# Automatische Erweiterung eines semantikbasierten Lexikons durch Bootstrapping auf großen Korpora

Chris Biemann, Rainer Osswald

Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik
Abteilung ASV
04109 Leipzig
Deutschland

Praktische Informatik VII
Informatikzentrum
FernUniversität in Hagen
Universitätsstraße 1
58084 Hagen
Deutschland

biem@informatik.uni-leipzig.de, Rainer.Osswald@fernuni-hagen.de

---

**Typ des Beitrags/Type of the paper** Vortrag/Lecture

# Automatische Erweiterung eines semantikbasierten Lexikons durch Bootstrapping auf großen Korpora

Chris Biemann, Rainer Osswald

Der Beitrag beschreibt am Beispiel von HaGenLex und des Wortschatz-Korpus, wie ein featurebasiertes semantisches Lexikon unter Zuhilfenahme großer, unstrukturierter Textkorpora automatisch erweitert werden kann. Die semantischen Klassen von Substantiven werden über die sie modifizierenden Adjektive vererbt und bisher unklassifizierten Substantiven zugewiesen. Hier erweist es sich als sinnvoll, mehrere Klassifikatoren für Einzeleigenschaften, aus denen die semantischen Klassen zusammengesetzt sind, zu kombinieren. Das Verfahren wird ausführlich evaluiert, Ausblicke für eine Erweiterung werden aufgezeigt.

The paper describes how a feature-based semantic lexicon can be automatically extended using large, unstructured text corpora. Experiments are carried out using HaGenLex and the Wortschatz corpus. The semantic classes of nouns are inherited via the adjectives that modify them. It turns out to be reasonable to combine several classifiers for single attributes into one for complex semantic classes. The method is evaluated thoroughly, and further steps are discussed.

## 1. Einführung

NLP-Anwendungen wie Text Retrieval oder Question Answering, die methodisch über statistische Assoziation und Pattern-Matching hinausgehen, erfordern es, umfangreiche natürlichsprachliche Korpora semantisch zu parsen. Dazu wird ein Computerlexikon benötigt, das neben morphosyntaktischen Angaben auch detaillierte semantische Information bereitstellt. Wenngleich die Erzeugung qualitativ hochwertiger semantischer Repräsentationen, die auch als Basis einer inferentiellen Weiterverarbeitung dienen können, vermutlich nicht ohne ein manuell erstelltes Lexikon möglich ist, steht aus Quantitätsgründen außer Frage, dass insbesondere im Substantivbereich Verfahren der automatischen semantischen Klassifikation hinzutreten müssen. Der hier vorgestellte Ansatz beschreibt, wie ein solches automatisches Klassifizieren mit Hilfe eines manuell erstellten Lexikons hinreichender Größe in Verbindung mit Kookkurrenzstatistiken über sehr großen Textbeständen gelingen kann.

## 2. Verwendete Ressourcen

### 2.1 Das semantikbasierte Computerlexikon HaGenLex

Als Basislexikon für die hier vorgestellten Klassifikationsexperimente dient das semantikbasierte Computerlexikon HaGenLex, das gegenwärtig circa 22’700 Lexikoneinträge enthält, davon etwa 11’300 Substantive und 6’700 Verben. Alle HaGenLex-Einträge sind sowohl morphosyntaktisch als auch semantisch annotiert. Grundlage der semantischen Beschreibung ist das sogenannte MultiNet-Paradigma, ein Wissensrepräsentationsformalismus, der speziell im Hinblick auf die Repräsentation natürlichsprachlicher Semantik entwickelt wurde (Helbig 2001). MultiNet stellt sowohl klassifikatorische als auch relationale Darstellungsmittel bereit, wobei zu letzteren insbesondere ein Inventar an semantischen Rollen zählt, das zur Charakterisierung von Valenzrahmen herangezogen wird. Eine ausführliche Darstellung des Aufbaus von HaGenLex-Einträgen findet sich in (Hartrumpf, Helbig & Osswald 2003).

Die hier beschriebenen Experimente beschränken sich auf die Klassifikation von Substantiven hinsichtlich ihrer *ontologischen Sorten* und *semantischen Features*. MultiNet definiert 45 hierarchisch angeordnete *ontologische Sorten*, wie *d* (*discrete object*), *abs* (*situational object*) und *at* (*attribute)*, wovon 17 für die Substantivklassifikation relevant sind (vgl. Abbildung 4). Zusätzlich sind Substantive in HaGenLex hinsichtlich 16 binärer *semantischer Features* wie HUMAN, ARTIFACT und MOVABLE klassifiziert (vgl. Abbildung 3).

Zwischen den semantischen Features untereinander und ebenso im Verhältnis der Features zu den ontologischen Sorten bestehen gewisse Abhängigkeiten. Beispielsweise sind Entitäten, die das Feature HUMAN+ tragen, stets von der Sorte *discrete object* und tragen zudem die Features ANIMATE+ und ARTIFICIAL–. Um diese Abhängigkeiten zu berücksichtigen und insbesondere um fehlerhafte Kombinationen durch den Lexikographen auszuschließen, sind die zulässigen Kombinationen von ontologischer Sorte und semantischen Features in HaGenLex explizit zu sogenannten *semantischen Klassen* zusammengefasst, über denen aufgrund der Sortenhierarchie und der Möglichkeit, den Wert einzelner semantischer Features unbestimmt zu lassen, wiederum eine Spezialisierungshierarchie definiert ist. Beispielsweise ist die semantische Klasse *prot-substance* unter anderem durch die ontologische Sorte *s* (*substance*) sowie die Feature-Ausprägungen MOVABLE+ und SPATIAL+ charakterisiert, aber hinsichtlich des Features ARTIFICIAL unbestimmt. Die entsprechenden Unterklassen sind in diesem Fall *nat-substance* und *art-substance*. Insgesamt sind 50 solche semantischen Klassen für HaGenLex-Substantive in Gebrauch, wovon die

(in der hier verwendeten Teilmenge) häufigsten 28 in Abbildung 5 aufgelistet sind.

## 2.2 Deutsches Korpus aus dem Projekt Deutscher Wortschatz

Als Textquelle verwenden wir das Deutsche Hauptkorpus des Projekts Deutscher Wortschatz[1] mit über 35 Millionen Sätzen oder 500 Millionen Tokens. Durch die Berechnung von statistisch signifikanten linken Nachbarschaftskookkurrenzen (siehe Biemann et al. 2004) und deren Filterung bezüglich Wortart werden Paare von Adjektiven und Substantiven ermittelt, die typischerweise miteinander auftreten

Kommen zwei Wörter A und B in einem Korpus nacheinander vor, so ist A linker Nachbar von B und B rechter Nachbar von A. Um statistisch signifikant häufig benachbart auftretende Wörter zu ermitteln, wird ein Signifikanzmaß (in vorliegendem Falle ein Maß der log-likelihood-Familie) angewendet, das unter Berücksichtigung der Häufigkeiten des benachbarten Auftretens sowie der Einzelhäufigkeiten der Wörter für jedes mögliche Wortpaar des Korpus einen Signifikanzwert berechnet. Dieser stellt ein Maß für das „Erstaunen“ über das gemeinsame Auftreten von A und B unter der Annahme statistischer Unabhängigkeit dar – je größer dieser Wert, desto unwahrscheinlicher stehen A und B lediglich zufällig hintereinander. Falls ein gewisser Schwellenwert überschritten ist, wird B als rechter Nachbarschaftskookkurrent von A bezeichnet, und A als linker Nachbarschaftskookkurrent von B. Die Gesamtheit der linken Nachbarschaftskookkurrenten A1, ... ,An von B heißt Nachbarschaftsprofil.

Im vorliegenden Beitrag werden lediglich Nachbarschaftsbeziehungen zwischen Adjektiven und Substantiven betrachtet. Im Folgenden wird die Gesamtheit der im linken Nachbarschaftsprofil von Substantiven vorkommenden Adjektive als Adjektivprofil (eines Substantivs) bezeichnet, analog heißen die Substantive, die im rechten Nachbarschaftsprofil eines Adjektivs vorkommen, Substantivprofil. Profile enthalten nur Wörter in unflektierter Form und bestehen aus den Vereinigungen der Profile der Vollformen des Substantivs bzw. Adjektivs. Tabelle 1 zeigt einige Beispiele für Adjektiv- und Substantivprofile.

Aus Tabelle 1 ist ersichtlich, dass es sowohl Adjektive gibt, die ausschließlich mit Substantiven desselben semantischen Typs kookkurrieren (*überbacken*, *erlegt*), sowie Adjektive, die ein weites Spektrum semantisch unterschiedlicher Substantive modifizieren (z.B. *ganz*). Für die spätere Anwendung zur Klassifikation erwies es sich als unwichtig, ob die Signifikanzwerte der Kookkurrenzen berücksichtigt werden oder nicht, weswegen diese nicht einbezogen wurden.

[1] Siehe www.wortschatz.uni-leipzig.de

| Wort | Adjektiv- bzw. Substantivprofil |
|---|---|
| Buch | neu, erschienen, erst, neuest, jüngst, gut, geschrieben, letzt, zweit, vorliegend, gleichnamig, herausgegeben, nächst, dick, veröffentlicht, ... |
| Käse | gerieben, überbacken, kleinkariert, fett, französisch, fettarm, löchrig, holländisch, handgemacht, grün, würzig, selbstgemacht, produziert, schimmelig, kontrolliert, ... |
| Camembert | gebacken, fettarm, reif |
| überbacken | Schweinesteak, Aubergine, Blumenkohl, Käse |
| erlegt | Tier, Wild, Reh, Stück, Beute, Großwild, Wildkatzen, Büffel, Rehbock, Beutetier, Wal, Hirsch, Hase, Grizzly, Wildschwein, Thier, Eber, Bär, Mücke, Feind, Drachen, Affe, Jäger, ... |
| ganz | Leben, Bündel, Stück, Volk, Wesen, Vermögen, Herz, Heer, Arsenal, Dorf, Land, Können, Berufsleben, Paket, Kapitel, Stadtviertel, Rudel, Jahrzehnt, ... |

*Tabelle 1: Ausschnitte aus korpusbasierten Adjektiv- und Substantivprofilen*

Wichtig erscheint an dieser Stelle lediglich, dass die Adjektiv-Substantiv-Paare mehrfach und typischerweise im Korpus auftreten.

Aus dem Wortschatz-Korpus konnten über 125'000 Substantive extrahiert werden, die mit einem oder mehreren von über 25'000 Adjektiven in Nachbarschaftskookkurrenzbeziehung stehen. Die Größe der Adjektivprofile folgt hierbei dem Zipf'schen Gesetz (vgl. Zipf 1929), aus dem insbesondere resultiert, dass etwa die Hälfte aller Substantive nur ein Adjektiv in ihrem Profil besitzen, ein Sechstel aller Substantive zwei Adjektive usw.

## 3. Verfahren

Unter der Annahme der *Distributional Hypothesis* (Harris 1968), aus der abgeleitet werden kann, dass semantischer Ähnlichkeit eine Funktion über die globalen Kontexte entspricht (vgl. Miller & Charles 1991), erfolgt die Klassifizierung unbekannter Substantive aufgrund der sie modifizierenden Adjektive, wobei diese Menge durch das Adjektivprofil des Substantivs approximiert wird.

### 3.1 Aufbau des Klassifikators für einzelne Eigenschaften

Für jede relevante semantische Eigenschaft von Substantiven wird ein Klassifikator folgendermaßen erstellt: für jedes Adjektiv, das mindestens ein Substantiv aus der Trainingsmenge modifiziert, wird ein Profil berechnet, das angibt, zu welchem Anteil das Adjektiv für welche Klasse spricht. Der Klassifikator ist nicht in der Anzahl der Klassen beschränkt.

```
Initialisieren der Adjektiv- und Substantivprofile;
Initialisieren der Startmenge;
Solange noch neue Substantive klassifiziert werden {
   Berechnung der Klassenwahrscheinlichkeiten der Adjektive;
   Für alle noch unklassifizierten Substantive s {
        Multipliziere die Klassenwahrsch. für jede Klasse;
        Weise die Klasse mit der höchsten Wahrsch. s zu;
   }
}
```

*Abbildung 1: Bootstrapping-Algorithmus zur Klassifikation semantischer Eigenschaften von Substantiven*

Die Klassifikation unbekannter Substantive erfolgt aufgrund der Profile der modifizierenden Adjektive; dieser Wechsel von Profilberechnung und Neuklassifikation wird im Expectation-Maximization-Bootstrapping-Stil (vgl. Dempster et al. 1977) iteriert, bis keine neuen Substantive mehr klassifiziert werden können.

Abbildung 1 beschreibt schematisch den Algorithmus, Abbildung 2 skizziert die verwendete Datenstruktur:

Abbildung 2 illustriert die Funktionsweise des Algorithmus am Beispiel eines Klassifikators mit zwei Klassen A und B. Ausgangspunkt ist eine Trainingsmenge, bestehend aus Substantiven und deren Klassen. Zunächst werden Adjektive, Substantive und deren Nachbarschaftskookkurrenzbeziehungen initialisiert,

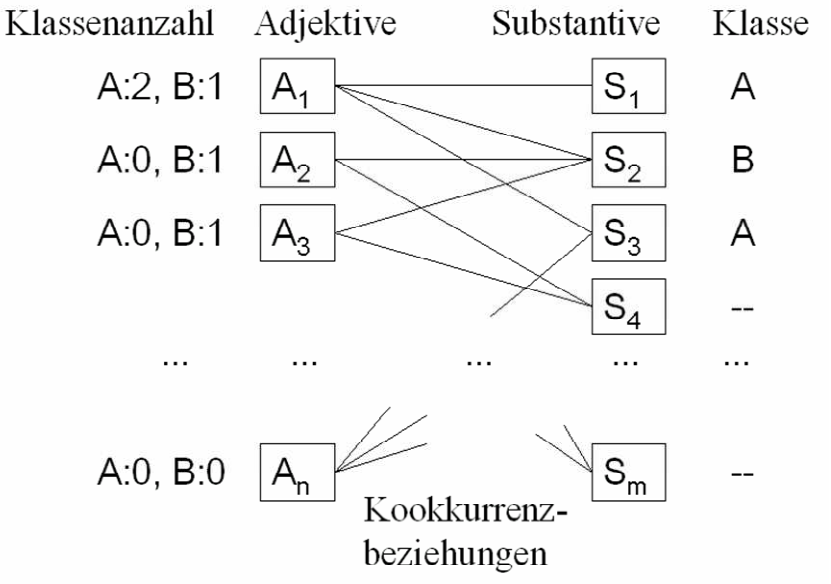


*Abbildung 2: Datenstruktur mit Beispielsituation*

den Substantiven aus der Trainingsmenge werden ihre Klassen zugeordnet.

Die folgenden Schritte werden iteriert, solange noch Substantive mit Klassen versehen werden: Jedem Adjektiv werden Klassenwahrscheinlichkeiten zugeordnet, die angeben, mit welcher Wahrscheinlichkeit (relativen Häufigkeit) das Adjektiv in den Adjektivprofilen einer Klasse zu finden ist, d.h. wie stark das Adjektiv für welche Klasse spricht. Die Wahrscheinlichkeit berechnet sich aus der Häufigkeitsverteilung (in Abb. 2 Klassenanzahl) pro Klasse, geteilt durch die Gesamtanzahl der Substantive je Klasse und in Summe normalisiert auf Eins. Das Teilen durch die Gesamtanzahl der Substantive je Klasse sorgt dafür, dass jede Klasse dieselbe Wahrscheinlichkeitsmasse besitzt, und ist für die Behandlung von sehr unterschiedlich großen Klassen unumgänglich, wie Vorversuche gezeigt haben. Da in den Iterationsschritten die Anzahl der klassifizierten Substantive ansteigt, müssen die Klassenwahrscheinlichkeiten der Adjektive zu Beginn jedes Schrittes neu berechnet werden.

Nun wird versucht, den noch unklassifizierten Substantiven Klassen zuzuweisen. Hierzu werden die Klassenwahrscheinlichkeiten der im Adjektivprofil des Substantivs enthaltenen Adjektive je Klasse multipliziert, wobei nur Adjektive betrachtet werden, deren Klassenwahrscheinlichkeiten nicht sämtlich null sind, also in mindestens einem Adjektivprofil eines schon klassifizierten Substantivs auftreten. Die Klasse mit dem höchsten Wert wird dem Substantiv zugewiesen. Zu beachten ist, dass im Falle von sich widersprechenden Klassenwahrscheinlichkeiten innerhalb eines Adjektivprofils (z.B. falls ein Adjektiv ausschließlich für Klasse A, ein anderes ausschließlich für Klasse B spricht) dem Substantiv keine Klasse zugewiesen werden kann. Ein möglicher Ausweg wäre *Smoothing*: die Klassenwahrscheinlichkeiten hätten per Definition immer einen (kleinen) Mindestwert. Versuche haben jedoch gezeigt, dass dies der Klassifikationsgenauigkeit abträglich ist, ohne dabei die Zahl der klassifizierbaren Substantive bemerkenswert zu erhöhen.

Im Beispiel von Abbildung 2 sind die Substantive $S_1$, $S_2$ und $S_3$ in der Trainingsmenge mit ihren Klassen A, B und A enthalten. Dies sorgt dafür, dass die Adjektive $A_2$ und $A_3$ mit einer Wahrscheinlichkeit von 100% für Klasse B sprechen, weswegen Substantiv $S_4$ diese Klasse zugewiesen wird.

Zur Erhöhung der Klassifikationsgenauigkeit kann gefordert werden, dass einem Substantiv nur dann eine Klasse zugewiesen wird, wenn die Anzahl der Adjektive, die für diese Klasse sprechen, einen Schwellenwert überschreitet, was bei den in Abschnitt 4 beschriebenen Versuchen auch erfolgte.

### 3.2 Kombination der Einzelklassifikatoren

Der Gesamtklassifikator soll gegebenen Substantiven eine semantische Klasse zuordnen. Ein solcher Klassifikator ließe sich im Prinzip nach dem in Abschnitt

3.1 beschriebenen Bauplan konstruieren. Erste Versuche in diese Richtung hatten allerdings zu unbefriedigenden Precision-Werten geführt (zwischen 60% Precision bei 45% Recall und 76% Precision bei lediglich 2,8% Recall). Der hier beschriebene Ansatz verwendet dagegen für jede ontologische Sorte und jedes semantische Feature einen gesonderten Klassifikator. Die Ergebnisse dieser Einzelklassifikatoren werden nach folgender Methode kombiniert:

1) Ermittle alle semantischen Klassen, die *kompatibel* mit allen Resultaten der Einzelklassifikatoren sind.
2) Wähle diejenigen unter den in 1) ermittelten Klassen aus, die *minimal* bezüglich der Spezialisierungsrelation auf der Menge der semantischen Klassen sind.
3) Enthält die in 2) ermittelte Menge genau ein Element, dann erkläre dieses zum Resultat, ansonsten verweigere eine Klassifikation.

Der Klassifikator ist somit *schwach* in dem Sinne, dass er nicht in jedem Fall eine Antwort liefert (was im Übrigen auch schon für die Einzelklassifikatoren der Fall ist). Die in Abschnitt 4.3 diskutierten Resultate beziehen sich auf das eben geschilderte Verfahren. Als Verbesserungsmöglichkeiten für künftige Experimente bieten sich folgende Modifikationen an:

- Enthält die in Schritt 2) ermittelte Menge mehr als ein Element, so wähle die speziellste semantische Klasse, die allgemeiner als alle Elemente der Menge ist.
- Wird in Schritt 1) keine kompatible Klasse gefunden, so ignoriere (sukzessive) die Resultate der am wenigsten zuverlässigen Einzelklassifikatoren bis sich eine kompatible Klasse findet; dann verfahre weiter wie in Schritt 2).

## 4. Evaluation

Zum Zwecke der Evaluation wurde eine Gesamtmenge von 6045 HaGenLex-Substantiven in jeweils 90% Trainingsmenge und 10% Testmenge aufgeteilt. Der *10-fold-cross-validation*-Methode folgend werden in je 10 Durchläufen jeweils 10% der Gesamtmenge zur Evaluation verwendet, so dass insgesamt jedes Wort einmal zur Evaluation herangezogen werden kann. Bei der Auswahl der Substantive wurde auf möglichst geringe Polysemie geachtet. Die Precision (Anzahl richtiger geteilt durch Anzahl erfolgter Klassifikationen) wurde aufgrund der Vereinigung der Testmengen berechnet, obwohl bei den Experimenten eine erheblich größere Zahl von Substantiven als die in den Testmengen vorhandenen klassifiziert wurden. Der Schwellenwert für die Mindestanzahl von

Adjektiven, die mit einem Substantiv kookkurrieren müssen, damit dieses klassifiziert wird, wurde in allen Versuchen auf 5 festgesetzt, was in jedem Versuch zur Klassifikation von über 20'000 Substantiven führte. Größere Schwellenwerte erbrachten deutliche Einbußen beim Recall (Anzahl gefundener richtiger geteilt durch Anzahl möglicher richtiger Klassifikationen) bei einer Verbesserung der Gesamt-Precision um maximal 1,8%. Da nur für 4726 der verwendeten Substantive aus HaGenLex mehr als 4 kookkurrierende Adjektive aus dem Korpus extrahiert werden konnten, ist der Gesamtrecall von vornherein auf 78,2% beschränkt.

In Abschnitt 4.1 wird die Klassifikation nach semantischen Features beschrieben, Abschnitt 4.2 widmet sich der Zuordnung der ontologischen Sorten. In Abschnitt 4.3 werden diese Einzelklassifikationen zu den komplexen semantischen Klassen kombiniert.

## 4.1 Evaluation der einzelnen semantischen Features

Da die semantischen Features mehr oder weniger orthogonal sind in der Hinsicht, dass das Vorhandensein eines bestimmten Features (Eigenschaft +) nur schwach mit dem Vorhanden- oder Nichtvorhandensein (Eigenschaft –) anderer Features korreliert, wurde für jedes der 16 Features ein binärer Klassifikatoren erstellt.

Abbildung 3 links zeigt die Verteilung in den Trainingsdaten für die einzelnen semantischen Features sowie den Anteil der kleineren Klasse (Bias).

Aus Abbildung 3 rechts ist ersichtlich, dass es dem Klassifikator möglich ist,

| **Name** | **Anzahl** | **+** | **-** | **Bias** |
|---|---|---|---|---|
| method | 6004 | 12 | 5992 | 0,0020 |
| instit | 6032 | 39 | 5993 | 0,0065 |
| mental | 9008 | 162 | 8846 | 0,0180 |
| info | 6015 | 119 | 5896 | 0,0198 |
| animal | 5995 | 143 | 5852 | 0,0239 |
| geogr | 6015 | 188 | 5827 | 0,0313 |
| thconc | 6028 | 518 | 5510 | 0,0859 |
| instru | 5932 | 969 | 4963 | 0,1634 |
| human | 5995 | 1313 | 4682 | 0,2190 |
| legper | 6009 | 1352 | 4657 | 0,2250 |
| animate | 6010 | 1505 | 4505 | 0,2504 |
| potag | 6015 | 1664 | 4351 | 0,2766 |
| artif | 5864 | 2204 | 3660 | 0,3759 |
| axial | 5892 | 2260 | 3632 | 0,3836 |
| movable | 5827 | 2345 | 3482 | 0,4024 |
| spatial | 6033 | 2910 | 3123 | 0,4823 |

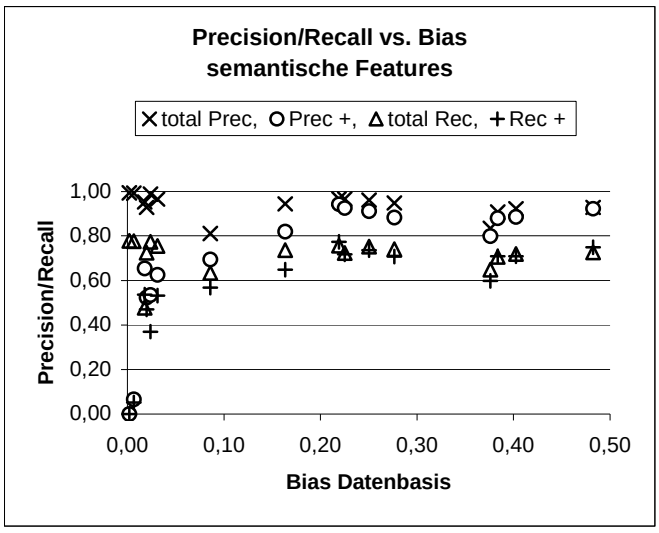


*Abbildung 3 links: Verteilung der Features in den Trainingsmengen, rechts: Precision und Recall gesamt und Precision und Recall der +-Eigenschaften in Abhängigkeit des Bias der Trainingsmenge*

mit hoher Genauigkeit den Substantiven aus der Testmenge die richtigen Features zuzuordnen, falls deren Bias nicht kleiner als 0,05 ist. In den anderen Fällen ergibt sich zwar eine hohe Precision für die jeweiligen Features (METHOD, INSTIT, MENTAL, INFO, ANIMAL und GEOGR), diese kommt aber hauptsächlich durch Zuweisen der häufigeren Klasse zustande, die seltene +-Eigenschaft wird nur unzureichend erkannt. Die Gesamtprecision beträgt 93,8% (86,8% für Eigenschaft +), der Gesamtrecall liegt bei 70,7% (69,2% für Eigenschaft +).

## 4.2 Evaluation der einzelnen ontologischen Sorten

Die ontologischen Sorten sind hierarchisch angeordnet (siehe Abschnitt 2.1) und deshalb nicht als orthogonal anzusehen. Dennoch wurden für jede der 17 Sorten binäre Trainingsmengen erstellt, die diejenigen Wörter enthalten, bei denen die jeweilige Sorte entweder zutrifft (Eigenschaft +) oder nicht (Eigenschaft –). Wörter, bei denen die Sorte unspezifiziert ist, wurden von der Trainingsmenge ausgeschlossen.
Abbildung 4 zeichnet ein ähnliches Bild wie Abbildung 3: Sorten, die in ihrem Bias über 0,1 liegen, können gut bis sehr gut differenziert werden, seltene Sorten führen zu Problemen. Bei den Sorten *ab* und *o* wurde Eigenschaft – zur Betrachtung in Abbildung 4 rechts herangezogen, da bei diesen Sorten Eigenschaft + die kleinere darstellt.

Die Gesamtprecision liegt bei 94,1% (89,5% für Eigenschaft +), der Ge-

| Name | Anzahl | + | - | Bias |
|---|---|---|---|---|
| re | 6033 | 7 | 6026 | 0,0012 |
| mo | 6033 | 8 | 6025 | 0,0013 |
| o- | 6033 | 5994 | 39 | 0,0065 |
| oa | 6045 | 41 | 6004 | 0,0068 |
| me | 6045 | 41 | 6004 | 0,0068 |
| qn | 6045 | 41 | 6004 | 0,0068 |
| ta | 6033 | 107 | 5926 | 0,0177 |
| s | 6010 | 224 | 5786 | 0,0373 |
| as | 6031 | 363 | 5668 | 0,0602 |
| na | 6033 | 411 | 5622 | 0,0681 |
| at | 6033 | 450 | 5583 | 0,0746 |
| io | 6033 | 664 | 5369 | 0,1101 |
| ad | 6031 | 1481 | 4550 | 0,2456 |
| abs | 6033 | 1846 | 4187 | 0,3060 |
| d | 6010 | 2663 | 3347 | 0,4431 |
| co | 6033 | 2910 | 3123 | 0,4823 |
| ab- | 6033 | 3082 | 2951 | 0,4891 |

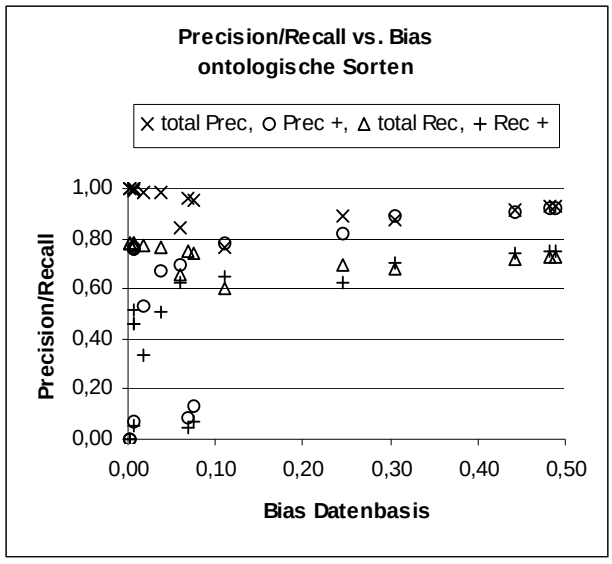


*Abbildung 4 links: Verteilung der +/– Eigenschaften in den Trainingsmengen rechts: Precision und Recall gesamt und Precision und Recall der Eigenschaften + in Abhängigkeit des Bias der Trainingsmenge*

samtrecall beträgt 73,6% (69,6% für Eigenschaft +).

### 4.3 Evaluation der zusammengesetzten semantischen Klassen

Nun wird untersucht, wie gut sich die Einzelklassifikatoren nach der in Abschnitt 3.2 beschriebenen Methode zu einem Klassifikator für semantische Klassen (im Sinne von Abschnitt 2.1) zusammensetzen lassen. Abbildung 5 präsentiert die Ergebnisse für die 28 semantischen Klassen, die mit einer Mindesthäufigkeit von 25 (≈ 0,41%) in der Trainingsmenge auftreten. Die restlichen Klassen, die insgesamt einen Anteil von 2,6% ausmachen, wurden im Kollektiv ausgewertet (Klasse „Rest“).

Bei diesem Versuch fällt auf, dass einige semantische Klassen gut erkannt werden, während andere sich offensichtlich nicht für das Verfahren eignen. Der bei den Einzelklassifikatoren beobachtbare Zusammenhang zwischen der Mächtigkeit der Klassen und der Genauigkeit des Klassifikators kann hier nicht

| **Klasse** | **Anz.** | **Prec** | **Rec** |
|---|---|---|---|
| nonment-dyn-abs-situation | 1421 | 89,19 | 34,27 |
| human-object | 1313 | 96,82 | 69,54 |
| prot-theor-concept | 516 | 53,71 | 18,22 |
| nonoper-attribute | 411 | 0,00 | 0,00 |
| ax-mov-art-discrete | 362 | 55,64 | 40,88 |
| nonment-stat-abs-situation | 226 | 36,84 | 6,19 |
| animal-object | 143 | 100,0 | 26,57 |
| nonmov-art-discrete | 133 | 57,41 | 23,31 |
| ment-stat-abs-situation | 126 | 51,28 | 15,87 |
| nonax-mov-art-discrete | 108 | 31,48 | 15,74 |
| tem-abstractum | 107 | 96,77 | 28,04 |
| mov-nonanimate-con-potag | 98 | 70,45 | 31,63 |
| art-con-geogr | 96 | 58,70 | 28,12 |
| abs-info | 94 | 42,31 | 11,70 |
| art-substance | 88 | 60,47 | 29,55 |
| nat-discrete | 88 | 100,0 | 31,82 |
| nat-substance | 86 | 57,14 | 9,30 |
| prot-discrete | 73 | 100,0 | 57,53 |
| nat-con-geogr | 63 | 65,00 | 20,63 |
| prot-substance | 50 | 100,0 | 40,00 |
| mov-art-discrete | 45 | 100,0 | 37,78 |
| meas-unit | 41 | 90,91 | 24,39 |
| oper-attribute | 39 | 0,00 | 0,00 |
| Institution | 39 | 0,00 | 0,00 |
| ment-dyn-abs-situation | 36 | 0,00 | 0,00 |
| plant-object | 34 | 100,0 | 8,82 |
| mov-nat-discrete | 27 | 22,22 | 22,22 |
| con-info | 25 | 40,00 | 8,00 |
| *Rest* | 157 | 39,24 | 19,75 |

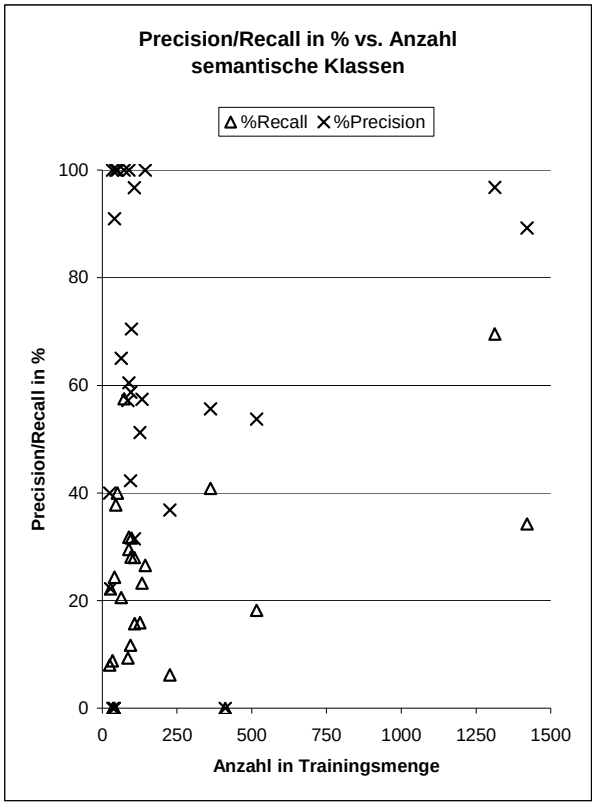


*Abbildung 5: Precision und Recall für die komplexen semantischen Klassen mit Mindesthäufigkeit 25 in der Gesamtmenge.*

beobachtet werden. Insgesamt jedoch ergibt sich eine erfreuliche Precision von 80,2%; der Gesamtrecall liegt aufgrund des häufigen Verweigerns der Klassifikation (siehe Abschnitt 3.2) bei lediglich 34,2%. Andererseits führt das vorgestellte Experiment immerhin zu einer Klassifikation von 6649 unbekannten Substantiven, wodurch der Umfang der mehr als verdoppelt wird.

## 5. Fehleranalyse

In der Tabelle von Abbildung 5 ist auffällig, dass Substantive der relativ großen Klasse *nonoper-attribute* in keinem einzigen Fall korrekt klassifiziert werden. Die Substantive dieser Klasse benennen Attribute wie Aufrichtigkeit, Pünktlichkeit und Eleganz, die *nicht* operational (durch Messung) bestimmbar sind. Das gleiche schlechte Klassifikationsverhalten zeigt übrigens auch die Klasse *oper-attribute* der operational bestimmbaren Attribute (Länge, Gewicht).[2] Die „Nichtlernbarkeit" dieser Klassen könnte zum einen daran liegen, dass Attributsubstantive typischerweise von Adjektiven modifiziert werden, welche eine Graduierung zum Ausdruck bringen (wie *gering* und *groß*), und diese möglicherweise zu unspezifisch sind.[3] Ein anderer Grund für die schlechte Klassifizierbarkeit solcher Substantive mag sein, dass ein erheblicher Anteil der nichtoperationalen Attribute in HaGenLex als semi-automatisch generierte deadjektivische Substantive auf *-heit* und *-keit* enstanden ist. Nicht wenige dieser Substantive, wie *Trockenheit*, sind aber eher als Zustandsbezeichnungen zu verstehen. Für eine genauere Begründung der fehlerhaften Leistung des Klassifikators im Fall von Attributsubstantiven ist eine detaillierte Analyse der zugehörigen Adjektivprofile erforderlich.

## 6. Ausblick

### 6.1 Semantische Klassen von Adjektiven

Eine möglicher Ansatz zur Verbesserung des Verfahrens, insbesondere im Hinblick auf das Sparse-Data-Problem, besteht darin, von einzelnen Adjektiven zu

---

[2] In der Tat weisen bereits die ontologische Sorte *at* (für *attribute*) sowie ihre Untersorten *oa* und *na* dieses Verhalten auf, was die Rekonstruktierbarkeit der betrachteten semantischen Klassen nach dem in Abschnitt 3.2 geschilderten Verfahren unmöglich macht.

[3] Zur Klassifikation von Attributsubstantiven sind nachgestellte Attribute (im grammatischen Sinn) vermutlich besser geeignet als attributive Adjektive. Dies soll durch künftige Experimente verifiziert werden.

abstrahieren und statt dessen semantische Adjektivklassen wie etwa *psychische Eigenschaft* als Merkmale zur Klassifikation zugehöriger Substantive zu verwenden. Die von HaGenLex momentan bereitgestellte sortale Klassifikation von Adjektiven ist dafür allerdings zu grob – so werden etwa graduierbare von nicht graduierbaren Qualitäten unterschieden, wobei erstere weiter in meßbare und nicht meßbare Qualitäten aufspalten (*groß* vs. *müde*).[4] Zusätzlich ist in HaGenLex aber vorgesehen, Adjektive mit Information über die Semantik der von ihnen modifizierbaren Substantive zu versehen – wie HUMAN+ für *berufstätig*.

Zur semantischen Charakterisierung von Adjektiven könnte man beispielsweise die in SIMPLE verwendete Klassifikation heranziehen (Peters & Peters 2000). Allerdings stellt sich die Frage, ob Adjektiv-Klassen wie *physical property* im Rahmen von HaGenLex nicht besser dadurch charakterisiert werden, dass diese Eigenschaften notwendigerweise von *physical objects* prädiziert werden – also über die ohnehin angelegte Möglichkeit der semantischen Restriktion modifizierbarer Substantive. Welcher Weg auch bevorzugt wird, er beinhaltet eine aufwändige semantische Annotation von Adjektiven.

Alternativ dazu könnte die Unterordnungsbeziehung von GermaNet (Kunze & Wagner, 2001) zur semantischen Abstraktion eingesetzt werden, wobei jedoch zunächst zu klären ist, ob die Qualität der GermaNet-Hierarchie im Bereich der Adjektive für diese Zwecke ausreicht.

### 6.2 Behandlung polysemer Substantive

Ein weiterer wichtiger Punkt für die Erweiterung des Ansatzes ist die Behandlung von Polysemie: Hat ein Wort mehrere Lesarten, die sich in mindestens einer Eigenschaft unterscheiden, so wird der hier vorgestellte Ansatz im günstigen Fall die häufigere Lesart zuordnen, im ungünstigen Fall das Wort nicht oder nur unterspezifiziert klassifizieren, da sich die Adjektive für manche Eigenschaften scheinbar widersprechen. Eine Möglichkeit, aus vormals einem Adjektivprofil mehrere zu erhalten, die die verschiedenen Bedeutungen widerspiegeln, wird in (Bordag 2003) für ungetypte (also nicht mit Wortarten versehene) Kookkurrenzen aufgezeigt und kann für die vorliegende Aufgabe folgendermaßen beschrieben werden: Unter der Annahme, dass innerhalb einzelner Sätze nur eine Lesart des Wortes vorkommt und die im Kontext befindlichen Wörter nur selten mit anderen Lesarten des Wortes kookkurrieren, jedoch häufig mit Kontextwörtern derselben Lesart, werden die Adjektivprofile in disjunkte Teil-

[4] Die vorgestellten Ergebnisse sollen künftig auch dahingehend genutzt werden, die diesbezüglichen Restriktionen für Adjektive aus Adjektiv-Substantiv-Kookkurrenzen zu extrahieren.

mengen zerlegt, die jeweils eine verschiedene Bedeutung des Substantives modifizieren.

Eine andere Möglichkeit bietet sich über die eigenschaftsspezifischen Adjektive an: In den nach Signifikanzstärke absteigend geordneten Substantivprofilen von eigenschaftsspezifischen Adjektiven finden sich Substantive der entsprechenden Eigenschaft an hohen Rängen. Werden die Profile mehrerer Adjektive, die für dasselbe Merkmal spezifisch sind, unter Summierung der Signifikanzstärken vereinigt, sollten sich die meisten diese Eigenschaft besitzenden Substantive in der Liste befinden. Kommt nun ein Substantiv mit hohem Rang in den Listen sich widersprechender Eigenschaften vor, sollte dieses dementsprechend mehrere Lesarten besitzen. Erste Versuche bestätigten dieses Bild, eine genauere Evaluation steht noch aus.

## Literaturverzeichnis


Biemann, Chr., Bordag, S., Heyer, G., Quasthoff, U. & Wolff, Chr. (2004): Language-independent Methods for Compiling Monolingual Lexical Data. *Proceedings of CicLING 2004*, LNCS 2945 (pp. 215-228). Berlin: Springer.

Bordag, S. (2003): Sentence Co-occurrences as Small-World-Graphs: A solution to Automatic Lexical disambiguation. *Proceedings of CicLING 2003*, LNCS 2588 (pp. 329-333). Berlin: Springer.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. 39*(B), 1-38.

Harris, Z. (1968): *Mathematical structures of language*. New York: Wiley Interscience.

Hartrumpf, S., Helbig, H. & Osswald, R. (2003): The Semantically Based Computer Lexicon HaGenLex - Structure and Technological Environment. *Traitement automatique des langues, 44*(2), 81-105.

Helbig, H. (2001): *Die semantische Struktur natürlicher Sprache – Wissensrepräsentation mit MultiNet*. Berlin: Springer.

Kunze, C. & Wagner, A. (2001). Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche. In I. Lemberg, B. Schröder & A. Storrer (Eds.), *Chancen und Perspektiven computergestützter Lexikographie* (pp. 229-246). Tübingen: Niemeyer.

Miller, G. & Charles, W. (1991): Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*(1), 1-28

Peters, I. & Peters, W. (2000). The Treatment of Adjectives in SIMPLE: Theoretical Observations. In *Proceedings of LREC 2000*.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology, 15*,1-95.