

The Importance of Feature Selection in a Linguistic Clustering of German Verbs

Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

Graduiertenkolleg
Sprachliche Repräsentationen und ihre Interpretation

Klausurtagung 20.-23.06.02

Overview

1. PhD Goals
2. Semantic verb classes
3. Clustering methodology
4. Small-scale semantic clustering of German verbs
5. Extension of verb classification
6. Problems and demands on feature description
7. Experiments on selectional preference definition
8. Discussion

PhD Goals

- Automatic acquisition of high-quality and large-scale lexical resource for NLP applications:
German semantic verb classes
- Theoretical investigation of relationship between verb behaviour and meaning components
- Development of clustering methodology suitable for the demands of natural language

Verb Classes: Hypothesis

meaning components ⇔ *syntactic behaviour*

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

Existing Verb Classes

- English (Levin 1993)
- Bangla, Korean (Jones et al. 1994)
- French (Saint-Dizier 1996)
- Spanish (Vázquez et al. 2000)

Class Usage

- Machine translation (Dorr 1997)
- Document classification (Klavans/Kan 1998)
- Word sense disambiguation (Dorr/Jones 1996)
- Smoothing → machine translation (Prescher et al. 2000)
- Probabilistic grammars (Riezler et al. 2000)

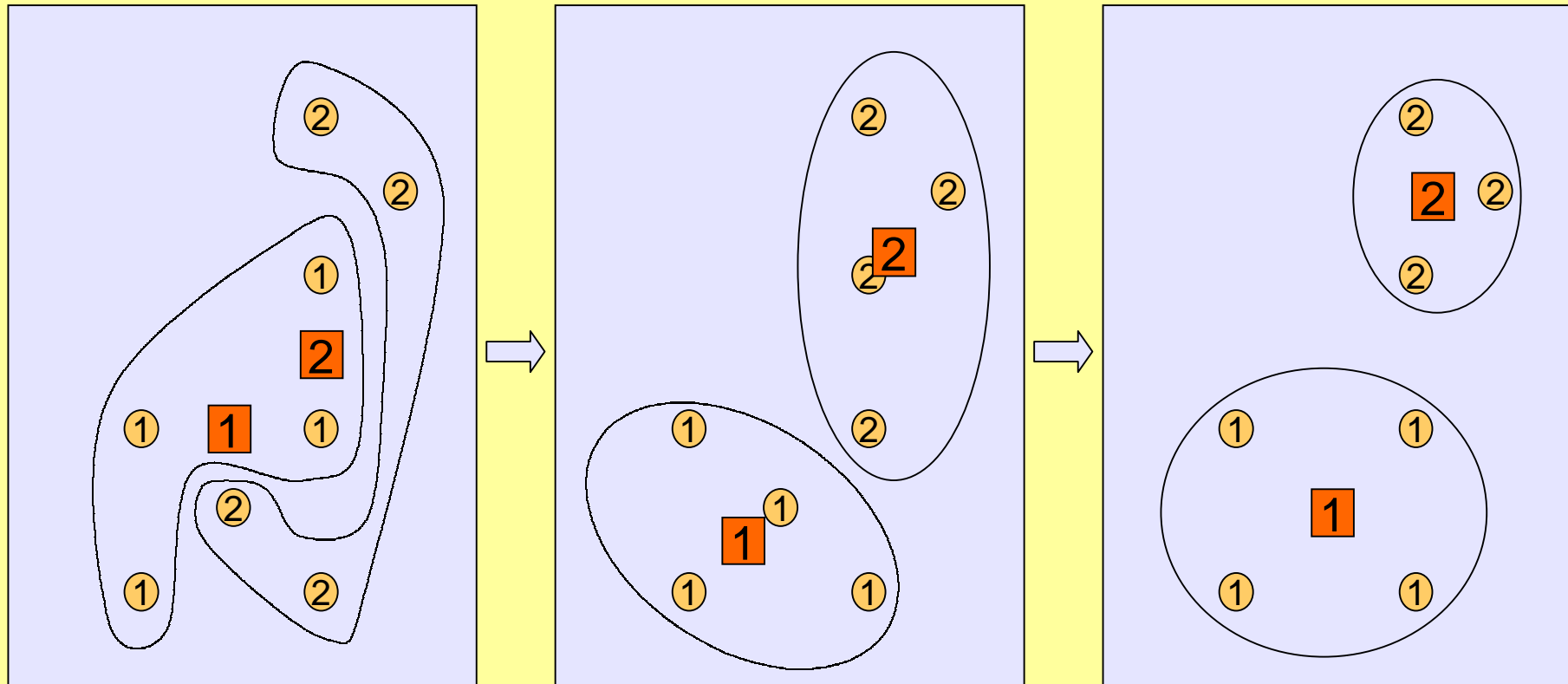
Clustering Methodology

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

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
 2. Calculation of cluster centroids
 3. Determining closest cluster (centroid)
 4. Re-arrangement of cluster membership

k-Means Clustering : Illustration



k-Means Parameters

- Number of clusters k
- Initial cluster assignment:
 - Random clusters
 - Agglomerative hierarchical clusters
- Distance measure:
 - Minkowski metric
 - Cosine
 - Kullback-Leibler divergence

Cluster Initialisation

- Random clusters
- Agglomerative hierarchical clusters
 - Single-linkage
 - Complete-linkage
 - Average verb distance
 - Distance between cluster centroids
 - Ward's method

Distance Measures

- Kullback-Leibler divergence

$$d(v1, v2) = D(p \parallel q) = \sum_i p_i \log(p_i/q_i)$$

- Information radius

$$d(v1, v2) = D(p \parallel (p+q)/2) + D(q \parallel (p+q)/2)$$

- Skew divergence

$$d(v1, v2) = D(p \parallel w * q + (1-w) * p)$$

with weight w set to 0.9

German Verb Choice

- 14 semantic verb classes for 57 German verbs
- Relation to (Levin 1993)
- Consistency with (Schumacher 1986)
- Class size: 2-6
- High and low frequency verbs: $8 \leq \text{freq} \leq 31,710$
- Basis: semantic intuition

German Semantic Verb Classes

1. *Aspect*: anfangen, aufhören, beenden, beginnen, enden
2. *Propositional Attitude*: ahnen, denken, glauben, vermuten, wissen
3. *Transfer of Possession (Obtaining)*: bekommen, erhalten, erlangen, kriegen
4. *Transfer of Possession (Supply)*: bringen, liefern, schicken, vermitteln, zustellen
5. *Manner of Motion*: fahren, fliegen, rudern, segeln
6. *Emotion*: ärgern, freuen
7. *Announcement*: ankündigen, bekanntgeben, eröffnen, verkünden
8. *Description*: beschreiben, charakterisieren, darstellen, interpretieren
9. *Insistence*: beharren, bestehen, insistieren, pochen
10. *Position*: liegen, sitzen, stehen
11. *Support*: dienen, folgen, helfen, unterstützen
12. *Opening*: öffnen, schließen
13. *Consumption*: essen, konsumieren, lesen, saufen, trinken
14. *Weather*: blitzen, donnern, dämmern, nieseln, regnen, schneien

Verb Description: Feature Definition

- Syntactic subcategorisation frames
- Prepositional phrases
- Selectional preferences
- Alternation behaviour
- Adjunct usage
- Morphological properties
- Auxiliary selection
- Voice
- Aktionsart

Subcategorisation Frame Elements

n	noun phrase (case: nominative)
a	noun phrase (case: accusative)
d	noun phrase (case: dative)
r	reflexive pronoun
p	prepositional phrase
x	expletive <i>es</i>
i	non-finite clause
s-2	finite verb second clause
s-dass	finite <i>dass</i> -clause
s-ob	finite <i>ob</i> -clause
s-w	indirect <i>wh</i> -question
k	copula construction

Examples:

- na
- np
- npr
- ns-dass

Lexical Verb Information: frame

glauben
`to think/believe`

ns-dass	0.28
ns-2	0.27
np	0.10
n	0.09
na	0.08
ni	0.05
nd	0.03
nad	0.02
nds-2	0.01

Prepositional Phrase Types

- Akk: an, auf, bis, durch, für, gegen, in, ohne, um, unter, vgl, über
- Dat: ab, an, auf, aus, bei, in, mit, nach, seit, unter, von, vor, zu, zwischen, über
- Gen: wegen, während
- Nom: vgl

Examples: Akk.an, Dat.nach, Gen.wegen, Nom.vgl

Lexical Verb Information: frame+pp

reden
`to talk`

np		0.36
np:Akk.über	`about`	0.12
np:Dat.von		0.11
np:Dat.mit	`with`	0.07
np:Dat.in	`in`	0.02

Random Clustering Input

- konsumieren kriegen vermuten
- anfangen
- ahnen bekanntgeben bestehen **fahren fliegen** liegen nieseln pochen
- aufhören **bekommen erhalten** essen insistieren regnen segeln vermitteln
- beginnen freuen interpretieren
- rudern saufen schneien ärgern
- eröffnen folgen glauben
- zustellen
- charakterisieren dämmern stehen
- blitzen verkünden wissen
- beschreiben **dienen** donnern schließen **unterstützen**
- beenden darstellen **liegen sitzen**
- ankündigen denken enden lesen schicken öffnen
- beharren bringen erlangen helfen trinken

Clustering Result: frame

- ahnen vermuten wissen
- denken glauben
- *anfangen aufhören* beharren bestehen pochen rudern
- *beginnen enden* fahren fliegen liegen segeln sitzen stehen
- dienen folgen helfen
- nieseln regnen schneien
- dämmern
- **blitzen donnern** insistieren
- freuen ärgern
- *lesen saufen* schließen öffnen
- **essen konsumieren** kriegen **trinken** verkünden
- ankündigen beenden bekanntgeben bekommen **beschreiben bringen**
charakterisieren darstellen *erhalten erlangen* eröffnen interpretieren
liefern schicken unterstützen *vermitteln*
- zustellen

Clustering Result: frame+pp

- ahnen vermuten wissen
- denken glauben
- *anfangen aufhören beginnen* beharren *enden* insistieren rudern
- liegen sitzen stehen
- dienen folgen helfen
- nieseln regnen schneien
- dämmern
- blitzen donnern segeln
- *bestehen* fahren fliegen *pochen*
- freuen ärgern
- essen konsumieren saufen trinken verkünden
- bringen eröffnen lesen liefern schicken *schließen* vermitteln *öffnen*
- ankündigen beenden bekanntgeben *bekommen* beschreiben charakterisieren darstellen *erhalten erlangen* interpretieren *kriegen* unterstützen
- zustellen

Extension of Verb Classification

- Number of verbs increases
- Number of verb classes increases

⇒ Extend feature description

⇒ Intuition: add selectional preferences

Problems on Number of Features

- Number of features in general
- Number of features compared to number of objects
- More features → more detailed information
- More features → sparse data problem
- More features → emphasis on few features

Problems on Feature Selection

- It's not always the linguistically most important features which carry the most important information for clustering (optimisation: nr, npr, nas-2, nrs-dass).
- PP information supports clustering (or due to verb choice?), but selectional preferences seem unstable.
- What is the optimal combination of features? Either feature combination is too general or too specific.
- Feature definition on different levels of verb description?
- Linguistic choice of features \leftrightarrow overfitting
- Finding of new/surprising verb properties

Problems on Selectional Preferences

- Means of preference role definitions: GermaNet
- Number of preference roles: 2 / 3 / 15 / 37,666
- Inclusion of selectional preferences:
 - (a) single slot description
 - independent argument assumption
 - probability distribution at risk
 - (b) slot combination description
 - number of features explodes: $15^1/15^2/15^3$
 - magnitude of probabilities varies severely
 - restriction to cut-off set of preference roles?

Limits on Data and Methodology

- Evaluation of corpus data for verb description
- Limit of the data in clustering:
manual classification / manual initialisation?
- Demands on clustering result:
fine-grained clusters or rough assignment?
- k-Means suitable clustering algorithm?

Linguistically plausible verb description
↕
Algorithmically useful feature definition

Verb-Frame Corpus Data on Selectional Preferences

nd	behagen, gefallen, guttun, mißfallen, schaden angehören, entstammen, beitreten helfen, assistieren, dienen	<i>please</i> <i>belong</i> <i>help</i>
nad	aufbürden, zumuten, überantworten bescheinigen, nachsagen ankreiden, anlasten, vorwerfen	<i>impose</i> <i>attest</i> <i>accuse</i>
nr	wohlfühlen, zurechtfinden verirren rächen	<i>get along</i> <i>get lost</i> <i>revenge</i>
ndr	widersetzen, zuwenden, nähern, widmen, hingeben	<i>dedicate</i>

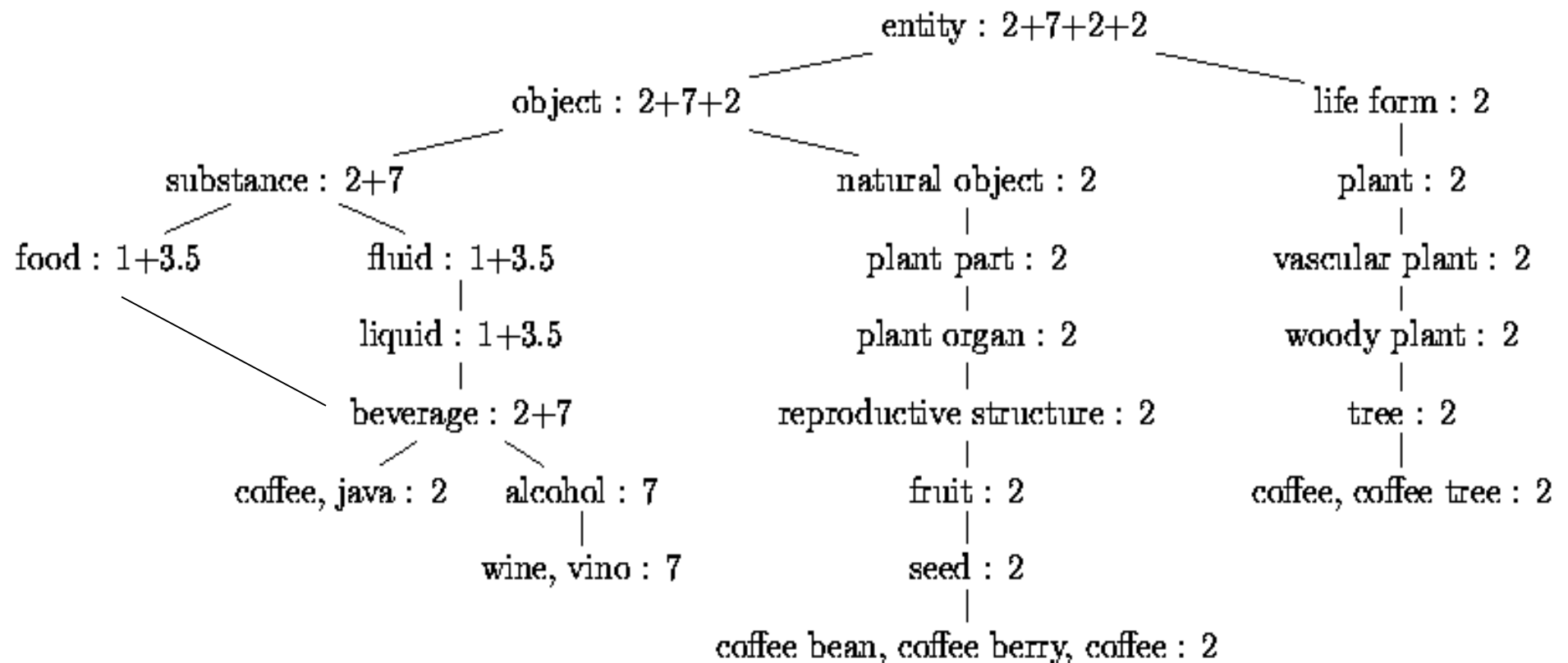
Experiment Setup on Feature Selection

- Filter frames according to linguistic analysis.
- Filter linguistically relevant frame arguments:
 - (a) n, na/na, nd/nd, nad, ns-dass
 - (b) n, na-n:a, nd-n:d, nad, ns-dass
- Distinguish number of selectional preference roles: 2/15.
- Define features on different levels of description.
- Try coarse clustering: 15 instead of 43 classes.
- Try clustering of large number of verbs:
844 with corpus frequency > 500.

Selectional Preference Roles: GermaNet Synsets

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

Selectional Preference Acquisition



Lexical Verb Information: frame+pp+pref(slot)

verfolgen

`to follow´

na	0.64
na_NP.Akk(Situation)	0.16
na_NP.Akk(Kognitives Objekt)	0.12
na_NP.Akk(Zustand)	0.09
na_NP.Akk(Sache)	0.07
na_NP.Akk(Attribut)	0.06
na_NP.Akk(Lebewesen)	0.05
na_NP.Akk(Ort)	0.05
na_NP.Akk(Struktur)	0.02
na_NP.Akk(Zeit)	0.01
na_NP.Akk(Kognitiver Prozess)	0.01

Examples:

Ziel	86
Strategie	27
Politik	25
Interesse	22
Konzept	17
Entwicklung	16
Kurs	14
Spiel	12
Plan	11
Spur	11
Programm	9
Weg	9
Projekt	9
Prozess	8
Zweck	7

Lexical Verb Information: frame+pp+pref(comb)

essen
`to eat`

na	0.42
na_Lebewesen:Nahrung	0.14
na_Lebewesen:Sache	0.07
na_Lebewesen:Lebewesen	0.05
na_Lebewesen:Attribut	0.02
na_Lebewesen:Zeit	0.01
na_Lebewesen:Substanz	0.01
na_Lebewesen:KognitivesObjekt	0.01
na_Lebewesen:Struktur	0.01
na_Situation:Nahrung	0.01
na_Sache:Nahrung	0.01
na_KognitivesObjekt:Nahrung	0.01
na_Struktur:Nahrung	0.01

Examples (n):

man	32
Mensch	5
Kind	5
Frau	3
Bürger	3

Examples (a):

Fleisch	34
nichts	29
etwas	19
Uhr	17
Brot	14
Fisch	13
Suppe	7
Eis	6
Sache	6
Gemüse	6
Kartoffel	5
Kuchen	5
Wurst	5
Pizza	4
Tier	4

Clustering Result: frame+pp+pref

168 verbs
43 classes

(part 1)

- ahnen *bemerken erfahren feststellen* vermuten wissen
- anfangen aufhören hasten riechen
- ankündigen anordnen bekanntgeben empfinden interpretieren scheuen
- basieren beharren beruhen *pochen*
- bedürfen dienen folgen helfen
- beenden beschreiben charakterisieren eröffnen registrieren unterstützen veranschaulichen
- beginnen bestehen enden existieren
- beibringen leihen schenken vermachen
- bekommen benötigen brauchen erhalten erlangen erneuern gründen herstellen kriegen
realisieren wahrnehmen
- bestimmen bringen darstellen erzeugen geben hervorbringen liefern produzieren schicken
stiften treiben vermitteln vernichten
- bilden erhöhen festlegen senken steigern vergrößern verkleinern
- blitzen insistieren rotieren
- demonstrieren lehren rufen verkünden
- denken folgern glauben versichern
- donnern eilen gleiten kriechen rennen starren
- drehen ergeben stützen ärgern
- entfernen schließen setzen spenden öffnen

Clustering Result: frame+pp+pref

168 verbs
43 classes

(part 2)

- **erhoffen wünschen**
- erkennen exekutieren **hören** lesen **sehen**
- **erwachsen resultieren**
- **essen konsumieren trinken**
- **fahren fliegen fließen gehen klettern laufen wandern**
- flüstern heulen schleichen
- freuen fühlen
- fürchten **versprechen** wollen **zusagen**
- **legen** präsentieren **stellen**
- **grinsen** grübeln **jammern klagen** kommunizieren **lachen lächeln schreien weinen**
- gähnen lamentieren
- leben **nachdenken** reden **spekulieren** sprechen verhandeln
- **liegen** segeln **sitzen stehen**
- **nieseln regnen schneien**
- phantasieren saufen
- **renovieren reparieren**
- töten unterrichten
- vorführen **zustellen überscheiben**

Clustering Result: frame+pp+pref

168 verbs
15 classes

- **ahnen bemerken erfahren erkennen feststellen fürchten hören lesen rufen unterrichten verkünden vermuten wissen**
- **anfangen aufhören beginnen enden** helfen korrespondieren rudern
- **ankündigen anordnen** beenden **bekanntgeben bekommen benötigen beschreiben bestimmen brauchen charakterisieren darstellen empfinden erhalten erlangen erneuern erzeugen eröffnen** exekutieren herstellen hervorbringen interpretieren krieg realisieren registrieren scheuen **sehen** stiften unterstützen veranschaulichen vermitteln **wahrnehmen**
- **basieren beharren beruhen insistieren pochen**
- bedürfen dienen **erwachsen folgen resultieren**
- beibringen **erhoffen leihen schenken vermachen wünschen**
- bilden bringen entfernen **geben** gründen **legen** lehren **liefern produzieren präsentieren** renovieren reparieren **schicken schließen setzen spenden stellen** treiben verkleinern vernichten versprechen vorführen wollen zusagen **zustellen öffnen überschreiben**
- **bestehen blitzen** demonstrieren donnern eilen existieren fahren fliegen fließen gehen gleiten gähnen hasten klettern kriechen laufen leben liegen rennen riechen **rotieren segeln sitzen starren stehen wandern**
- dekorieren dämmern **essen konsumieren trinken**
- **denken** folgern **glauben** versichern
- drehen ergeben **erhöhen** festlegen **senken steigern** stützen **vergrößern**
- **ekeln freuen fühlen ärgern**
- **eliminieren** erniedrigen **töten** ängstigen
- flüstern grinsen grübeln heulen jammern klagen kommunizieren lachen lamentieren lächeln nachdenken phantasieren reden saufen schleichen schreien spekulieren sprechen verhandeln weinen
- **niesel**n regnen **schneien**

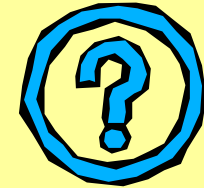
Clustering Result: frame+pp+pref

844 verbs
100 classes

- betragen einsparen investieren kosen kosten
- beziffern schätzen veranschlagen
- ausweiten durchsetzen erhöhen reduzieren steigern verbessern verdoppeln verringern verschieben verschärfen verändern vollziehen zurückziehen
- malen proben raten singen spielen
- entscheiden handeln streiten verfügen verhandeln
- ergänzen erklären erläutern erzählen kommentieren verkünden
- appellieren dürfen können müssen sollen vermögen versuchen
- behaupten betonen einräumen versichern
- ahnen bezweifeln merken weißes wissen

- abhalten ablösen ansprechen befragen behandeln benennen bestrafen betreuen einbeziehen empfangen entlassen ermorden ernennen erschießen festnehmen gebärden kennenlernen schicken treiben töten umbringen unterbringen unterrichten verhaften verletzen versetzen versorgen vertreiben wählen zitieren

Discussion



- Careful usage of selectional preferences helps clustering.
- Fine role distinction on selected frames appears optimal.
- Selectional preferences on selected arguments outperform preference combinations on frame arguments.
- Feature description on different levels improves clustering.
 - Linguistically plausible and methodologically useful selection of features for verb description.
- k-Means only useful for small sets of objects.
 - Alternative: classification.

Further Work

- Extension of manual verb classification
- Re-run of k-Means clustering
- Application of classification technique: SVM
- Application of soft clustering technique: LSC
- Variation on number of clusters
- Definition of NLP application for automatic verb classes