

**Cross-Linguistic Comparison of Semantic Verb
Classes in English, German and Spanish**

Ph.D. Dissertation Proposal

Sabine Schulte im Walde

Institute for Natural Language Processing

Task Definition

1. cross-linguistic construction of semantic verb classes
2. empirical evidence from large-coverage corpus data

Motivation

- central role of verb in meaning and structure of sentences
- idiosyncratic versus generalised lexical verb information
- cross-linguistic comparison of verb classes and verb features
- empirical verification of close relationship between lexical syntax and semantics
- support of NLP-tasks
(lexicography, parsing, machine translation, information retrieval)

English Verb Classes and Alternations

(Levin 1993)

- Verbs of Motion
- Verbs of Communication
- Verbs of Contact by Impact
- Verbs of Sending and Carrying
- Verbs of Change of Possession
- Verbs of Learning
- etc.

English Verb Class *Vehicle Names*

Verb class *Vehicle Names* (sub-class of *Motion Verbs*):

balloon, bicycle, bike, boat, bobsled, bus, cab, canoe, caravan, chariot, coach, cycle, dogsled, ferry, gondola, helicopter, jeep, jet, kayak, moped, motor, motorbike, motorcycle, parachute, punt, raft, rickshaw, rocket, skate, skateboard, ski, sled, sledge, sleigh, taxi, toboggan, tram, trolley, yacht

Alternation Behaviour of Vehicle Names Verbs

- (1) *Intransitive Use*, possibly followed by a path:
 - a. They skated.
 - b. They skated along the canal/over the bridge.
- (2) *Induced Action Alternation* (some verbs):
 - a. He skated Penny around the rink.
 - b. Penny skated around the rink.
- (3) *Locative Preposition Drop Alternation* (some verbs):
 - a. They skated along the canals.
 - b. They skated the canals.
- (4) *Resultative Phrase*:

Penny skated her skate blades blunt.

Causative Alternation

- (5)
- a. Mary closes the door.
The door closes.
 - b. Maria schließt die Tür.
Die Tür schließt sich.
 - c. Maria cierra la puerta.
La puerta se cierra.

Object-Drop Alternation

- (6)
- a. John eats fish.
John eats.
 - b. Jupp ist Fisch.
Jupp ist.
 - c. Juan come pescado.
Juan come.

Constructing Semantic Verb Classifications

1. identification of verbs
2. lexical decomposition and morphological properties of verbs
3. description of verb alternation behaviour
4. assigning verbs to classes

Identification of Verbs

- which verbs?
- how many verbs?
- restrictions on verbs (e.g. frequency, polysemy)?
- how to do cross-linguistic selection
(e.g. by translation, or independently)?

Lexical Decomposition

- aspectual class
(state, activity, accomplishment, achievement)
- first-order aspect calculus (BECOME, CAUSE, DO, AT)

$[DO(x, [\phi(x, y)])] CAUSE [BECOME \psi]$

$[DO(x, [\phi(x, y)])] CAUSE [BECOME < STATE > (y)]$

$[DO(x, [break_k(x, y)])] CAUSE [BECOME broken(y)]$

Morphological Properties

- origin, e.g. *enter*, *pass*, *separate* originate from Romance
- morphemes and morphological change in alternations,
e.g. German *enden* - *beenden* in causative/inchoative alter-
nation
- verb lexicalisation patterns:
 - manner (e.g. *float*)
 - path (e.g. *subir*)

Description of Verb Alternation Behaviour

- which alternations?
- description of alternations?
- how to determine the alternations for each verb (e.g. lexicon, introspection, lexical decomposition factors)?
- cross-linguistical mapping of alternations?

Assigning Verbs to Classes

- which classes?
- how many classes?
- how detailed/fine?
- cross-linguistic mapping?
- multiple assignment?

Providing Empirical Evidence for the Classification System

1. construction of (restricted) corpora
for English, German and Spanish
2. definition and training of context-free grammars
for English, German and Spanish
3. extraction of syntactic and semantic information
from probabilistic models
4. application of clustering techniques,
e.g. latent class analysis, neural networks, decision trees

Related Work

- Levin classification (1993)
 - definition of syntactic signatures for verbs, based on example sentences from Levin (Dorr and Jones, 1996)
 - definition of meaning components, event structure and diathesis alternation for grouping verbs into three Levin classes (Fernandez et al., 1998)
 - contribution of syntactic verb features to classify verbs by machine learning techniques (Stevenson and Merlo, 1999)
 - automatic verification of Levin classes (Schulte im Walde, 1998)
 - empirical investigation of alternation types in corpus data (Lapata, 1999)

- cross-linguistic classifications
 - Bangla, Korean and German (Jones et al., 1994)
 - Japanese, Hindi, Bengali and Greek (Nomura et al., 1994)
 - English and German (Frense and Bennett 1996)
 - French (Saint-Dizier, 1996)
 - Spanish and Catalan (Fernandez et al., 1998)

Inference of Semantic Classes

1. Induction of subcategorisation frames for verbs from a large corpus
2. Definition of selectional preferences for the subcategorisation frames
3. Clustering of the verbs into semantic verb classes

Induction of Subcategorisation Frames (1)

- British National Corpus (BNC)
- English context-free grammar
- Robust statistical head-entity parser (Carroll/Rooth 1998)
 - 5.5 million maximum probability (Viterbi) parses
- LISP extraction tool
 - subcategorisation frame tokens, annotated with lexical heads:
proved subj*distinction ap*difficult
took subj*this obj*forms
argued subj*he pp*against*type
serve subj*comparison obj*us pp*as*example
excelled subj*nobody obj*him pp*in*judgement

Induction of Subcategorisation Frames (2)

- Lemmatisation
- Generalisation
 - 88 subcategorisation frame types with frequency > 2000
- Joint frequencies of BNC-verbs and subcategorisation frame types:

give	subj	758
give	subj:adv	105
give	subj:ap	58
give	subj:obj	9,982
give	subj:obj:adv	498
give	subj:obj:ap	60
give	subj:obj:as	53
give	subj:obj:obj	13,430
...		

Selectional Preferences for Subcategorisation Frames

- Goal: preferential ordering on conceptual classes for the argument slots in the frame types
- Basis: lexical heads in frame tokens
- Example:
 - $drink < \text{subj:obj} >$
 - $\Rightarrow drink < \text{subj:}\{coffee, milk, beer, \dots\} >$
 - $\Rightarrow drink < \text{subj:beverage} >$
- Selectional preference (Resnik 1993/1997):
amount of information a verb provides about its semantic argument classes
- Conceptual classes \approx WordNet synsets

Lexical Heads in Frame Tokens

give	subj:obj:obj	13430	35855
combination		5	
doctor		28	
government		32	
people		37	
result		26	
a_little		16	
all		22	
boy		26	
employee		21	
performance		16	
rise		64	
advantage		65	
benefit		41	
cheque		15	
grin		31	
idea		133	
name		116	

Selectional Preference over WordNet Classes – Examples

- *break* <subj:pp.into> ⇒ *break* <offender:smile>
- *drive* <subj:obj> ⇒ *drive* <person:artifact>
- *eat* <subj:obj> ⇒ *eat* <living entity:food>
- *swim* <subj:pp.in> ⇒ *swim* <fish:body of water>

Clustering Verbs into Semantic Verb Classes

- Verbs: 153 different verbs with 226 verb senses from 30 different semantic classes in Levin's classification
- Information:
 - (A) syntactic information about the subcategorisation frames
 - (B) ditto, refined by their selectional preferences
- Clustering approaches:
 - Iterative clustering (Hughes 1994)
 - Unsupervised latent class analysis (Rooth 1998)

Resulting Clusters (1): *Desire*

Verb	Frame	Probability
need	subj:to	0.38
	subj:obj	0.32
	subj	0.10
	subj:obj:to	0.05
	subj:obj:pp.for	0.02
like	subj:to	0.34
	subj:obj	0.34
	subj	0.14
	subj:obj:adv	0.04
	subj:obj:obj	0.03
want	subj:to	0.53
	subj:obj	0.15
	subj	0.11
	subj:obj:to	0.10
	subj:to:adv	0.02
desire	subj:obj	0.25
	subj	0.24
	subj:to	0.20
	subj:obj:to	0.07
	subj:sent	0.02

Resulting Clusters (2): *Manner of Motion*

Verb	Frame	Probability
roll	subj(PhysObject)	0.24
	subj(PhysObject):adv	0.10
	subj(Agent):obj(PhysObject)	0.07
	subj(LifeForm):obj(PhysObject)	0.07
fly	subj(Agent):obj(Part)	0.05
	subj(PhysObject)	0.34
	subj(PhysObject):adv	0.12
	subj(LifeForm):obj(PhysObject)	0.07
move	subj(LifeForm):pp.to(LifeForm)	0.05
	subj(LifeForm):pp.to(Agent)	0.04
	subj(PhysObject)	0.20
	subj(PhysObject):adv	0.11
	subj(Part)	0.09
	subj(Group):adv	0.04
	subj(Part):adv	0.04
		0.04