853798714284]
0.346258083833868 Courtyard [0.936239746913258]
0.336917049051949 Motel [0.909670246384697]
0.334869698962761 Stralsund [0.943212760292422]

Schmuck

5.23812679913946 Schmuck [0.972137431215577]
2.75941195121208 Gold [0.91402025629692]
2.11391609180784 Silber [0.97287739037992]
1.95449016843618 Ring [0.912088328825923]
1.64405744348888 Uhren [0.943974117168756]
1.05471891776668 Piercing [0.977725752867927]
0.983783338378835 Anhänger [0.958316353656124]
0.95861553811623 Kette [0.977591828039233]
0.854361732046223 Swarovski [0.985879242700639]
0.83752348845416 QVC [0.990269559267013]
0.83447827418751 Silberschmuck [0.988685528638553]
0.81378873019939 Ringe [0.973311616794978]
0.776350507744696 Armband [0.986778769402551]
0.736494026901781 Perlen [0.977077927279865]
0.664662796259044 Ohrringe [0.989281093997866]
0.646033250157187 Sterling [0.954587055102969]
0.527717719385297 Edelsteine [0.971256254894063]
0.476665597856169 Ohrstecker [0.992741624697757]
0.473978644091478 Juwelier [0.98359597483858]
0.472008211330705 Ketten [0.966705852705878]
0.466186478173875 Diamanten [0.974739028918641]
0.434211728374053 Collier [0.986500168072613]
0.409581318864386 Ohrschmuck [0.992777543820692]
0.401341327319334 Titan [0.959710801483323]
0.367754405260697 Modeschmuck [0.96528475563816]
0.360141369594073 Armbänder [0.983980887470088]
0.357006590201934 Goldschmuck [0.986114656282537]
0.327987489543272 Diamant [0.961262560959184]
0.327808359292292 Tattoos [0.957080561737198]
0.326464882409947 Trauringe [0.985528841081403]
0.310074464445332 Goldschmiede [0.981770887141097]
0.305954468672806 Kristall [0.958335888368764]
0.296818825872857 Tchibo [0.937350388753967]
0.270934504606334 Platin [0.9774531358938]
0.267441464712236 Steine [0.919353783084456]
0.257947561410328 Zirkonia [0.98660017972254]
0.257678866033859 Schmuckstücke [0.989326404188709]
0.25749973578288 Strass [0.967149711116231]
0.255171042520148 Fossil [0.953002554175583]
0.252842349257416 Weißgold [0.991149384825327]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

Versicherung

10.6011953714714 Versicherung [0.984956631105957]
7.55291954552836 Krankenversicherung [0.982286870711837]
5.81643557199206 Versicherungen [0.974075984475858]
4.98328994306009 Vergleich [0.950855447053427]
2.98515476446153 Altersvorsorge [0.982890241092975]
2.69285972457216 Lebensversicherung [0.985977041128417]
2.66915658737684 Krankenkassen [0.970429143192329]
2.46014991906885 Rentenversicherung [0.986034088616657]
2.34556292986491 Rente [0.97390392714576]
2.29907335138788 Versicherungsvergleich [0.987876580235702]
1.74499614987716 Autoversicherung [0.991860423384201]
1.65673142905036 Krankenkasse [0.969041687204333]
1.6093251546597 Unfallversicherung [0.98766861861378]
1.44942719601054 Haftpflichtversicherung [0.990053199279937]
1.33405446745206 PKV [0.988941029082691]
1.29136262918865 Vorsorge [0.956746872986378]
1.23858713587529 Versicherer [0.987349171416092]
1.19720486320278 Haftpflicht [0.98532218220792]
1.18031146708014 Allianz [0.971282405039701]
1.1285836262395 Versicherungsmakler [0.976179011657321]
1.10422570624872 Berufsunfähigkeitsversicherung [0.986779671392018]
1.09715405205785 Rechtsschutzversicherung [0.9933145174995]
1.06651021723074 Unfall [0.963497592881756]
1.03665212175817 Berufsunfähigkeit [0.988255888334443]
1.02133020434461 Hausratversicherung [0.987308763720765]
1.01556811574465 Versicherten [0.971076060972378]
0.987150542422068 Krankenversicherungen [0.976131442400953]
0.957161490390408 Kfz-Versicherung [0.99149000380145]
0.934113135990529 Rechtsschutz [0.985725417050085]
0.919576957931515 Versicherungsschutz [0.987394517148197]
0.864968072790894 Arbeitnehmer [0.921500543224847]
0.813240231950257 Kündigung [0.912326108985889]
0.739773602300645 Versicherte [0.97227182520496]
0.733749600582495 Riester [0.974540093460985]
0.721177770909834 Commented [0.976253366126557]
0.705332027259917 Arbeitgeber [0.911478533254111]
0.691581588555444 Riester-Rente [0.986306302202122]
0.645222966637507 Versicherungsnehmer [0.990085794007683]
0.62885339675123 Schäden [0.901936336683129]
0.618769741701283 Versicherungsvergleiche [0.985955739676004]

Wein

8.92803737136743 Wein [0.985543018302143]
4.10971310146124 Weine [0.987141943203397]
3.27494408195416 Weingut [0.994265178714841]
2.78621404290631 Sauvignon [0.997505018037237]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

2.75630683948256 Riesling [0.998008037860744]
2.56728067699799 Cabernet [0.9973765489015]
2.41132092956643 Bordeaux [0.98672723704802]
2.27421114485617 Blanc [0.99162785025845]
2.1486851361199 Pinot [0.997441830548847]
1.7711593464223 Chardonnay [0.997243724628574]
1.72240218027725 Chateau [0.983961688023868]
1.68375449134587 Merlot [0.99746053052216]
1.63067973597415 Rotwein [0.987968472145468]
1.52642575220829 Domaine [0.989550931938164]
1.20292248137118 Noir [0.991114415627002]
1.07465848922288 Winery [0.991658228137533]
1.06391716968336 Weißwein [0.988407301519471]
1.01115833547458 Cru [0.994660599924227]
0.969035513750995 Champagne [0.982488240139341]
0.961137484677824 Winzer [0.98985943258178]
0.939233617381561 Vino [0.984306154435128]
0.89711079565798 Vineyards [0.986835017588281]
0.737149380162678 Weinbau [0.983585221577246]
0.733884861479101 Shiraz [0.99571394761031]
0.729040736980889 Vineyard [0.984558124624269]
0.671753699436818 Rebsorten [0.99639653285729]
0.66554058323259 Syrah [0.996027900486355]
0.65142943795519 Sekt [0.969787971763311]
0.649955139194864 Ros [0.984670285611462]
0.646164085239742 Chianti [0.984874316730121]
0.641635881904457 Jahrgang [0.9732260302734]
0.63931912670966 DOC [0.945084828789496]
0.581189632731117 Loire [0.980588812804698]
0.576977350558759 Spätlese [0.997991741522808]
0.574976516526889 Rheingau [0.989903904452025]
0.566446645127864 AOC [0.980758474006771]
0.559180458380546 Rioja [0.990389706019316]
0.555705325588351 Trauben [0.984694798349603]
0.551282429307374 Reserva [0.986079212359352]
0.546754225972089 Weinen [0.973043845317644]

Aerzte

5.63947907701879 Ärzte [0.967182834630622]
4.91739383573665 Patienten [0.973003710176192]
4.12837151651821 Arzt [0.963854792663603]
3.96980447836891 Medizin [0.946341066154551]
2.73624975639987 Behandlung [0.946819721792177]
2.5659414948151 Therapie [0.958990600468905]
2.32337023950885 Klinik [0.971275391299761]
1.77310630406864 Praxis [0.904179388405206]
1.74938177011653 Chirurgie [0.97569270525347]
1.52708530785609 Presseportal [0.960820598504029]
1.41711725124645 Prof [0.915667555250156]

Tabelle B.1: **Top40-Terme aller 20 Branchen im "Test2"**

1.22387123872328 Psychotherapie [0.964036644874641]
1.19905731290603 Erkrankungen [0.953243371145401]
1.11347952972164 Krankheit [0.912247088004395]
1.11099813713991 Kliniken [0.95892213449465]
1.09610978164956 Diabetes [0.926076682796255]
1.08370281874093 Homöopathie [0.94999874216522]
1.06512263526313 Medikamente [0.927767212424714]
1.06354906923569 Krankenhaus [0.923841947487994]
1.05434976015222 Facharzt [0.98043118121237]
1.03764575155329 Patient [0.954750564185284]
1.01446591353376 Krankheiten [0.907358352561351]
0.959814754965509 Innere [0.950051870082834]
0.945047443015727 Psychiatrie [0.978209613903735]
0.882770541391444 Studien [0.913637096570809]
0.870726709104532 Krebs [0.934831657575274]
0.838892257934101 Schmerzen [0.934545980520458]
0.822732945267742 Diagnose [0.946563972975951]
0.796284931652765 Ärzten [0.963503179549302]
0.761727000819465 Erkrankung [0.949979107779256]
0.760456043643459 Diagnostik [0.966727038668522]
0.738244553948501 Symptome [0.950031154708001]
0.703202448952913 Plastische [0.97565961114255]
0.699268533884324 Klinische [0.974400578089282]
0.653393032007543 Neurologie [0.979009742382554]
0.634328674367457 Betroffenen [0.901103054725739]
0.612177706442785 Orthopädie [0.972797104502158]
0.611875097591355 Allgemeinmedizin [0.976783706394343]
0.596321002627855 Störungen [0.952947429415262]
0.587847954787817 Gesundheitswesen [0.905700457438071]

Literaturverzeichnis

[Arp95] Antti Arppe. *Term Extraction from Unrestricted Text. NODALIDA-95.* Helsinki, [www2.lingsoft.fi/doc/nptool/term-extraction.html -28.11.2006-], 1995.

[Asm05] Jørg Asmussen. Automatic detection of new domain-specific words using document classification and frequency profiling. *In: Proceedings of the Corpus Linguistics 2005 Conference, Vol. I, Birmingham*, 2005.

[BJL01] Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme. *Recent Advances in Computational Terminology.* John Benjamins Publishing Company, Amsterdam, Netherlands, 2001.

[Bla97] Ingeborg Blank. *Computerlinguistische Anaylse mehrsprachiger Fachtexte.* CIS, Dissertation, Oettingenstr.67, 80538 München, 1997.

[Bou92] Didier Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. *In Proceedings of Coling92*, pages 977–981, 1992.

[Bußmann90] Hadumod Bußmann. *Lexikon des Sprachwissenschaft.* 2. völlig neu bearbeitete Aufl., Kröner, Stuttgart, 1990.

[Dro84] Günter Drosdowski. *Grammatik der deutschen Gegenwartssprache. - DUDEN Bd.4.* DUDEN, Mannheim, 1984.

[Dro03] Patrik Drouin. Term extraction using non-technical corpora as a point of leverage. *In Terminology, 9(1)*, pages 99–115, 2003.

[Dro04] Patrik Drouin. *Detection of Domain Specific Terminology Using Corpora Comparision.* In: Proceedings of the fourth International Conference on Language Resources and Evaluation, Lissabon, 2004.

[Dun93] Ted Dunning. Accurate methode for the statistics of surprise and coincidence. *In: Computational Linguistics 19*, pages 61–74, 1993.

[Fle82] Wolfgang Fleischer. Wortbildung der deutschen gegenwartssprache. *5., unveränd. Aufl. Tübingen:Niemeyer*, pages 230–233, 1982.

[Fle92] Wolfgang Fleischer. *Wortbildung der deutschen Gegenwartssprache.* Tübingen:Niemeyer, 1992.

[GB04] Franz Guenthner and Xavier Blanco. *Multi-lexemic Expressions: An Overview.* Lingvisticae Investigationes Supplementa, Amsterdam/Philadelphia, [seneca.uab.es/filfrirom/BLANCO/PUBLIC/multilex.pdf -30.11.2006-], 2004.

[GM94] Franz Guenthner and Petra Maier. *Das CISLEX-Wörterbuchsystem.* CIS-Bericht-94-76, Oettingenstr.67, 80538 München, 1994.

[Gro93] Maurice Gross. Local grammars and their representation by finite automata. *In M. Hoey, editor, Data, Description, Discourse, HarperCollins, London*, pages 26–38, 1993.

[Gro97] Maurice Gross. The construction of local grammars. *In Finite-State Language Processing, The MIT Press, Cambridge, Massachusetts*, pages 329–354, 1997.

[Gro99] Maurice Gross. A bootstrap method for constructing local grammars. *Contemporary Mathematics: Proceedings of the Symposium, University of Belgrad*, pages 229–250, 1999.

[GT94] Gregory Grefenstette and Pasi Tapanainen. *What is a word, What is a sentence? Problems of Tokenization.* Rank Xerox Research Centre, Meylan, France, April 22 1994.

[HF97] Birgit Hamp and Helmut Feldweg. Germanet - a lexical-semantic net for german. *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid*, pages 9–15, 1997.

[HFH01] Munpyo Hong, Sisay Fissaha, and Johann Haller. Hybrid filtering for extraction of term candidates form german technical texts. *Conference TIA-2001, Nancy*, Mai 2001.

[Hof88] Lothar Hoffmann. *Vom Fachwort zum Fachtext: Beiträge zur angewandten Linguistik*. Tübingen: Gunter Narr Verlag, 1988.

[Jac01] Christian Jacquemi. *Spotting and Discovering Terms through Natural Language Processing*. The MIT Press, Cambridge, Massachusetts London, England, 2001.

[KG03] Adam Kilgarriff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational Linguisitics 29(3)*, pages 333–347, [acl.ldc.upenn.edu/J/J03/J03-3001.pdf], 2003.

[Kil01] Adam Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics 6:1*, pages 1–37, October 2001.

[Koß96] Gerhard Koß. *Namenforschung: eine Einführung in die Onomastik - 2. Aufl., S. 34-67*. Tübingen: Niemeyer, 1996.

[Kun01] Claudia Kunze. Fortentwicklungen des lexikalisch-semantischen Wortnetzes GermaNet. *Vortrag am IMS Stuttgart*, pages 1–31, November 2001.

[Lan96] Stefan Langer. *Selektonsklassen und Hyponymie im Lexikon. Semantische Klassifizierung von Nomina für das elektronische Wörterbuch CISLEX*. Dissertation, CIS-Bericht-96-94, Oettingenstr.67, 80538 München, 1996.

[Lan98] Stefan Langer. Zur Morphologie und Semantik von Nominalkomposita. *Tagungsband KONVENS 98*, pages 83–97, 1998.

[Lez99] Wolfgang Lezius. Automatische extrahierung idiomatischer bigramme aus textkorpora. *Tagungsband des 34. Linguistischen Kolloquiums*, 1999.

[Luh58] H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal*, pages 159–165, April 1958.

[Lyo68] John Lyons. *Einführung in die moderne Linguistik 7. Auflage C.H. Beck - Die englische Originalausgabe: Introduction to Theoretical Linguistics* -. Cambrige University Press, 1968.

[Mal05] Friederike Mallchok. *Automatic Recognition of Organization Names in English Business News.* Studien zur Informations- und Sprachverarbeitung Bd.9. CIS, Oettingenstr.67, 80538 München, 2005.

[MM05] Petra Maier-Meyer. *Lexikon und automatische Lemmatisierung.* CIS-Bericht-95-84, Oettingenstr.67, 80538 München, 2005.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundation of statistical natural language processing.* second printing, The MIT Press Cambridge, Massachusetts London, England, 1999.

[Mug00] H. Bergenholtz/J. Mugdan. *Wortstrukturen, in Sprachwissenschaft:ein Reader/Ludger Hoffmann(Hrsg.).-2., verbesserte Aufl..* Berlin;New York:de Gruyter, 2000.

[Mül97] W. Müller. *Sinn- und sachverwandte Wörter: Synonymwörterbuch der deutschen Sprache.* der 2 Aufl. DUDEN, Mannheim, 1997.

[Noh00] Holger Nohr. *Automatische Dokumentindexierung. - Eine Basistechnologie für das Wissensmanagement.* Fachhochschule Stuttgart, Wolframstrasse 32 D-70191 Stuttgart, 2/2000.

[Obe05] Otto Oberhausen. *Automatisches Klassifizieren.* Peter Lang GmbH, Frankfurt am Main, 2005.

[O'S06] S. O'Shaughnessy. *Dynamische Erkennung domänenspezifischen Vokabulars.* Magisterarbeit im Studiengang Computerlinguistik, Oettingenstr.67, 80538 München, 2006.

[Por80] M.F. Porter. An algorithm for suffix stripping. *Program*, pages 130–137, July 1980.

[QW02] Uwe Quasthoff and Christian Wolff. The poisson collocation measure and its applications. *International Workshop on Computational Approaches to Collocations*, 2002.

[RG00] P. Rayson and R. Garside. Comparing corpora using frequency profiling. *In Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, pages 1–6, 2000.

[RS97] Ellen Riloff and Jessica Shepherd. A corpus-based apporach for buildung semantic lexicons. *In: Proceedings of the second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, pages 117–124, 1997.

[RS99] Ellen Riloff and Jessica Shepherd. A corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction. *In: Journal of Natural Language Engineering, Bd. 5, Nr. 2*, pages 147–156, 1999.

[Sal89] G. Salton. *Automatic text processing. Chapter 9.* Addison-Wesley, 1989.

[SB87] Gerard Salton and Chris Buckley. Term weighting approaches in automatrc text retreival. *Department of Computer Science Cornell University*, Nov. 1987.

[Sch92] Thea Schippan. *Lexikologie der deutschen Gegenwartssprache.* Max Niemeyer, Tübingen, 1992.

[Sen98] Jean Senellart. Tools for locating noun phrases with finite state transducers. *The computational treatment of nominals. Proceedings of the Workshop, COLING-ACL'98*, pages 80–84, 1998.

[Sto58] W.G. Stock. Natürlichsprachige suche - more like this! lexis-nexis. *PASSWORD*, pages 21–27, Nov. 1958.

[SWY75] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *In: Communications of the ACM, Bd. 18, Nr. 11*, pages 613–620, 1975.

[Vog03] David Vogel. Using generic corpora to learn domain-specific terminology. *Workshop on Link Analysis for Detecting Complex Behavior, Washington, DC, USA*, 2003.

[WM06] Renè Witte and Jutta Mülle. *Text Mining: Wissensgewinnung aus natürlichsprachigen Dokumenten.* Universität Karlsruhe (TH), IPD, Interner Bericht 2006-5, März 2006.

[Xea02] Feiyu Xu and et al. An domain adaptive approach to automatic acquisition of domain relevant terms und their relations with bootstrapping. *Proc. Of the 3rd International Conference on Language Resources and Evaluation*, 2002.

[XK] Feiyu Xu and Daniela Kurz. Text mining for the extraction of domain relevant terms and term collocations. *[Stand: 04.08.2007, www.coli.uni-saarland.de/publikationen/softcopies/Kurz:2002:TME.pdf].*

Danksagung

Herzlich danke ich meinem Doktorvater Prof. Dr. Franz Guenthner für hilfreiche Diskussionen und wertvolle Ratschläge. Nachdem ich bei ihm den Magister (M.A.) im Hauptfach "Computerlinguistik" absolviert hatte, arbeite ich bei ihm als wissenschaftliche Hilfskraft am CIS - eine durchaus nötige finanzielle Hilfe, die meine Promotion ermöglicht. Dafür bin ich ihm zutiefst dankbar. Ich danke auch den Mitarbeitern und Studenten am CIS, die über das Thema diskutiert und viel zur Ausarbeitung der Dissertation beigetragen haben.
Auch bei der Korrekturarbeit an meiner Dissertation habe ich viel gelernt. Ohne die tatkräftige Unterstützung von Dr. Ingmar Thilo, Annette Gotscharek (M.A.), Christopher Hak (Dipl. Math.), Sintje Göritz (Dipl. Inf.) bei der Korrektur wäre sie nicht in dieser Form vollendet worden. Dafür bin ich ihnen sehr dankbar.

Ich bedanke mich bei Jesus Christus und meiner Familie für alles.

München, am 8. Oktober 2007

Kim, Dae-Woo

Lebenslauf

Daewoo Kim

10.06.1965	Geburt in Seoul, Korea
03.1972-02.1978	Chang-Chun Grundschule
03.1978-02.1981	Kwang-Sung Mittelschule
03.1981-02.1984	Soong Moon Oberschule
03.1985-02.1991	Inchon-Universität im Fach Germanistik absolviert Wehrdienst (von Mai.1986 - Oktober.1987) abgeleistet.
03.1991	Ankunft in Würzburg (Deutschland)
26.10.1992	bestandene Prüfung zum Nachweis deuscher Sprachkenntnisse
05.11.1992	Immatrikulation als Germanistik-Student an der Universität Würzburg.
12.10.1993	Einschreibung in Computerlinguistik als Hauptstudiengang an der LMU.
06.10.1995	Zwischenprüfung für das Hauptfach Computerlinguistik
06.1997-02.1998	Online-Abteilung bei "Oscar Rothacker Verlagsbuchhandlung" Aufgabe: Gestaltung von Webseiten für Internet-Buchhandel
03.1998-04.1998	Zentralabteilung Technik bei Siemens (als Werkstudent) Aufgabe: HTML-Seiten-Erstellung eines automatischen Inhaltsverzeichnisses aus einem langen Word-Dokument
20.02.2001	Abschluß in Computerlinguistik mit dem Magister (M.A.)
05.2001	Seit Mai 2001 arbeite ich bei Prof. Dr. Franz Guenthner als Wissenschaftliche Hilfskraft am CIS der LMU München.
01.04.2005	Doktorand der Computerlinguistik an der LMU.
13.02.2008	Promotion zum Dr. phil. am Centrum für Informations- und Sprachverarbeitung (CIS) der Ludwig-Maximilians-Universität München Titel: "Semantische Analyse und automatische Gewinnung von branchen-spezifischem Vokabular für E-Commerce"