Experiments on the Automatic Induction of German Verb Classes

Sabine Schulte im Walde
Institute for Natural Language Processing (IMS)
University of Stuttgart
Germany

Computer Laboratory
University of Cambridge
March 2003

Goals

1. Automatic acquisition of high-quality and large-scale lexical resource for NLP applications: German semantic verb classes

2. Theoretical investigation of relationship between verb behaviour and meaning components

3. Development of clustering methodology suitable for the demands of natural language

Overview

1. Theoretical and practical aspects of verb classes
2. German verbs and verb classes
3. Clustering methodology
4. Statistical grammar model
5. Corpus-based empirical lexical acquisition $\rightarrow$ verb description
6. Clustering experiments and results

Verb Classes

verb meaning components $\leftrightarrow$ *verb behaviour*

To a certain extent, the lexical meaning components of a verb determine its behaviour, particularly with respect to the choice of its arguments.

Verb Class Usage

- Redundancy reduction in verb descriptions by generalising over the common properties of verbs
- Prediction of properties and refinement of vague properties

$\rightarrow$ Possible NLP applications: parsing, language modelling, machine translation, information extraction

Verb Classes for NLP Applications

- Machine translation (Dorr 1997)
- Document classification (Klavans and Kan 1998)
- Word sense disambiguation
  (Dorr and Jones 1996, Prescher et al. 2000)
- Subcategorisation acquisition and filtering (Korhonen 2002)

German Verb Classes: Constitution

- 168 German verbs $\rightarrow$ 43 semantic classes
- Class size: 2-7 verbs ($\varnothing$ 3.9 verbs per class)
- Ambiguity: 8 verbs with 2 senses
- High and low frequency verbs: 8 $\leq$ freq $\leq$ 71,604
- Basis: semantic intuition
- Relation to Levin (1993), consistency with Schumacher (1986)

German Semantic Verb Classes (1)

1. *Aspect* : anfangen, aufhören, beenden, beginnen, enden
2. *Propositional Attitude* : ahnen, denken, glauben, vermuten, wissen
3. *Desire*
   (a) *Wish* : erhoffen, wollen, wünschen
   (b) *Need* : bedürfen, benötigen, brauchen
4. *Transfer of Possession (Obtaining)* : bekommen, erhalten, erlangen, kriegen
5. *Transfer of Possession (Giving)*
   (a) *Gift* : geben, leihen, schenken, spenden, stiften, vermachen, überschreiben
   (b) *Supply* : bringen, liefern, schicken, vermitteln, zustellen
6. *Manner of Motion*
   (a) *Manner of Locomotion* : gehen, klettern, kriechen, laufen, rennen, schleichen, wandern
   (b) *Rotation* : drehen, rotieren
   (c) *Rush* : eilen, hasten
   (d) *Means* : fahren, fliegen, rudern, segeln
   (e) *Flotation* : fließen, gleiten, treiben

German Semantic Verb Classes (2)

7. *Emotion*
   (a) *Origin* : ärgern, freuen
   (b) *Expression* : heulen, lachen, weinen
   (c) *Objection* : ängstigen, ekeln, fürchten, scheuen
8. *Manner of Look on the Face* : gähnen, grinsen, lachen, lächeln, starren
9. *Perception* : empfinden, erfahren, fühlen, hören, riechen, sehen, wahrnehmen
10. *Manner of Articulation* : flüstern, rufen, schreien
11. *Moaning* : heulen, jammern, klagen, lamentieren
12. *Communication* : kommunizieren, korrespondieren, reden, sprechen, verhandeln
13. *Statement*
    (a) *Announcement* : ankündigen, bekanntgeben, eröffnen, verkünden
    (b) *Constitution* : anordnen, bestimmen, festlegen
    (c) *Promise* : versichern, versprechen, zusagen
14. *Observation* : bemerken, erkennen, erfahren, feststellen, realisieren, registrieren
15. *Description* : beschreiben, charakterisieren, darstellen, interpretieren
16. *Presentation* : darstellen, demonstrieren, präsentieren, veranschaulichen, vorführen
17. *Speculation* : grübeln, nachdenken, phantasieren, spekulieren

German Semantic Verb Classes (3)

18. *Insistence* : beharren, bestehen, insistieren, pochen
19. *Teaching* : beibringen, lehren, unterrichten, vermitteln
20. *Position*
    (a) *Bring into Position* : legen, setzen, stellen
    (b) *Be in Position* : liegen, sitzen, stehen
21. *Production* : bilden, erzeugen, herstellen, hervorbringen, produzieren
22. *Renovation* : dekorieren, erneuern, renovieren, reparieren
23. *Support* : dienen, folgen, helfen, unterstützen
24. *Quantum Change* : erhöhen, erniedrigen, senken, steigern, vergrößern, verkleinern
25. *Opening* : öffnen, schließen
26. *Existence* : bestehen, existieren, leben
27. *Consumption* : essen, konsumieren, lesen, saufen, trinken
28. *Elimination* : eliminieren, entfernen, exekutieren, töten, vernichten
29. *Basis* : basieren, beruhen, gründen, stützen
30. *Inference* : folgern, schließen
31. *Result* : ergeben, erwachsen, folgen, resultieren
32. *Weather* : blitzen, donnern, dämmern, nieseln, regnen, schneien

Clustering Methodology

1. Statistical acquisition of lexical verb information
2. Automatic verb clustering by standard technique k-Means
3. Clustering evaluation against manual verb classification

Lexical Acquisition Framework

- Lexicalised probabilistic context-free grammar (Carroll/Rooth 1998)
- Unsupervised training by the *EM-Algorithm* (Baum 1972)
- Robust statistical parser **LoPar** (Schmid 2000)
- 35 million words of German newspaper corpora
- Lexicalised grammar rules and lexical choice parameters
- Corpus-based empirical lexical acquisition

Corpus-Based Statistical Lexical Acquisition

D1 coarse syntactic definition of subcategorisation

D2 syntactico-semantic definition of subcategorisation with prepositional preferences

D3 syntactico-semantic definition of subcategorisation with prepositional and selectional preferences

Subcategorisation Frame Elements

n	noun phrase: NP_{Nom}
a	noun phrase: NP_{Akk}
d	noun phrase: NP_{Dat}
r	reflexive pronoun
p	prepositional phrase
x	expletive *es*
i	subordinated non-finite clause
s-2	subordinated finite verb second clause
s-dass	subordinated finite *dass*-clause
s-ob	subordinated finite *ob*-clause
s-w	indirect *wh*-questions
k	copula constructions

Examples: **nad, np:auf**$_{Akk}$**, nai**

Lexical Verb Information (1):

Subcategorisation Frame Definition

Frame	Freq	Prob
ns-dass	1,929	0.27945
ns-2	1,888	0.27358
np	687	0.09951
n	608	0.08811
na	555	0.08046
ni	346	0.05015
nd	234	0.03392
nad	160	0.02325
nds-2	70	0.01011

probability distribution for *glauben* 'to think/believe'

(probability values $>$1%)

Lexical Verb Information (2):

Subcategorisation Frame Definition + PPs

Refined Frame		Freq	Prob
np:über$_{Akk}$	acc / 'about'	480	0.11981
np:von$_{Dat}$	dat / 'about'	463	0.11568
np:mit$_{Dat}$	dat / 'with'	280	0.06983
np:in$_{Dat}$	dat / 'in'	81	0.02031

refined **np** probability distribution for *reden* 'to talk'

with total joint probability p(*reden*,**np**) = 0.35820

(probability values $>$1%)

Selectional Preferences: Nominal Level

Noun		Freq
Ziel	‘goal’	86.30
Strategie	‘strategy’	27.27
Politik	‘policy’	25.30
Interesse	‘interest’	21.50
Konzept	‘concept’	16.84
Entwicklung	‘development’	15.70
Kurs	‘direction’	13.96
Spiel	‘game’	12.26
Plan	‘plan’	10.99
Spur	‘trace’	10.91
Programm	‘program’	8.96
Weg	‘way’	8.70

Nominal arguments for *verfolgen* ‘to follow’ in **na**

Selectional Preferences: Nominal Level

Noun		Freq
Uhr	‘o’clock’	85.38
Prozeß	‘process’	77.14
Kampf	‘fight’	69.77
Verhandlung	‘negotiation’	65.51
Krieg	‘war’	64.39
Tag	‘day’	52.12
Zeit	‘time’	51.94
Arbeit	‘work’	46.86
Geschichte	‘history’	46.10
Karriere	‘career’	42.49
Spiel	‘game’	37.79
Diskussion	‘discussion’	31.25

Nominal arguments for *beginnen* ‘to begin’ in **n**

GermaNet Top Level Nodes

Lebewesen	‘creature’
Sache	‘thing’
Besitz	‘property’
Substanz	‘substance’
Nahrung	‘food’
Mittel	‘means’
Situation	‘situation’
Zustand	‘state’
Struktur	‘structure’
Physis	‘body’
Zeit	‘time’
Ort	‘space’
Attribut	‘attribute’
Kognitives Objekt	‘cognitive object’
Kognitiver Prozess	‘cognitive process’

Selectional Preferences: GermaNet Top Level Nodes

Synset	Freq	Prob
Situation	140.99	0.244
Kognitives Objekt	109.89	0.191
Zustand	81.35	0.141
Sache	61.30	0.106
Attribut	52.69	0.091
Lebewesen	46.56	0.081
Ort	45.96	0.080
Struktur	14.25	0.025
Kognitiver Prozess	11.77	0.020
Zeit	4.58	0.008
Besitz	2.86	0.005
Substanz	2.08	0.004
Nahrung	2.00	0.003
Physis	0.50	0.001

GermaNet arguments for *verfolgen* ‘to follow’ in **na**

Selectional Preferences: GermaNet Top Level Nodes

Synset	Freq	Prob
Situation	1,102.26	0.425
Zustand	301.82	0.116
Zeit	256.64	0.099
Sache	222.13	0.086
Kognitives Objekt	148.12	0.057
Kognitiver Prozess	139.55	0.054
Ort	107.68	0.041
Attribut	101.47	0.039
Struktur	87.08	0.034
Lebewesen	81.34	0.031
Besitz	36.77	0.014
Physis	4.18	0.002
Substanz	3.70	0.001
Nahrung	3.29	0.001

GermaNet arguments for *beginnen* ‘to begin’ in **n**

Verb Description

D1 coarse syntactic definition of subcategorisation

D2 syntactico-semantic definition of subcategorisation with prepositional preferences

D3 syntactico-semantic definition of subcategorisation with prepositional and selectional preferences

D1 → D2 → D3

$D1$		$D2$		$D3$	
np	0.43	n	0.28	n(Situation)	0.12
n	0.28	np:um$_{Akk}$	0.16	np:um$_{Akk}$(Situation)	0.09
ni	0.09	ni	0.09	np:mit$_{Dat}$(Situation)	0.04
na	0.07	np:mit$_{Dat}$	0.08	ni(Lebewesen)	0.03
nd	0.04	na	0.07	n(Zustand)	0.03
nap	0.03	np:an$_{Dat}$	0.06	np:an$_{Dat}$(Situation)	0.03
nad	0.03	np:in$_{Dat}$	0.06	np:in$_{Dat}$(Situation)	0.03
nir	0.01	nd	0.04	n(Zeit)	0.03
ns-2	0.01	nad	0.02	n(Sache)	0.02
xp	0.01	np:nach$_{Dat}$	0.01	na(Situation)	0.02

Example: *beginnen* ‘to begin’

D1 → D2 → D3

$D1$		$D2$		$D3$	
na	0.42	na	0.42	na(Lebewesen)	0.33
n	0.26	n	0.26	na(Nahrung)	0.17
nad	0.10	nad	0.10	na(Sache)	0.09
np	0.06	nd	0.05	n(Lebewesen)	0.08
nd	0.05	ns-2	0.02	na(Lebewesen)	0.07
nap	0.04	np:auf$_{Dat}$	0.02	n(Nahrung)	0.06
ns-2	0.02	ns-w	0.01	n(Sache)	0.04
ns-w	0.01	ni	0.01	nd(Lebewesen)	0.04
ni	0.01	np:mit$_{Dat}$	0.01	nd(Nahrung)	0.02
nas-2	0.01	np:in$_{Dat}$	0.01	na(Attribut)	0.02

Example: *essen* ‘to eat’

D1 → D2 → D3

$D1$		$D2$		$D3$	
n	0.34	n	0.34	n(Sache)	0.12
np	0.29	na	0.19	n(Lebewesen)	0.10
na	0.19	np:in$_{Akk}$	0.05	na(Lebewesen)	0.08
nap	0.06	nad	0.04	na(Sache)	0.06
nad	0.04	np:zu$_{Dat}$	0.04	n(Ort)	0.06
nd	0.04	nd	0.04	na(Sache)	0.05
ni	0.01	np:nach$_{Dat}$	0.04	np:in$_{Akk}$(Sache)	0.02
ns-2	0.01	np:mit$_{Dat}$	0.03	np:zu$_{Dat}$(Sache)	0.02
ndp	0.01	np:in$_{Dat}$	0.03	np:in$_{Akk}$(Lebewesen)	0.02
ns-w	0.01	np:auf$_{Dat}$	0.02	np:nach$_{Dat}$(Sache)	0.02

Example: *fahren* ‘to drive’

k-Means Clustering

- k-Means algorithm (Forgy 1965)
- Unsupervised hard clustering: n objects $\rightarrow$ k clusters
- Iterative re-organisation of cluster membership:
  1. Initial cluster assignment: agglomerative hierarchical clusters
  2. Calculation of cluster centroids
  3. Determining closest cluster (centroid)
  4. Re-arrangement of cluster membership
- Similarity measure: skew divergence

  $d(v_1, v_2) = D(p \,||\, w * q + (1-w) * p)$ with weight w set to 0.9
- Number of clusters: 43 (= manual)

Clustering Evaluation

Gold standard : hand-constructed verb classes

- Evaluation = Similarity between two partitions on verb set
- Evaluation measure: adjusted Rand index (Hubert/Arabie 1985):
  - Agreement between verb pairs in the partitions
  - Correction for chance in comparison to random partition
  - Typical range: $0 \leq R_{adj} \leq 1$

$$R_{adj} = \frac{\sum_{i,j}\binom{t_{ij}}{2} - \frac{\sum_i \binom{t_{i.}}{2} \sum_j \binom{t_{.j}}{2}}{\binom{n}{2}}}{\frac{1}{2}\left(\sum_i \binom{t_{i.}}{2} + \sum_j \binom{t_{.j}}{2}\right) - \frac{\sum_i \binom{t_{i.}}{2} \sum_j \binom{t_{.j}}{2}}{\binom{n}{2}}} \quad (1)$$

Feature Variation

D1 –

D2 amount of PP information:

- arguments according to standard German grammar
- chosen PPs: 30 most frequent PPs with at least 10 verbs
- all kinds of PPs in relevant frame types

D3 a) role choice: nominal / 15 / 2-3

b) role integration: separately / combined

c) role means other than GermaNet

Clustering Results on *D1* and *D2*

Distribution		R_{adj}
$D1$		0.094
$D2$	pp_{arg}	0.151
	pp_{chosen}	0.151
	pp_{all}	0.160

Clustering Results on Varying *D3*

Single Slots		Slot Combinations	
n	0.125	na	0.137
na	0.176	n/na	0.128
na	0.164	nad	0.088
nad	0.144	n/na/nad	0.118
nad	0.115	nd	0.150
nad	0.161	n/na/nd	0.124
nd	0.152	n/na/nad/nd	0.161
nd	0.143	n/na/nd/nad/ns-dass	0.182
np	0.133	np/ni/nr/ns-2/ns-dass	0.131
ni	0.148	all NP	0.158
nr	0.136	all NPs+PPs	0.176
ns-2	0.121		
ns-dass	0.156		

Cluster Analysis (D3) – Examples

(a) beginnen$_{1}$ bestehen$_{37}$ enden$_{1}$ existieren$_{37}$ laufen$_{8}$ liegen$_{31}$ sitzen$_{31}$ stehen$_{31}$

(b) eilen$_{10}$ gleiten$_{12}$ kriechen$_{8}$ rennen$_{8}$ starren$_{16}$

(c) fahren$_{11}$ fliegen$_{11}$ fließen$_{12}$ klettern$_{8}$ segeln$_{11}$ wandern$_{8}$

(d) bilden$_{32}$ erhöhen$_{35}$ festlegen$_{22}$ senken$_{35}$ steigern$_{35}$ vergrößern$_{35}$ verkleinern$_{35}$

(e) töten$_{39}$ unterrichten$_{29}$

(f) nieseln$_{43}$ regnen$_{43}$ schneien$_{43}$

(g) dämmern$_{43}$

Large-Scale Cluster Analysis (D3) – Examples ↓ →

(a) *anhören* 'to listen', *auswirken* 'to affect', *einigen* 'to agree', *lohnen* 'to be worth', *verhalten* 'to behave', *wandeln* 'to promenade'

(b) *beschleunigen* 'to speed up', ***bilden***, *darstellen* 'to illustrate', *decken* 'to cover', *erfüllen* 'to fulfil', ***erhöhen*** 'to raise', *erledigen* 'to fulfil', *finanzieren* 'to finance', *füllen* 'to fill', *lösen* 'to solve', *rechtfertigen* 'to justify', ***reduzieren*** 'to reduce', ***senken*** 'to lower', ***steigern*** 'to increase', ***verbessern*** 'to improve', ***vergrößern*** 'to enlarge', ***verkleinern*** 'to make smaller', ***verringern*** 'to decrease', ***verschieben*** 'to shift', ***verschärfen*** 'to intensify', ***verstärken*** 'to intensify', ***verändern*** 'to change'

Large-Scale Cluster Analysis (D3) – Examples ↑

(a) ***ahnen*** ‘to guess’, ***bedauern*** ‘to regret’, ***befürchten*** ‘to fear’, ***bezweifeln*** ‘to doubt’, ***merken*** ‘to notice’, ***vermuten*** ‘to assume’, *weißen* ‘to whiten’, ***wissen*** ‘to know’

(b) ***basieren*** ‘to be based on’, ***beruhen*** ‘to be based on’, ***resultieren*** ‘to result from’, ***stammen*** ‘to stem from’

(c) ***befragen*** ‘to interrogate’, ***entlassen*** ‘to release’, ***ermorden*** ‘to assassinate’, ***erschießen*** ‘to shoot’, ***festnehmen*** ‘to arrest’, ***töten*** ‘to kill’, ***verhaften*** ‘to arrest’

(d) ***beziffern*** ‘to amount to’, ***schätzen*** ‘to estimate’, ***veranschlagen*** ‘to estimate’

Discussion

- Clustering methodology for automatic basis of large-scale verb classification
- Step-wise refinement of features improves clustering
- Linguistic intuition and clustering results do not necessarily align
- Difficulty of selecting and encoding verb features
  ($\rightarrow$ selectional preferences)
- Idiosyncratic properties of verbs
- Specific properties of desired verb classification

Future Work

- Extension of number of verbs and verb classes
- Soft clustering approach
- Classification approach
- Classification task on learned classification
- Application task, e.g. class information for parsing improvement
- Human judgement task on clustering verbs