

Auf- und Ausbau des Computerlexikons HaGenLex

Rainer Osswald

Intelligente Informations- und Kommunikationssysteme
Fakultät für Mathematik und Informatik
FernUniversität in Hagen





HaGenLex

Hagen German Lexicon

[Hartrumpf/Helbig/Osswald 2003]

- ▶ Computerlexikon für den Allgemeinwortschatz des Deutschen
- ▶ Im Aufbau seit 1996 an der FernUniversität in Hagen
- ▶ 26.000 Einträge (13.500 Substantive; 7.500 Verben; 3.300 Adjektive; 600 Adverbien)
- ▶ Einträge sind morphosyntaktisch und semantisch beschrieben (inklusive syntaktischer und semantischer Valenzrahmen)

Ursprünglicher Zweck

- ▶ Lexikalisch-semantische Ressource zur Unterstützung eines semantischen Parsers



Vortragsüberblick

- ▶ Motivation: Semantisches Parsen
- ▶ Hintergrund: Der MultiNet-Formalismus
- ▶ Semantische Angaben in HaGenLex
- ▶ Eintragsaufbau, Merkmalsarchitektur, Datenformate
- ▶ Lexikonwerkbank
- ▶ Korpusbasierte Erweiterung
- ▶ Derivationssemantik im Lexikon
- ▶ Präpositionalverben



Motivation: Semantisches Parsen

WOCADI = WOrd ClAss based Disambiguating parser

[Hartrumpf 2003]

- ▶ ermittelt semantische Repräsentation natürlichsprachlicher Ausdrücke (Phrasen, Sätze, Texte)
- ▶ verwendet die lexikalische Information in HaGenLex
- ▶ basiert auf Wortklassenfunktionen [Helbig 1986]
- ▶ hat Module zur Kompositumanalyse, Disambiguierung und Koreferenzauflösung



Motivation: Semantisches Parsen

Anwendungen

- ▶ Natürlichsprachliche Schnittstellen zu Datenbanken
 - ▷ SQL-Datenbanken
 - ▷ Multimedia-Datenbanken (Wetterdaten) [Knoll/etal 1998]
 - ▷ Bibliographie-Datenbanken (Z39.50) [Leveling 2006]
- ▶ Text-Retrieval (CLEF-Task) [Leveling 2004, Leveling/Hartrumpf 2005]
- ▶ Fragebeantwortung (CLEF-Task) [Hartrumpf 2004, 2005, 2006]
- ▶ Antwortvalidierung (CLEF-Task) [Glöckner 2006]
- ▶ Lesbarkeitsbeurteilung von Texten (EU-Projekt BenToWeb)
[Hartrumpf/Helbig/Leveling/Osswald 2006]



Hintergrund: Der MultiNet-Formalismus

MultiNet = Multilayered Extended Semantic Networks [Helbig 2006]

- ▶ Formalismus zur Wissensrepräsentation / semantische Interlingua
- ▶ Grundidee semantischer Netze: Knoten repräsentieren Konzepte, Kanten repräsentieren Relationen

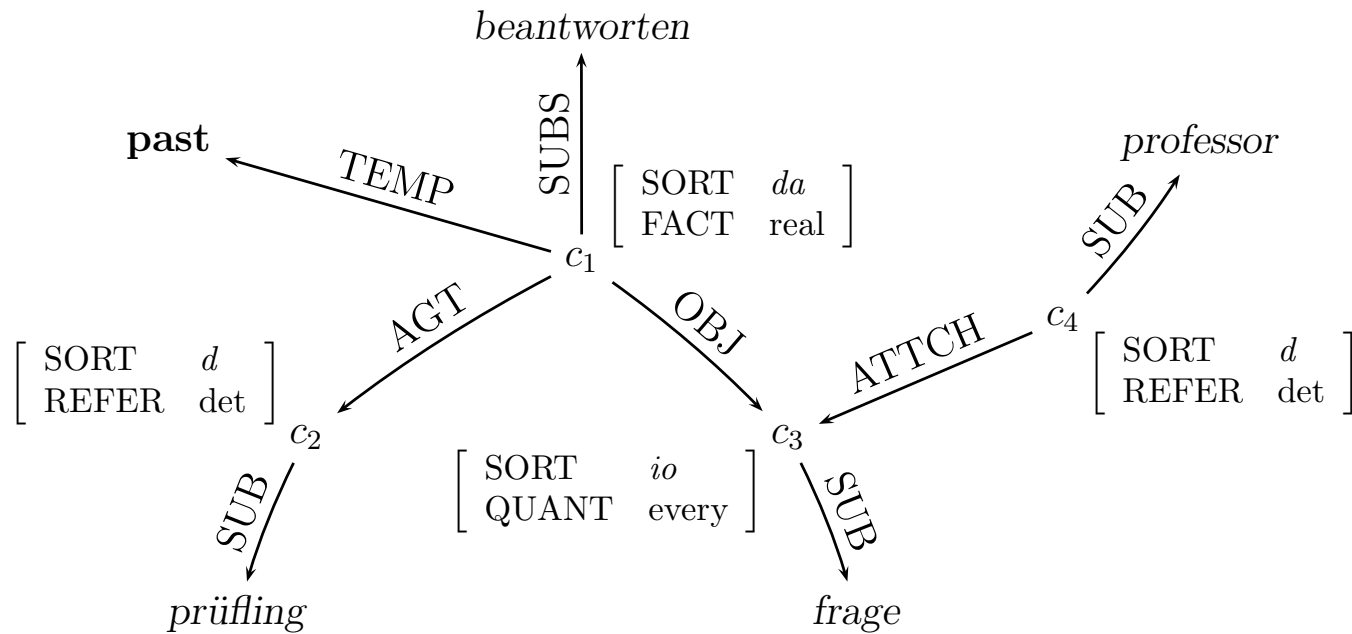
Darstellungselemente von MultiNet

- ▶ Hierarchie ontologischer Sorten: *substance, action, relation, . . .*
- ▶ über 100 semantische Relationen und Funktionen, insbes.
 - ▷ semantische Rollen: AGT (Agens), AFF (affiz. Objekt), . . .
 - ▷ lexikalische Relationen: SYNO, ANTO, . . .
- ▶ “Schichten”-Merkmale: REFER (Referenzdeterminiertheit), FACT (Faktizität), . . .



Hintergrund: Der MultiNet-Formalismus

Beispiel Der Prüfling beantwortete jede Frage des Professors.



Logische Darstellung

$$\begin{aligned} & \text{SUBS}(c_1, \text{beantworten}) \wedge \text{TEMP}(c_1, \text{past}) \wedge \text{SORT}(c_1) = da \wedge \text{FACT}(c_1) = \text{real} \\ & \wedge \text{AGT}(c_1, c_2) \wedge \text{SUB}(c_2, \text{prüfling}) \wedge \text{SORT}(c_2) = d \wedge \text{REFER}(c_2) = \text{det} \\ & \wedge \text{OBJ}(c_1, c_3) \wedge \text{SUB}(c_3, \text{frage}) \wedge \text{SORT}(c_3) = io \wedge \text{QUANT}(c_3) = \text{every} \\ & \wedge \text{ATTCH}(c_4, c_3) \wedge \text{SUB}(c_4, \text{professor}) \wedge \text{SORT}(c_4) = d \wedge \text{REFER}(c_4) = \text{det} \end{aligned}$$



Semantische Angaben in HaGenLex

Lexikoneintrag \simeq lexikalisiertes MultiNet-Konzept

- ▶ Klassifikation durch ontologische Sorten und semantische Merkmale
- ▶ Semantisch beschriebene Valenzrahmen mit semantischen Merkmalen als Selektionsrestriktionen
- ▶ Angabe kompatibler semantischer Relationen für Adjunktinterpretation (COMPAT-R)
- ▶ Zusätzliche semantische Angaben durch allgemeine MultiNet-Ausdrücke (NET)



Semantische Angaben in HaGenLex

Skizzenhafte Beschreibung zweier Verbeinträge

beantworten.1.1

<i>action, MENTAL –</i>		
AGT	ORNT	OBJ
HUMAN +	HUMAN +	INFO +
NP[nom]	NP[dat] optional	NP[acc]

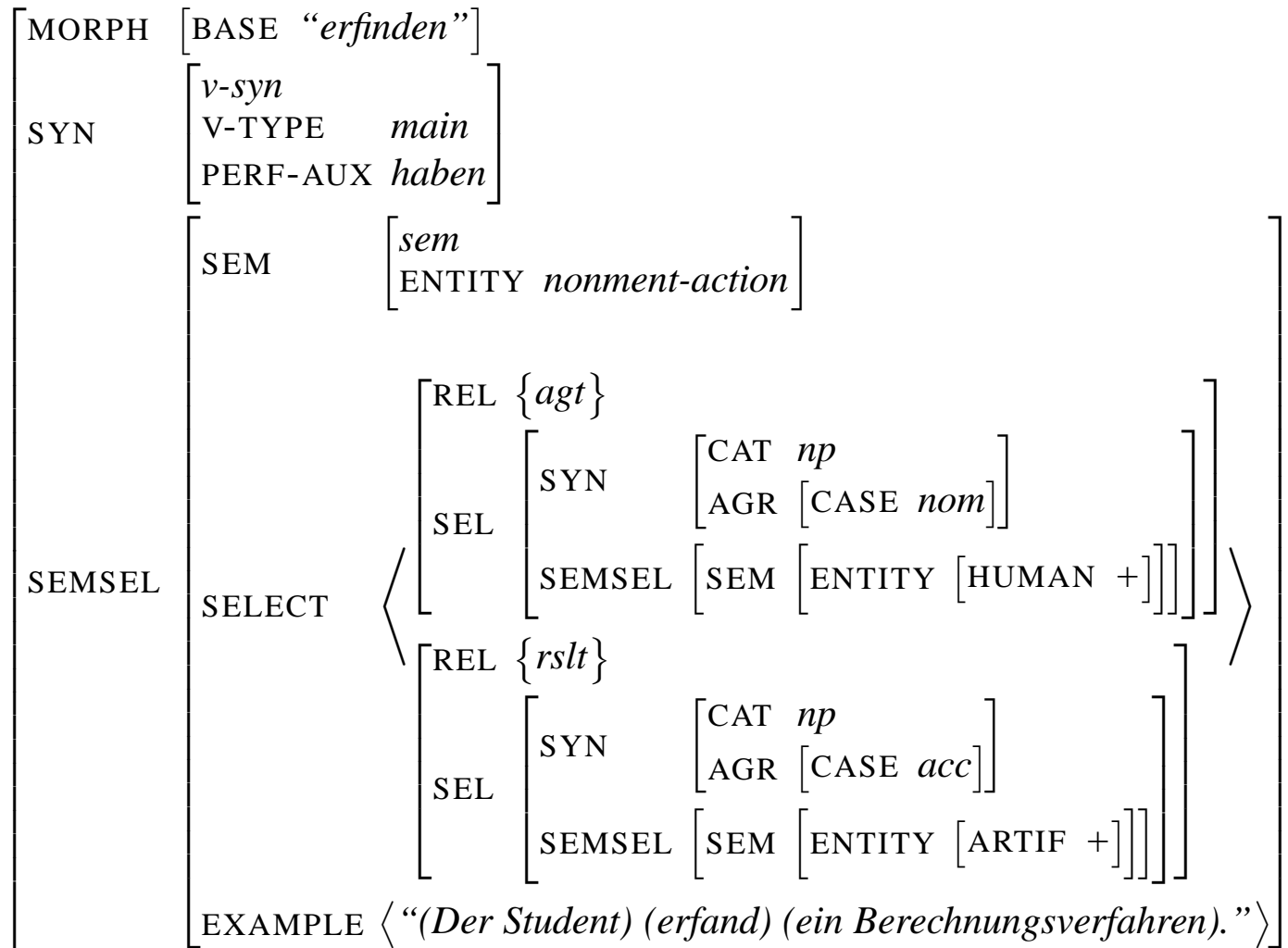
erfinden.1.1

<i>action, MENTAL –</i>	
AGT	RSLT
HUMAN +	ARTIF +
NP[nom]	NP[acc]



Aufbau der Lexikoneinträge

Beispiel Verbeintrag zu *erfinden.1.1* (vereinfacht)





Merkmalsarchitektur

- ▶ Typhierarchie ($word \leq sign$, $acc \leq case$, . . .)
- ▶ Merkmalsdeklarationen (Ausschnitt)

sign $\left[\begin{array}{ll} \text{MORPH} & \textit{morph} \\ \text{SYN} & \textit{syn} \\ \text{SEMSEL} & \textit{semsel} \end{array} \right]$

word $\left[\begin{array}{ll} \text{G-ID} & \textit{string} \\ \text{ORIGIN} & \textit{string} \end{array} \right]$

sem $\left[\begin{array}{ll} \text{SEM} & \textit{sem} \\ \text{C-ID} & \textit{string} \\ \text{DOMAIN} & \textit{domain} \\ \text{SELECT} & \textit{list(select-element)} \\ \text{COMPAT-R} & \textit{set(rel)} \end{array} \right]$

sem $\left[\begin{array}{ll} \text{ENTITY} & \textit{entity} \\ \text{NET} & \textit{net} \\ \text{LAY} & \textit{lay} \\ \text{MOLEC} & \textit{boolean} \end{array} \right]$

select-element $\left[\begin{array}{ll} \text{REL} & \textit{set(rel)} \\ \text{OBLIG} & \textit{boolean} \\ \text{SEL} & \textit{sign} \end{array} \right]$



IBL-Darstellung

IBL = Inheritance-Based Lexicon

[Hartrumpf 1996]

- ▶ **Klassen** als benannte Menge von Attribut-Wert-Spezifikationen
- ▶ Vererbungshierarchie über Klassen
- ▶ Disjunktionen und Defaults

Beispiel Nicht-lexikalische Klasse *agt-select*

```
agt-select [  
  vselect  
  rel {agt}  
  oblig +  
  sel [  
    syn (np-nom-syn mit-dat-pp-syn) ?np-nom-syn  
    semsel sem entity potag +]]
```



IBL-Darstellung

Beispiel IBL-Darstellung für Eintrag *erfinden.1.1*

```
“erfinden.1.1” [  
  verb  
  semsel [  
    v-nonment-action  
    select <  
      [ agt-select  
        sel semsel sem entity human +]  
      [ rslt-select  
        oblig —  
        sel syn np-acc-syn  
        sel semsel sem entity artif +] >  
    example ” (Der Student) (erfand) (ein Berechnungsverfahren).” ]]
```



Datenformate

- ▶ Basisdarstellung der Lexikoneinträge in Scheme (LISP-Dialekt)
Verarbeitung (Expandierung, Parserzugriff) ebenfalls in Scheme
- ▶ XML-Darstellung der Merkmalsstrukturen (u.a. nach ISO-Norm)
Motivation: Einsatz von XML-Technologien
 - ▷ XML-Editoren
 - ▷ Transformation bzw. Extraktion (XSLT, XQuery)
 - ▷ XML-Datenbanken

Lexikonwerkbank

The screenshot shows the LIApplus Lexikonwerkbank interface. The main window displays the search results for the lexeme 'informieren.1.1'. The interface includes a menu bar (Datei, Bearbeiten, Anzeige), a search input field, and a list of 18385 entries. The search results are displayed in a table format with the following fields:

Wortart	Verb
GermaNet-ID	"1 2"
Erstellungsinformation	"DS 1997-11-10"
Begriffklasse	nichtmentale Handlung
Konzept-ID	"informieren.1.1"
kompatible Relationen	{dur tlim}
Beispielsatz bzw. -phrase	"(Der Minister) (informiert) (das Parlament) (über das Gesetz)."
1. Argument	Handelnder
Syntax	NP / Nom
juristische Person ?	ja
2. Argument	Objekt d
Argument notwendig ?	nein
menschlich ?	ja
3. Argument	mentaler
Argument notwendig ?	nein
Syntax	(ueber-a

An overlay dialog box titled "Argumente notwendig?" is displayed, asking the user to mark all acceptable sentences. The dialog contains the following text and options:

Markieren Sie bitte alle akzeptablen Sätze:

- Der Minister informiert über das Gesetz
- Der Minister informiert das Parlament

The dialog box has "OK" and "Abbrechen" buttons.



Lexikonwerkbank

LIAnet

Beispielsatz bzw. -phrase

Stelligkeit ändern

1. Argument	Lexem	2. Argument	3. Argument
Der Minister	informiert	das Parlament	über das Gesetz

OK Netz erzeugen Abbrechen

Relation/
Funktion:

- *ALTN1
- *ALTN2
- *AN
- *AUF
- *BEI
- *COMP
- *FLP
- *HINTER
- *IN
- *ITMS
- *MODP
- *MODQ
- *MODS
- *NEBEN
- *OPMINUS
- *OPPLUS
- *OPTIMES
- *OPDIV
- *OPPOWER
- *ORD

Arität

lia_exmpl.net



Anbindung an andere Ressourcen

CELEX

- ▶ Verwendung der Angaben zur Flexion und Derivation

GermaNet (= “Deutsches WordNet”)

- ▶ Systematische Verlinkung von HaGenLex-Einträgen mit GermaNet-Lesarten
- ▶ Automatische Projektion von GermaNet-Relationen auf HaGenLex [Osswald 2004]
- ▶ Automatische Extraktion formaler Bedeutungspostulate aus GermaNet-Glosses [Glöckner, Hartrumpf, Osswald 2005]



Weitere Aufgaben

Kontinuierliche Erweiterung der Lexikons bzgl.

- ▶ Umfang des Wortmaterials
- ▶ Lesartenabdeckung
- ▶ Derivationsbeziehungen
- ▶ Bedeutungsbeziehungen
- ▶



Korpusbasierte Erweiterung

Zusammenarbeit mit dem WORTSCHATZ-Projekt der Uni Leipzig

[Biemann & Osswald 2006]

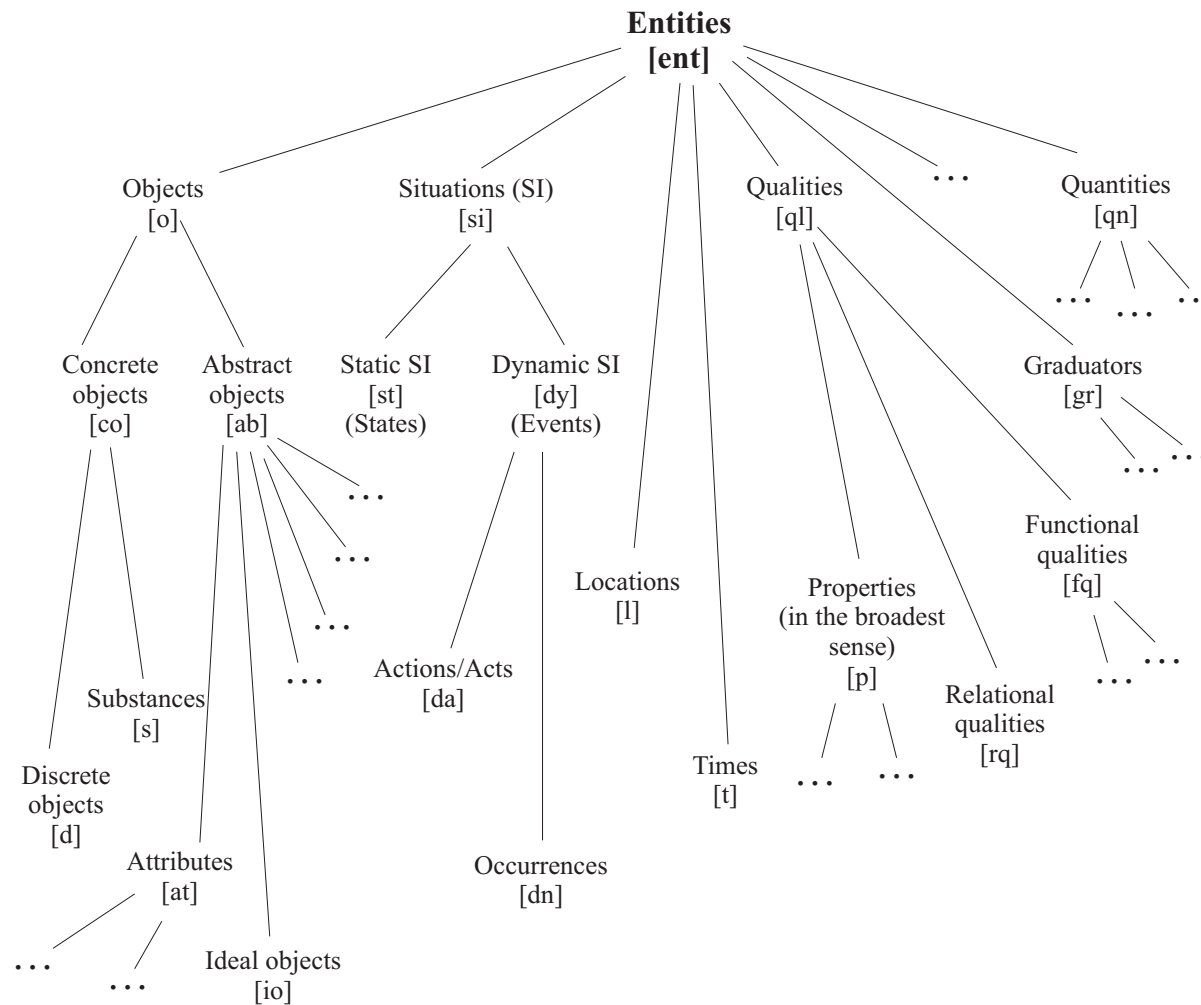
Aufgabenstellung

Automatische semantische Kategorisierung von Substantiven
unter Verwendung von

- ▶ Statistik über Adjektiv-Substantiv-Kookkurrenzen
- ▶ klassifizierte Substantive in HaGenLex



Ontologische Sorten





Semantische Merkmale

Name	Meaning	Examples	
		+	–
ANIMAL	animal	<i>fox</i>	<i>person</i>
ANIMATE	living being	<i>tree</i>	<i>stone</i>
ARTIF	artifact	<i>house</i>	<i>tree</i>
AXIAL	object having a distinguished axis	<i>pencil</i>	<i>sphere</i>
GEOGR	geographical object	<i>the Alps</i>	<i>table</i>
HUMAN	human being	<i>woman</i>	<i>ape</i>
INFO	(carrier of) information	<i>book</i>	<i>grass</i>
INSTIT	institution	<i>UNO</i>	<i>apple</i>
INSTRU	instrument	<i>hammer</i>	<i>mountain</i>
LEGP	juridical or natural person	<i>firm</i>	<i>animal</i>
MENTAL	mental object or situation	<i>pleasure</i>	<i>length</i>
METHOD	method	<i>procedure</i>	<i>book</i>
MOVABLE	object being movable	<i>car</i>	<i>forest</i>
POTAG	potential agent	<i>motor</i>	<i>poster</i>
SPATIAL	object having spatial extension	<i>table</i>	<i>idea</i>
THCONC	theoretical concept	<i>mathematics</i>	<i>pleasure</i>



Semantische Sorten

Kombination von Sorten und Merkmalen zu 86 **semantischen Sorten**

Beispiele

▶ *art-substance* (künstliche Substanz):

SORT	<i>s</i>
ANIMATE	—
ARTIF	+
INFO	—
MOVABLE	+
...	

Beispiellexeme: *Papier, Bier*

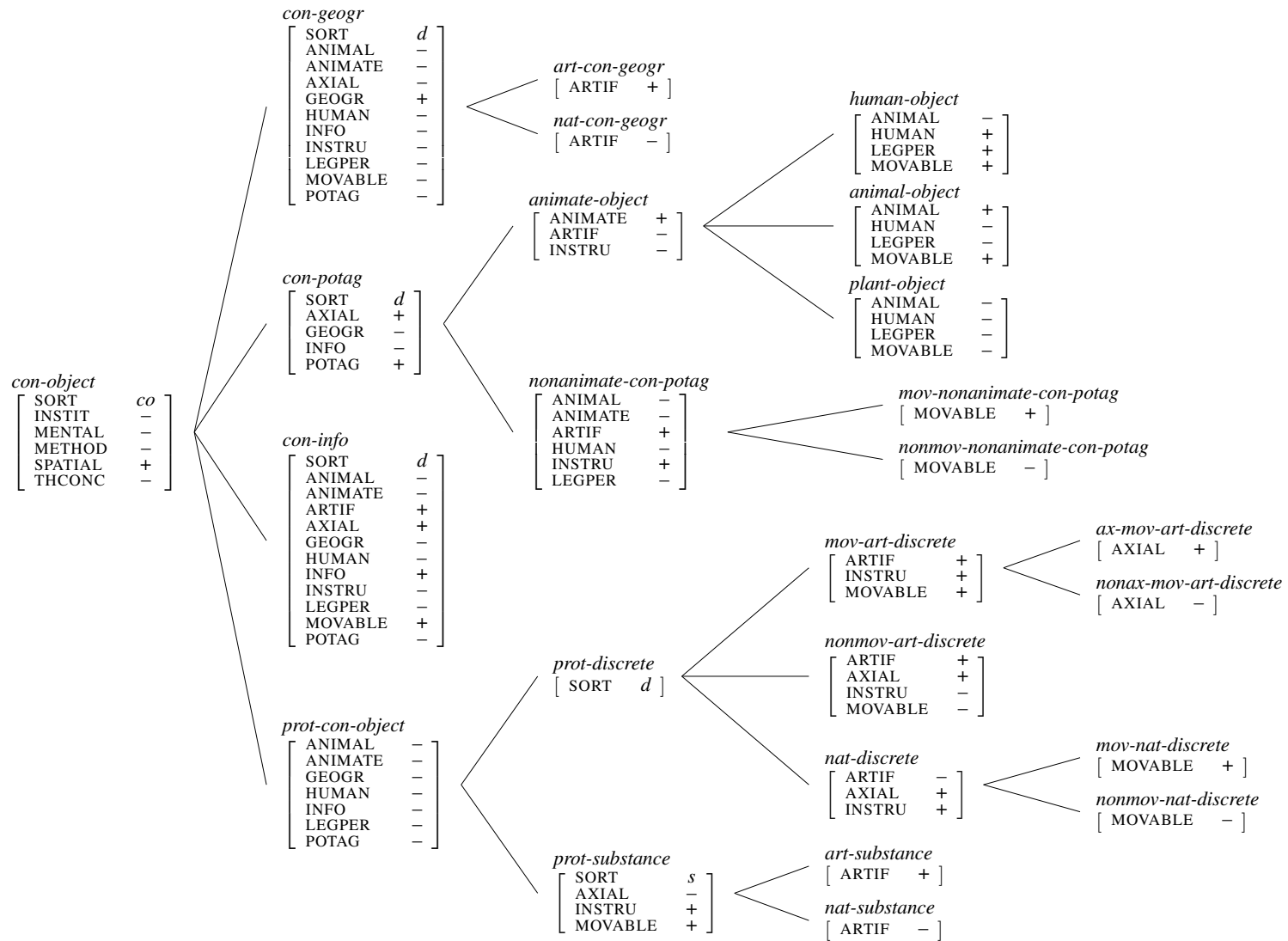
▶ *con-info* (konkretes Informationsobjekt):

SORT	<i>d</i>
ANIMATE	—
ARTIF	+
INFO	+
MOVABLE	+
...	

Beispiellexeme: *Plakat, Zertifikat*



Semantische Sorten



Experimenteller Ansatz

▶ Trainingsdaten:

6045 monoseme HaGenLex-Substantive (mit semantischer Sorte)

2 Mio Adjektiv-Substantiv-Paare (WORTSCHATZ-Korpus)

▶ Berechne Kookkurenzprofile für Substantive und Adjektive

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, . . .

erlegt Tier, Wild, Reh, Stück, Beute, Großwild, Wildkatzen, Büffel, Rehbock, Beutetier, Wal, Hirsch, Hase, Grizzly, Wildschwein, Thier, Eber, . . .

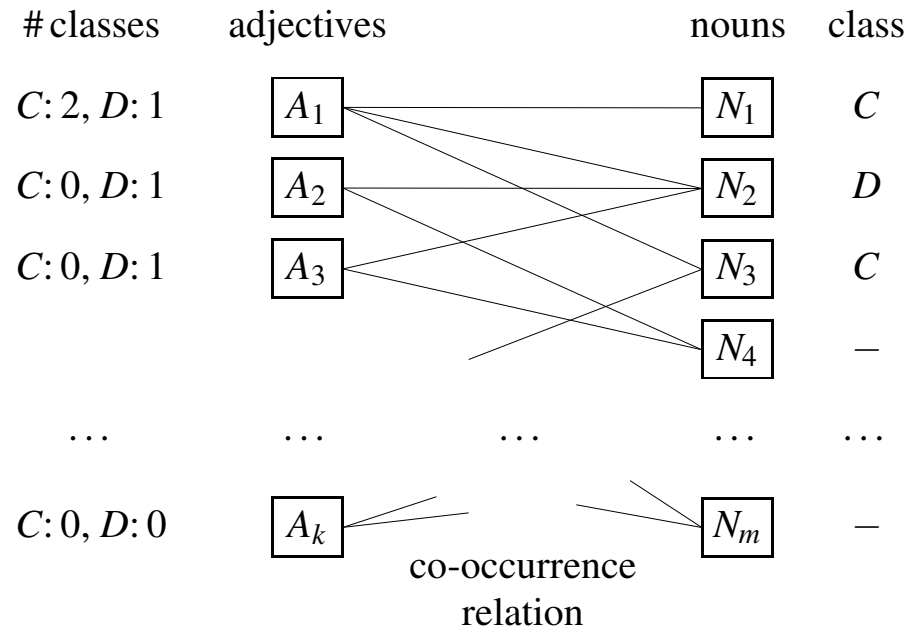
ganz Leben, Bündel, Stück, Volk, Wesen, Vermögen, Herz, Heer, Arsenal, Dorf, Land, Können, Berufsleben, Paket, Kapitel, Stadtviertel, Rudel, . . .



Experimenteller Ansatz

Algorithmus

```
Initialize adjective and noun profiles;  
Initialize the training set;  
As long as new nouns get classified:  
    Calculate adjective class probabilities;  
    For each unclassified noun n:  
        Multiply class probabilities class-wise;  
        Assign class with highest probability to noun n;
```





Ergebnisse

▶ Semantische Merkmale

Precision: 93,8% (87,6% für Wert +)

Recall: 75,8% (76,9% für Wert +)

▶ Ontologische Sorten

Precision: 93.3% (90.4% für Wert +)

Recall: 79.2% (76.3% für Wert +)

▶ Semantische Sorten

Precision: 82.3%

Recall: 32.8%



Derivationssemantik im Lexikon

Aufgabenstellung:

Halbautomatische Konstruktion lexikalischer Einträge für abgeleitete Substantive aus den Einträgen ihrer Grundwörter

[Osswald/Helbig 2005]

Spezielle Probleme:

- ▶ Ermittlung der syntaktischen und semantischen Valenzen abgeleiteter Einträge
- ▶ Vorhersage von Resultatsobjektlesarten deverbaler Substantive
Beispiele: *Erfindung, Absperrung, Verletzung*

Mögliche Anwendung: Resolution von Bridging-Referenzen

Derivationssemantik im Lexikon

Beispiel Substantiveintrag zu *Erfinder*

MORPH	[BASE "Erfinder"]
SYN	[<i>n-syn</i> CAT <i>n</i> AGR [GEND <i>masc</i>]]
DV-STATUS	[<i>derived</i> DV-ROOT "erfinden.1.1" DV-TYPE <i>er-suff</i> DV-SEM <i>nomen-agentis</i>]
SEM	[<i>sem</i> ENTITY [SORT <i>d</i> HUMAN <i>+</i>] NET ((AGT <i>erfinden.1.1 c</i>) (RSLT <i>erfinden.1.1 x1</i>))]
SEMSEL	[REL {} SELECT < [SEL [SYN [<i>np-syn</i> CAT <i>np</i> AGR [CASE <i>gen</i>]]] ∨ [PP-syn P-CASE <i>dat</i> P-FORM "von"] > OBLIG -]
EXAMPLE	"der Erfinder des Reißverschlusses"

Derivationssemantik im Lexikon

Beispiel Substantiveintrag zu *Erfindung* (Resultatslesart)

MORPH	[BASE "Erfindung"]
SYN	[<i>n-syn</i> CAT <i>n</i> AGR [GEND <i>fem</i>]]
DV-STATUS	[<i>derived</i> DV-ROOT "erfinden.1.1" DV-TYPE <i>ung-suff</i> DV-SEM <i>result-object</i>]
SEM	[<i>sem</i> ENTITY [SORT <i>o</i> ARTIF <i>+</i>] NET ((RSLT erfinden.1.1 c) (AGT erfinden.1.1 x1))]
SEMSEL	[REL {} SELECT < [SEL [SYN [<i>np-syn</i> CAT <i>np</i> AGR [CASE <i>gen</i>]]] ∨ [pp-syn P-CASE <i>dat</i> P-FORM "von"] > OBLIG -]
EXAMPLE	"die Erfindung des Forschers"



Präpositionalverben

Traditionelle Klassifikation präpositionaler Komplemente

- ▶ Präpositionalobjekte

bestehen aus, leiden an

- ▶ Prädikative Komplemente

halten für, betrachten als, ernennen zu

- ▶ Obligatorische Adverbialergänzungen (mit regulärer Semantik)

wohnen/sich aufhalten in/auf/. . .

(Beachte: *hineingelangen in*)



Präpositionalverben

Mögliche Varianten präpositionaler Objekte

- ▶ obligatorisches Komplement & eindeutige Präposition
verhelfen zu, abzielen auf
- ▶ fakultatives Komplement & eindeutige Präposition
warten (auf)

Beachte: *zählen* vs. *zählen auf*

Problem: Lesartenbestimmung

- ▶ obligatorisches Komplement & variierende Präposition
stimmen für / gegen

Konsequenz: Präposition trägt Bedeutung

Beachte: *bestehen auf* vs. *bestehen aus*

Präpositionalverben

Beispiel

[Osswald/Helbig/Hartrumpf 2006]

ernennen zu

<i>c</i> : action, MENTAL —		
<i>x</i> ₁ : AGT NP[nom]	<i>x</i> ₂ : OBJ NP[acc]	<i>x</i> ₃ : — PP[zu[dat]]

- ▶ Semantischer Valenzrahmen in logischer Notation

$\text{AGT}(c, x_1) \ \& \ \text{OBJ}(c, x_2)$

- ▶ Zusätzliche MultiNet-Spezifikation

$\text{RSLT}(c, y) \ \& \ \text{SUBR}(y, \text{RPRS}') \ \& \ \text{ARG1}(y, x_2) \ \& \ \text{ARG2}(y, x_3)$

- ▶ Paraphrase: *x*₁ tut etwas mit dem Resultat, dass *x*₂ ein *x*₃ repräsentiert.

Präpositionalverben

Semantische Klassifikation von Präpositionalverben (Skizze)

- ▶ Abstraktion über semantische Spezifikation (MultiNet)
- ▶ Zusätzliche semantische Gruppierung

Vorläufige Ergebnisse

- ▶ 150 Hauptklassen, 750 Unterklassen (insges. über 3000 Lesarten)

Beispielklasse

- ▶ “Bewertung von Attributen” (*etwas an jmdm/etwas schätzen, ...*)

schätzen, mögen, genießen, lieben, verehren, bewundern

hassen, fürchten, verabscheuen, verachten

kritisieren, bemängeln, beanstanden, reklamieren, rügen, tadeln,

bemäkeln, bekritteln, missbilligen, verurteilen, anprangern, monieren

loben, rühmen, preisen, würdigen