



Lexikalische Ressourcen

Claudia Kunze

kunze8@googlemail, kunze@cl.uni-heidelberg.de

Computational Lexicography

Themen

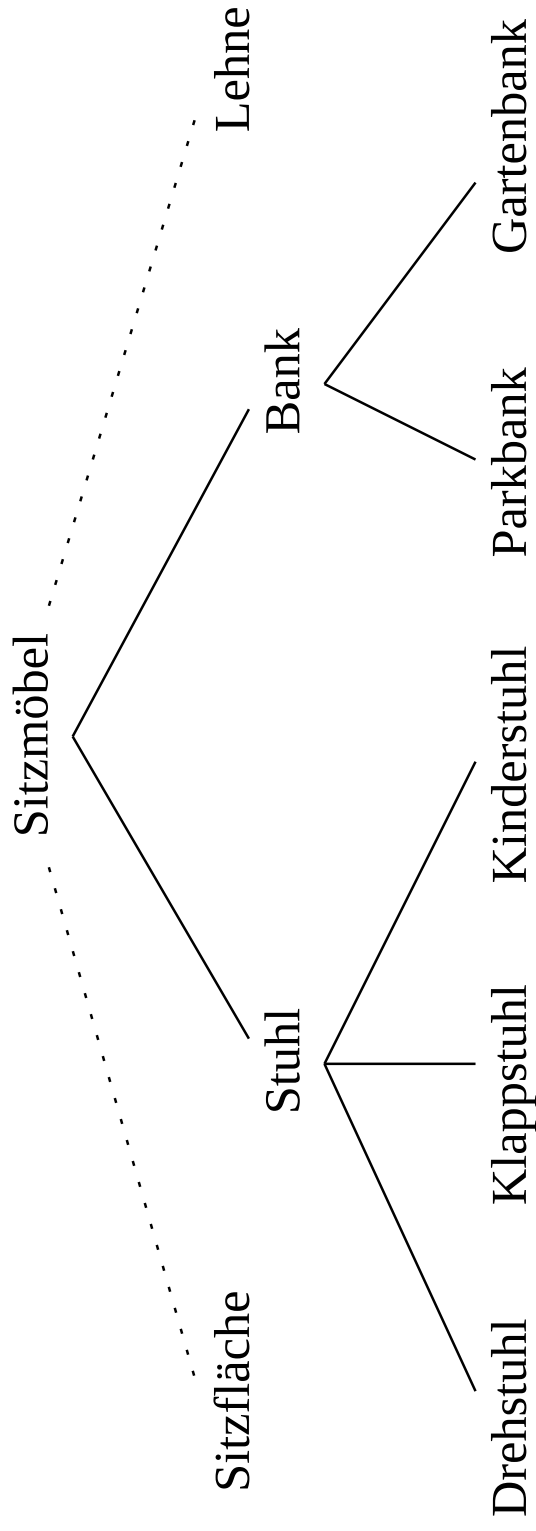
- Typen lexikalischer Datenbanken
- GermaNet als monolinguales lexikalisch-semantisches Wortnetz
- EuroWordNet als polylinguales lexikalisch-semantisches Wortnetz
- FrameNet (Referatsfolien von Eva Mújdríczová aus SoSe 2008)
- Ontologien

Diese Lehreinheit behandelt GermaNet, Euro-WordNet und FrameNet.

GermaNet, ein monolinguales Wortnetz

- Entwicklung eines lexikalisch-semanticischen Wortnetzes für das Deutsche als Grundlagenressource für die Lesartendisambiguierung (seit 1996)
- Aufbau einer Ressource, die den deutschen Grundwortschatz (Nomina, Verben und Adjektive) abdeckt, und die wichtigsten lexikalisch-semanticischen Relationen zwischen den Eintragungswörtern abbildet
- technische und strukturelle Orientierung am Modell des Princeton WordNet, der “Mutter aller Netze”, mit einigen Modifikationen zur Erfassung der Besonderheiten der deutschen Sprache (z.B. Umlaute)

GermaNet-Ausschnitt *Sitzmöbel*



Quellen

Manueller Aufbau des GermaNet (‘from scratch’), unter Berücksichtigung folgender Quellen:

- Thesaurus Wehrle/Eggers (1981);
- Duden–8 (‘sinn– und sachverwandte Wörter’);
- andere monolinguale allgemeinsprachliche Wörterbücher und Enzyklopädien;
- Princeton WordNet, PAROLE Lexika;
- Korpora (ECI Corpus, Tübinger Zeitungskorpus, DEREKO, Wortwarte);
- Wissen und Intuition der Lexikographen.

Aktuelle Abdeckung des GermaNet

Datenabdeckung am 30.04.2008 (Version 5.1)

- 57 776 Synsets
- 81 773 Lexical Units
- 72 057 Literals
 - 1.42 Lexical Units per Synset
 - 1.13 Lesarten per Literal
- Konzeptuelle Relationen: 68 997
 - davon 62590 Hyperonymierelationsinstanzen
 - davon 3925 Meronymierelationsinstanzen
- Lexikalische Relationen: 12 042
 - davon 1742 Antonymierelationsinstanzen
 - davon 1681 Pertonymierelationsinstanzen

Synsets und Relationen

Das *Synset* dient in Wortnetzen als zentrale Organisationseinheit. Lexical Units mit identischen oder sehr ähnlichen Bedeutungen werden zu Synsets zusammengefasst. In GermaNet gibt es zwei grundlegend unterschiedliche Arten von Relationen:

- **Lexikalische Relationen:** Antonymie, Derived_from, Participle gelten zwischen *Lexical Units*;
- **Konzeptuelle Relationen:** Hyponymie, Hyperonymie, Meronymie, Holonymie, kausale Relation, Entailment, assoziative Relation gelten zwischen *Synsets*.

In GermaNet ist eine Wortbedeutung durch die Gesamtheit der Relationen, die sie zu anderen Konzepten und Wortbedeutungen aufweist, bestimmt.

Lexikalische Relationen in GermaNet

- **Synonymie** wird durch das gemeinsame Enthaltensein (mindestens) zweier Lexical Units im selben Synset realisiert: {*Zündholz*, *Streichholz*} 'match', {*erschaffen*, *erzeugen*} 'create', {*fleißig*, *arbeitsam*, *tüchtig*, *emsig*} 'busy';
- **Antonymie** dient der Kodierung lexikalischer Kontraste, wie z.B. *Geburt* 'birth' and *Tod* 'death', *starten* 'start' and *landen* 'land', *schön* 'beautiful' and *hässlich* 'ugly';

Lexikalische Relationen, Forts.

- **Derived_from** erfasst Derivationsrelationen, z.B. deverbale Nominalisierungen (*Beobachtung* ‘observation’, abgeleitet aus *beobachten* ‘observe’) oder denominal Adjektive (*mathematisch* ‘mathematical’, abgeleitet aus *Mathematik* ‘mathematics’);
- **Participle** verbindet Partizipien, die in GermaNet enthalten sind, mit ihrer verbalen Grundform, z.B. *provokierend* ‘provoking’ mit *provizieren* ‘provoke’ oder *entschlossen* ‘determined’ mit *entschließen* ‘decide’.

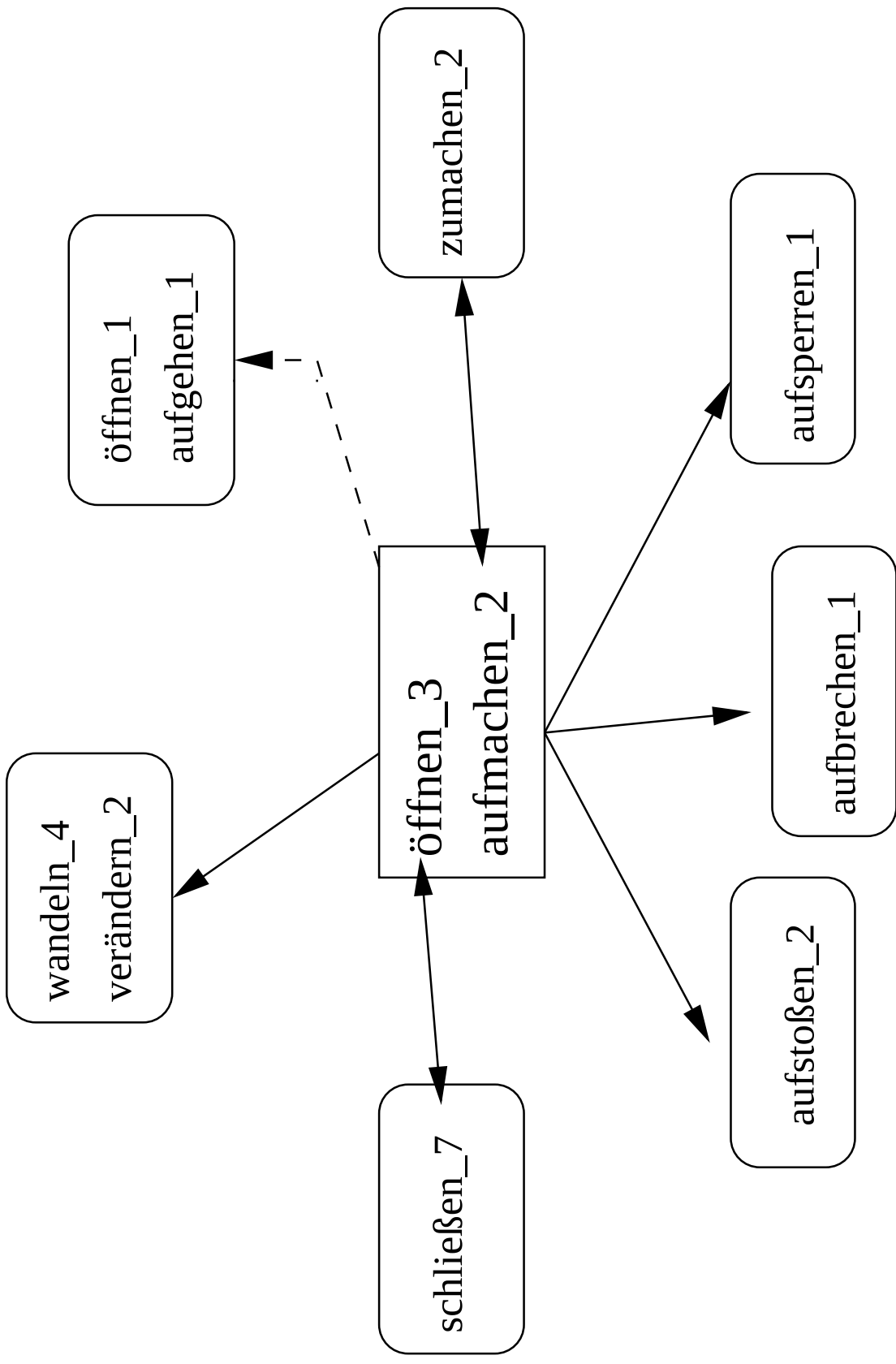
Konzeptuelle Relationen in GermaNet

- **Hyponymie/ Hyperonymie** fungiert als zentrale Relation für die Repräsentation aller Wortarten in GermaNet und “ergibt” die taxonomische Struktur von Wortnetzen. Konzepte werden mit übergeordneten Termen verknüpft wie *Hammer* ‘hammer’ mit *Werkzeug* ‘tool’. Ko-Hyponyme eines gegebenen Konzepts sollten inkompatibel zueinander sein, wie die Hyponyme *Säugling* ‘baby’, *Kleinkind* ‘small child’, *Schulkind* ‘school child’ des Konzepts *Kind* ‘child’;
- **Meronymie/ Holonymie** wird nur auf Nomina angewendet. Prominente Beispiele beziehen sich auf Körperteile oder konstitutive Teile von Artefakten wie *Dach* ‘roof’ als Teil von *Gebäude* ‘building’.

Konzeptuelle Relationen, Forts.

- Weitere Typen sind Mitgliedschaft in einer Gruppe (*Anführer* 'leader' einer *Bande* 'gang'), Material in einer Komposition (*Fensterscheibe* 'window-pane' aus *Glas* 'glass') und geographisches Enthaltensein (*Bretagne* als Teil von *France*);
- Die **CAUSE**-Relation kann wortartenübergreifend angewendet werden. Typische Beispiele sind lexikalische Resultative wie z.B. die Beziehung zwischen *töten* 'kill' und *sterben* 'die' oder *öffnen* 'to open' und *offen* 'open';
- **Entailment** für rückwärtsgerichtete Präsupposition zwischen verbalen Prädikaten, z.B. zwischen *gelingen* 'succeed' und *versuchen* 'try'.

Relationen des kausativen Verbs 'to open'





Relationen/Pointers

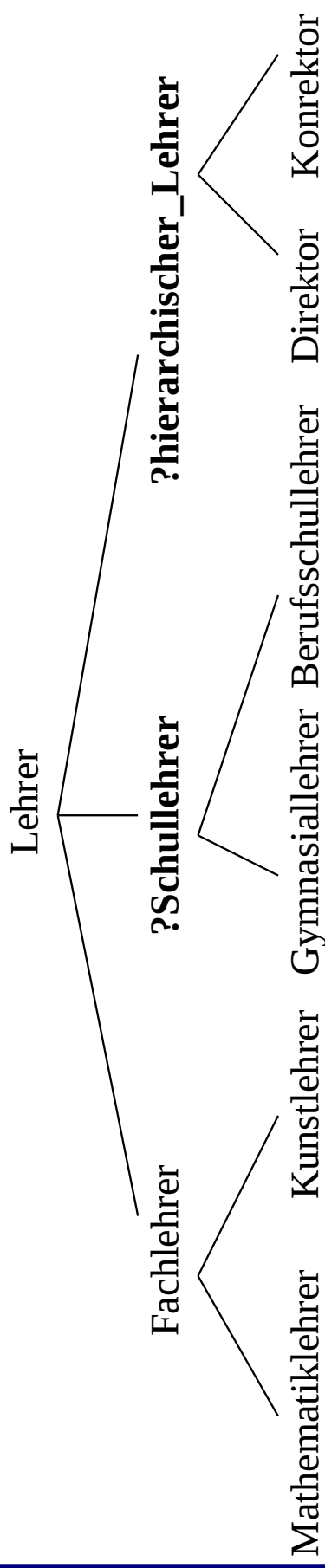
Pointers - Overview

Name	Symbol	Valid Class			Reverse Pointer		Lex
		N	A	V	Name	Symbol	
<u>Antonym</u>		y	y	y	<u>Antonym</u>		y
<u>Hyperonym</u>	@	y	y	y	<u>Hyponym</u>	~	n
<u>Hyponym</u>	~	y	y	y	<u>Hyperonym</u>	@	n
<u>Meronym</u>	%	y	n	n	<u>Holonym</u>	#	n
<u>Holonym</u>	#	y	n	n	<u>Meronym</u>	%	n
<u>Entailment</u>	*	n	n	y			n
<u>Cause</u>	>	n	n	y			n
<u>Also See</u>	^	y	y	y			n
<u>Derived From / Pertonym</u>	\	y	y	y			y
<u>Participle</u>	<	n	y	n			y
<u>Regular Polysemy (not implemented)</u>	.	y	y	y			n
<u>Selectional Restriction (not implemented)</u>		y	y	y			y

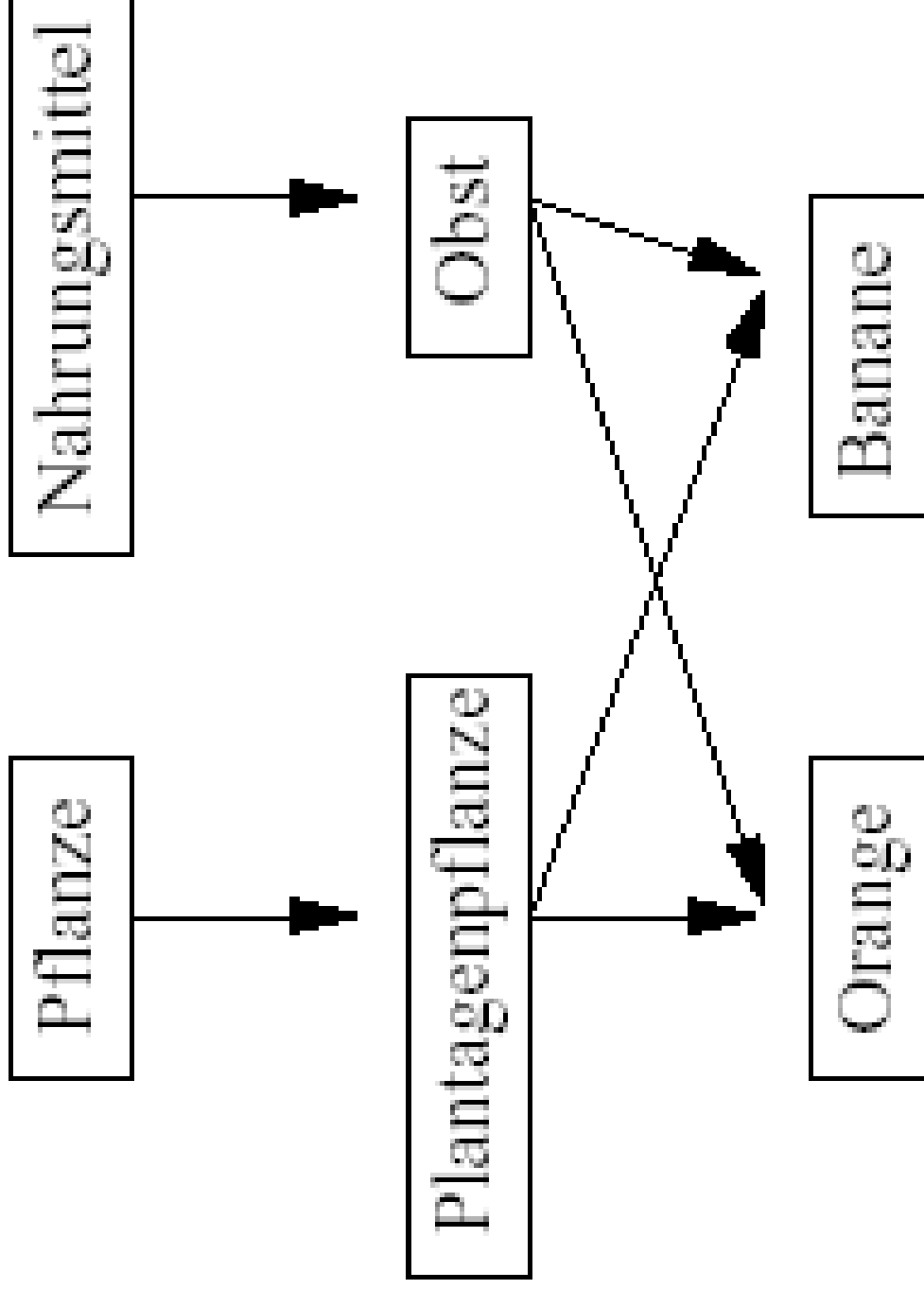
Artifizielle (nicht-lexikalisierte) Konzepte

Um unmotivierte Ko-Hyponymie zu vermeiden, verwendet GermaNet intermediäre Konzeptknoten, die keine lexikalische Realisierung im Deutschen haben. $\Rightarrow$ Künstliche Konzepte führen zu tieferen konzeptuellen Hierarchien. Falls notwendig, können die nicht-lexikalisierten Elemente leicht herausgefiltert werden.

Beispiel aus der nominalen Taxonomie Lehrer ‘teacher’, Mathematiklehrer ‘math teacher’, Gymnasiallehrer ‘high school teacher’ etc.



Kreuzklassifikation von Konzepten



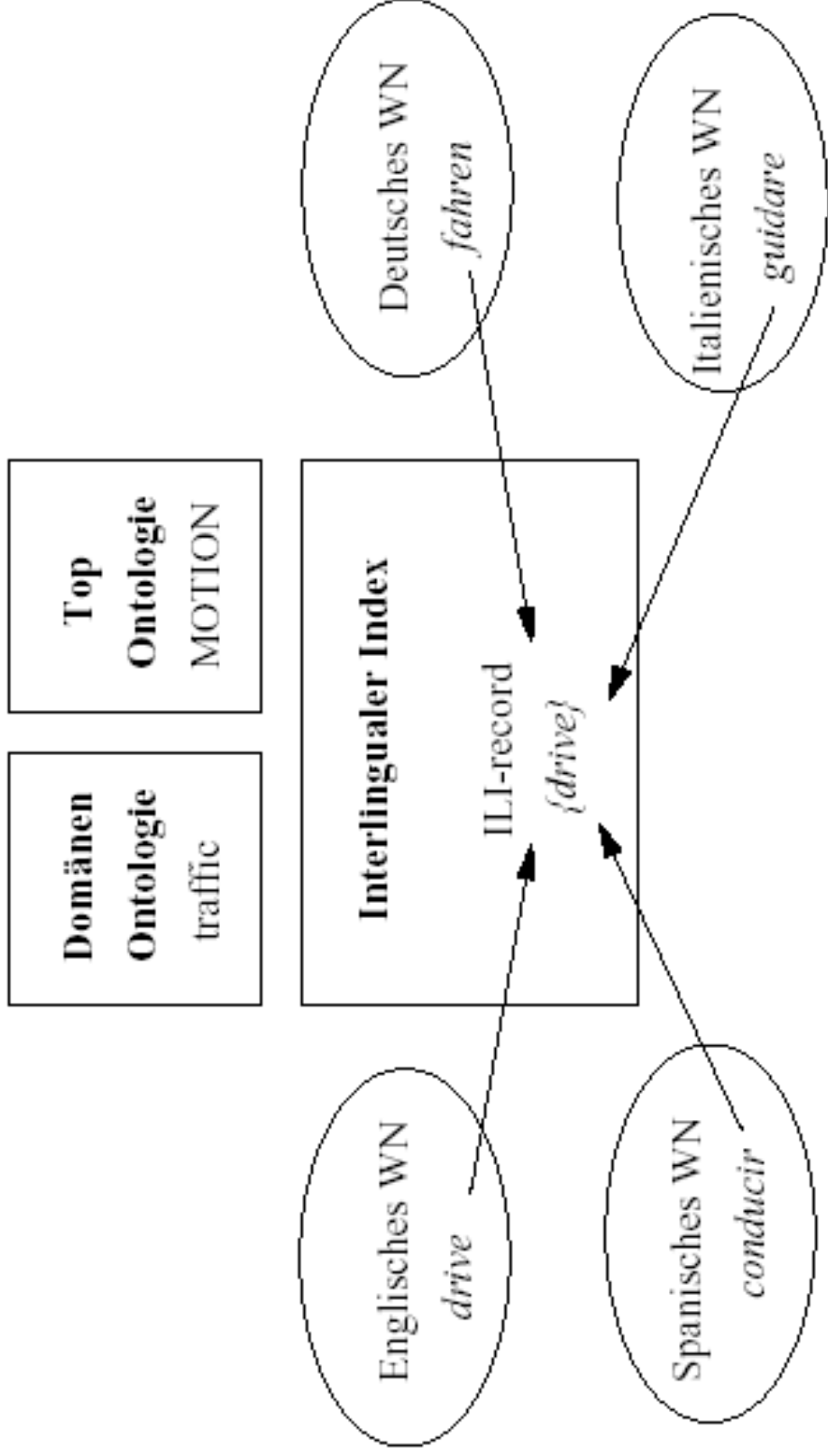
EuroWordNet, ein polylinguales Wortnetz

- Entwicklung einer multilingualen lexikalisch–semantischen Datenbank für 8 europäische Sprachen: Italienisch, Holländisch, Englisch, Spanisch, Französisch, Deutsch, Tschechisch and Estnisch;
- Kodierung generischer and häufiger Konzepte (7 500 – 15 000 Konzepte pro Sprache) und der lexikalisch–semantischen Relationen wie Hyponymie, Meronymie, Antonymie, etc. zwischen Konzepten and Lexical Units der jeweiligen Sprachen;
- Verknüpfung der sprachspezifischen Wordnetze über Äquivalenzrelationen, für den multilingualen Zugriff auf die Konzepte;

EuroWordNet, Forts.

- Anbindung (disambiguerter) Lesarten;
- EuroWordNet wurde zu einer multilingualen Referenz-Ressource für mehrere nachfolgende Initiativen: BalkaNet (Balkan- und andere osteuropäische Sprachen), CoreNet (zur Integration des Chinesischen, Japanischen and Koreanischen), MultiWordNet (Spanisch, Italienisch, Baskisch), etc.;
- Das EuroWordNet Lesarten-Inventar (ILI) fungiert als Quasi-Standard für monolinguale und multilinguale Wortnetz-(Neu-)Entwicklungen.

EuroWordNet-Architektur



Die EuroWordNet-Architektur

- **Sprachunabhängige Komponenten:**
 - Unstrukturierte Liste von “ILI–records”, die den sog. Interlingual Index (ILI) konstituieren;
 - Ontologie, aus einer Top-Ontologie und Domänen-Ontologie bestehend, dient der ausgewogenen Entwicklung individueller Wortnetze.
- **Sprachspezifische Komponenten:**
 - Monolinguale Wortnetze für 8 europäische Sprachen;
 - Sprach-interne Relationen zwischen Konzepten und Lexical Units der jeweiligen Sprachen.

- Monolinguale Wordnetze werden über den Interlingualen Index mittels Äquivalenzrelation verknüpft.

Multilinguale Wordnetze werden in Anwendungen wie dem cross-lingualen Information Retrieval und in der kontrastiven Forschung über Lexikalisierungsmuster verschiedener Sprachen eingesetzt.

Aktivitäten

- Terminologische Erweiterung von Wortnetzen:
 - Akquisition adäquater Subrelationen aus domänenspezifischen Korpora;
 - Erweiterung von Wortnetzen mit domänenspezifischem Vokabular
- Verknüpfung von Wortnetzen mit anderen lexikalischen Ressourcen, z.B. morphologischen und syntaktischen Ressourcen (Language Resources Grid architectures);
- Verknüpfung mit Ontologien wie SUMO;
- Konvertierung der GermaNet-Daten nach OWL (Web Ontology Language);

Weitere Arbeiten

- Gebrauch von Wortnetzen in virtuellen Kursen zur Computerlinguistik (eLearning);
- Standardisierung des Lesarten-Inventars von Wordnetzen (ILI) und die formale Spezifikation von Wortnetzen.