Vernetzungsstrategien und Zugriffsstrukturen in kollaborativ erstellten Lexika

Christian M. Meyer und Iryna Gurevych

TECHNISCHE UNIVERSITÄT DARMSTADT

Arbeiten am UKP Lab zu Sprachressourcen, Bedeutungsalignierung, Standardisierung von Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, Elisabeth Niemann.

DFG-Forschungsnetzwerk „Internetlexikografie“
2. Arbeitstreffen, Berlin, 5.–6. Dezember 2011.

UKP

Forschungsprofil UKP Lab

TECHNISCHE UNIVERSITÄT DARMSTADT

Ubiquitous Knowledge Processing Lab
und Web Research Center Darmstadt

UBIQUITOUS KNOWLEDGE PROCESSING

WeRC

Informatik-orientierte Forschung im Bereich Computerlinguistik und Sprachtechnologie mit Kernkompetenzen in

- Methoden: semantische Ähnlichkeit, Lesartendisambiguierung, Text Mining,…
- Ressourcen: kollaborativ erstellte Lexika auf Wiki-Basis

WIKIPEDIA
The Free Encyclopedia

Wiktionary
[ˈwɪkʃənrɪ] *n.*,
a wiki-based Open Content dictionary

OMEGAWIKI

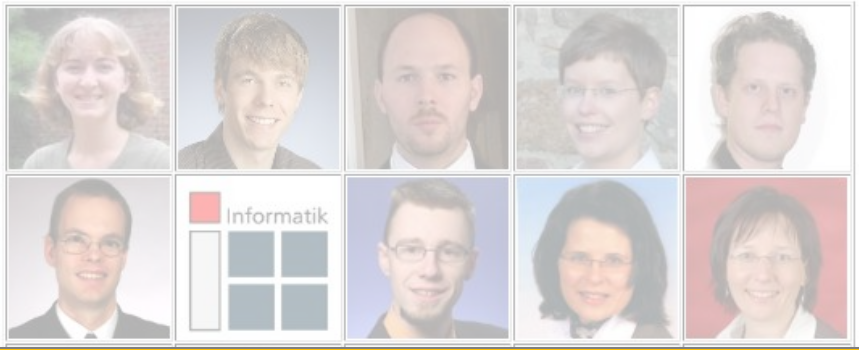
Forschungsprofil UKP Lab

TECHNISCHE UNIVERSITÄT DARMSTADT

Ubiquitous Knowledge Processing Lab
und Web Research Center Darmstadt

UBIQUITOUS KNOWLEDGE PROCESSING

WeRC



Der kollaborative Erstellungsprozess in Wikis eröffnet neue Forschungsgebiete für die Lexikographie:

Collaborative Lexicography

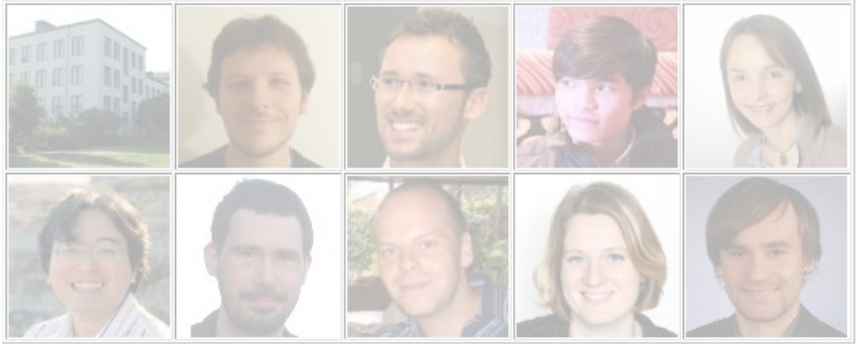
Mining,…

- Ressourcen: kollaborativ erstellte Lexika auf Wiki-Basis

WIKIPEDIA
The Free Encyclopedia

Wiktionary
[ˈwɪkʃənrɪ] *n.*,
a wiki-based Open Content dictionary

OMEGAWIKI



Teil I

VERNETZUNGSSTRATEGIEN

Automatische Alignierung von Wortbedeutungen

Wiktionary

TECHNISCHE UNIVERSITÄT DARMSTADT

Wiktionary
[ˈvɪkʃəˌnɛʁi], *n*
Das freie Wörterbuch
ein Wiki-basiertes freies Wörterbuch

Hauptseite
Themenportale
Zufällige Seite
Inhaltsverzeichnis

Mitarbeit
Eintrag erstellen
Autorenportal
Wunschliste
Literaturliste
Letzte Änderungen

Hilfe

Werkzeuge
Was linkt hierher?
Änderungen an verlinkten Seiten
Spezialseiten
Druckversion
Beständige URL

In anderen Sprachen
Asturianu
Aymar aru
Azərbaycanca
Беларуская
Brezhoneg
Català
Cymraeg
Česky

Eintrag | Diskussion | Lesen | Bearbeiten | Versionsgeschichte | Suche

Ihre **Spenden** helfen, Wiktionary zu betreiben.

Wasser

Wasser (Deutsch) [Bearbeiten]

Dieser Eintrag war 2006 5. **Wort der Woche**.

Dieser Eintrag oder Abschnitt bedarf einer Überarbeitung. Hilf bitte mit, ihn zu verbessern, und entferne anschließend diese Markierung.
Folgendes ist zu überarbeiten: Übers.-Abschnitt (Form, Zuordnung); Herkunft (formal; belegen)

Substantiv, *n* [Bearbeiten]

Silbentrennung:
Was·ser, Plural 1: Was·ser, Plural 2: Wäs·ser

Aussprache:
IPA: [ˈvasɐ], Plural 1: [ˈvasɐ], Plural 2: [ˈvɛsɐ]
Hörbeispiele: Wasser (Info) Wasser (Bairisch) (Info), Plural: Wässer (Info)

Bedeutungen:
[1] *kein Plural:* die chemische Verbindung (Diwasserstoffoxid), der Stoff H_2O in flüssigem Aggregatzustand, die aus Wasserstoff und Sauerstoff zusammengesetzt ist
[2] *auch Plural möglich: siehe Plural 1, poetisch, gehoben:* für Gewässer
[3] *beide Pluralformen, übertragen, umgangssprachlich, zum Teil synonym:* für sehr viele Flüssigkeiten, Lösungen, Emulsionen, die in ihrer Konsistenz dem Wasser ähneln sowie Gewässer und Wässer, die ihrer Herkunft nach, ihrem Vorkommen nach, ihrem Verwendungszweck nach und Ähnlichem benannt werden
[4] *nur im Plural 2 üblich oder umgangssprachlich Wässerchen:* ein alkoholisches Getränk, welches aus vergorenen Früchten oder anderen Teilen der Pflanze gebrannt wurde
[5] *kein Plural:*ein Reinheitsmaß für Diamanten
[6] *Medizin:* krankhafte Ansammlung von Körperflüssigkeiten im Gewebe
[7] *umgangssprachlich, kurz für* Mineralwasser, Tafelwasser

Abkürzungen:
[1] H_2O
[2] Wa, Wa.

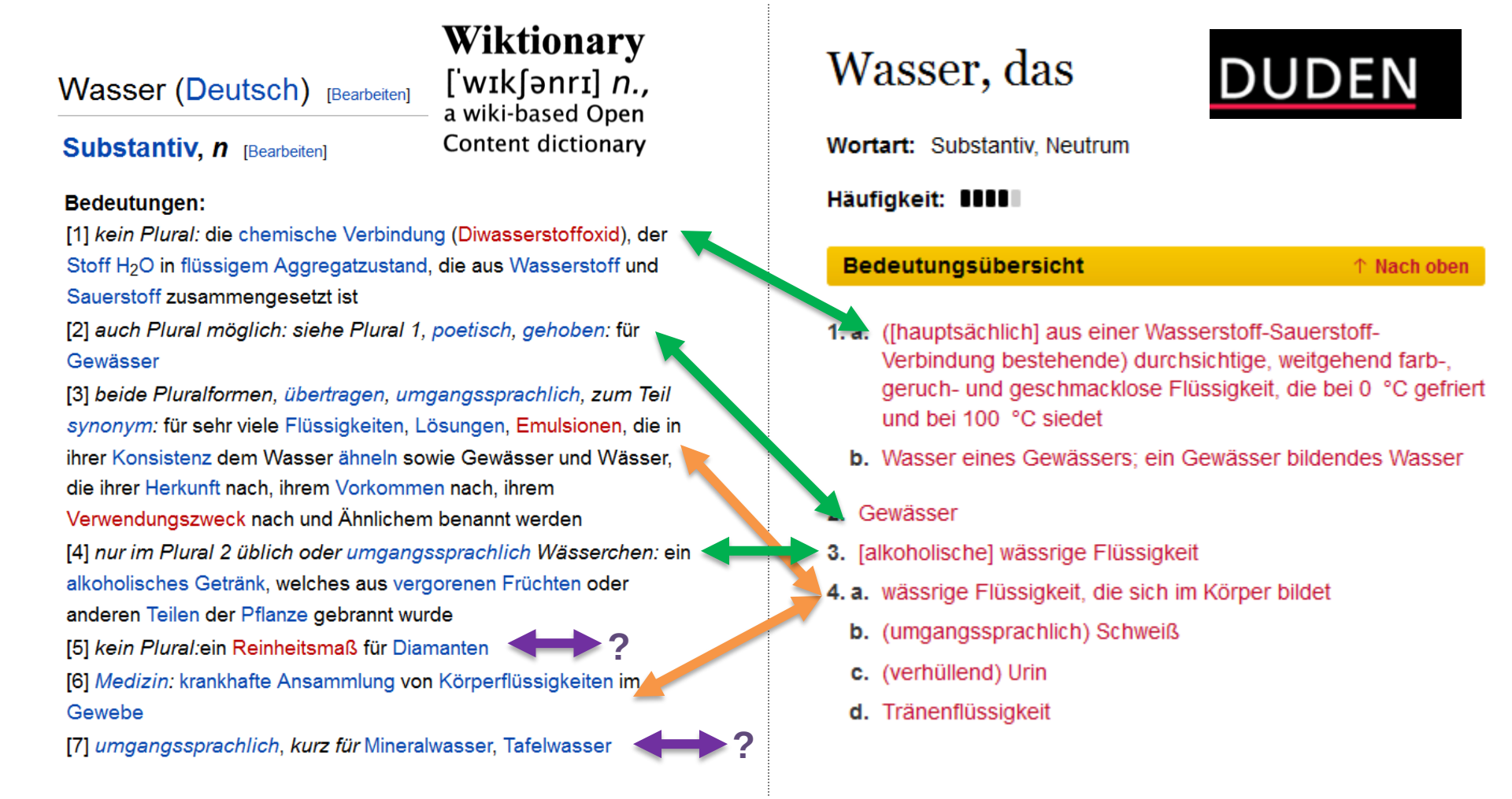
Herkunft:

379.694 Einträge in Englisch

85.574 Einträge in Deutsch

UKP

Bedeutungsalignierung



Warum Bedeutungen alignieren?

TECHNISCHE UNIVERSITÄT DARMSTADT

Analysieren

- Abdeckungsgrad von Wortbedeutungen
- Überlappung zwischen Wörterbüchern

Erweitern

- Neue Wortbedeutungen finden
- Bedeutungsparaphrasen überarbeiten

Anwenden

- Vernetzung auf Wörterbuchportalen auch auf Bedeutungsebene
- Sprachtechnologische Anwendungen

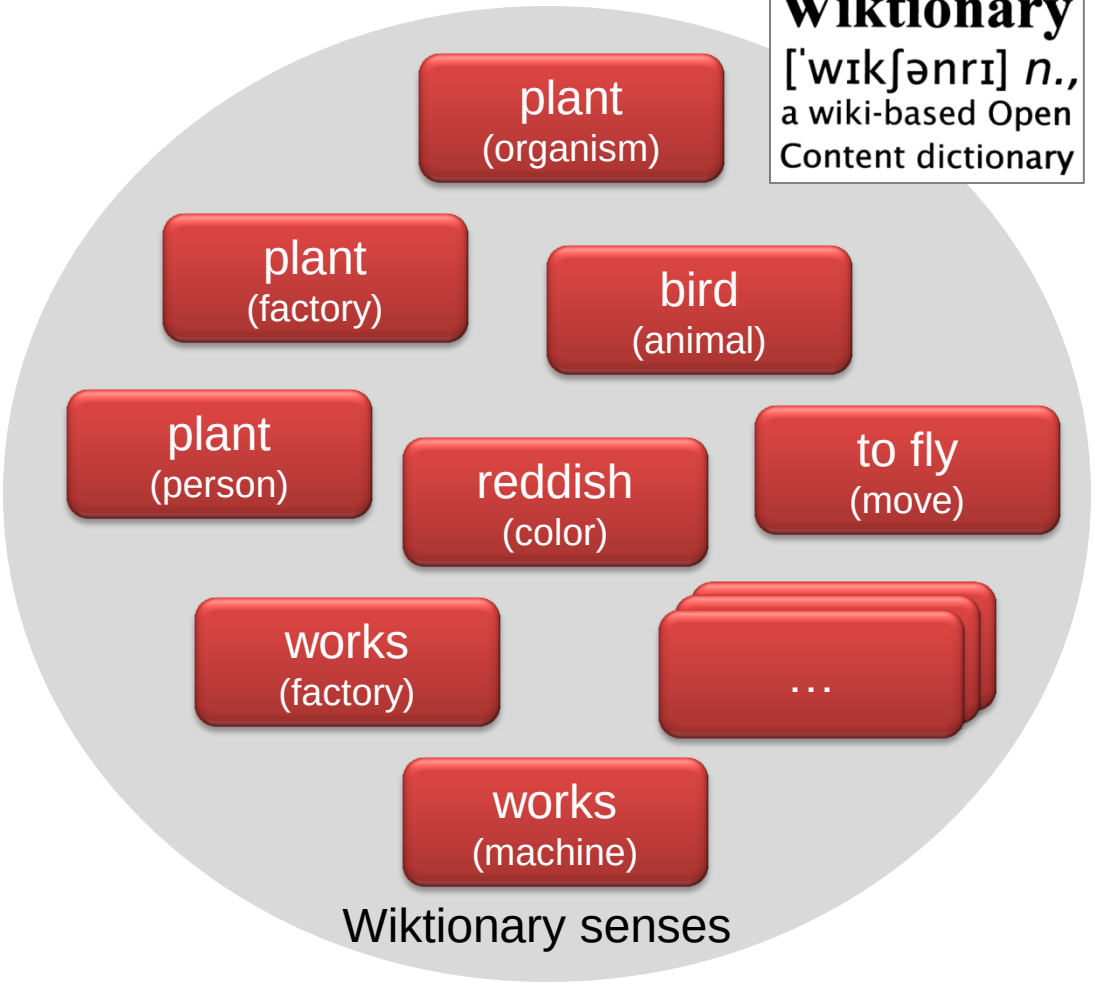
Beispiel Wiktionary–WordNet

TECHNISCHE UNIVERSITÄT DARMSTADT

Zwei-Schritt-Verfahren:

1. *Kandidatenextraktion*
2. *Disambiguierung*

WordNet synsets



Wiktionary senses

(Meyer&Gurevych, 2011)

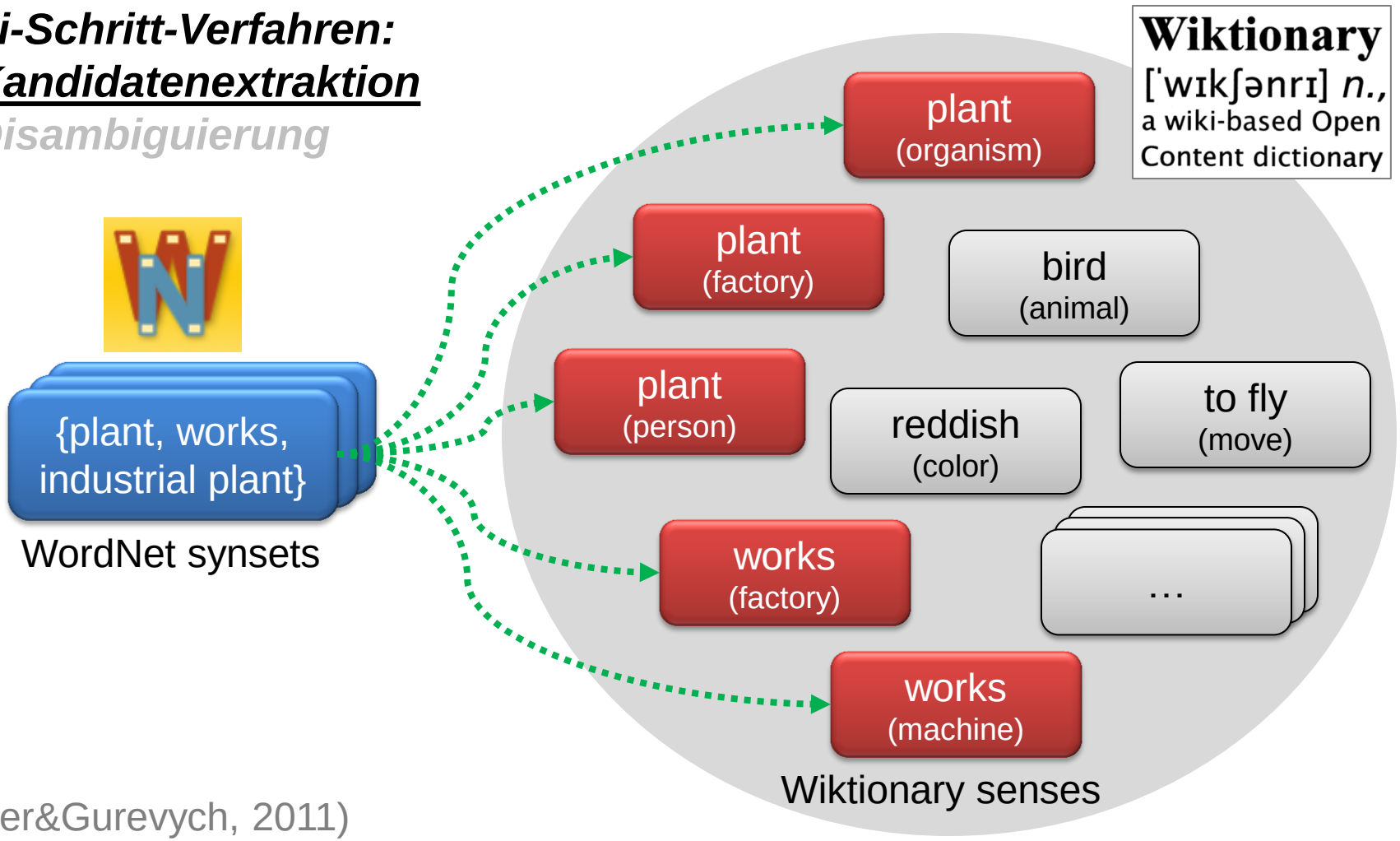
UKP

Beispiel Wiktionary–WordNet

TECHNISCHE UNIVERSITÄT DARMSTADT

Zwei-Schritt-Verfahren:

1. ***Kandidatenextraktion***
2. *Disambiguierung*



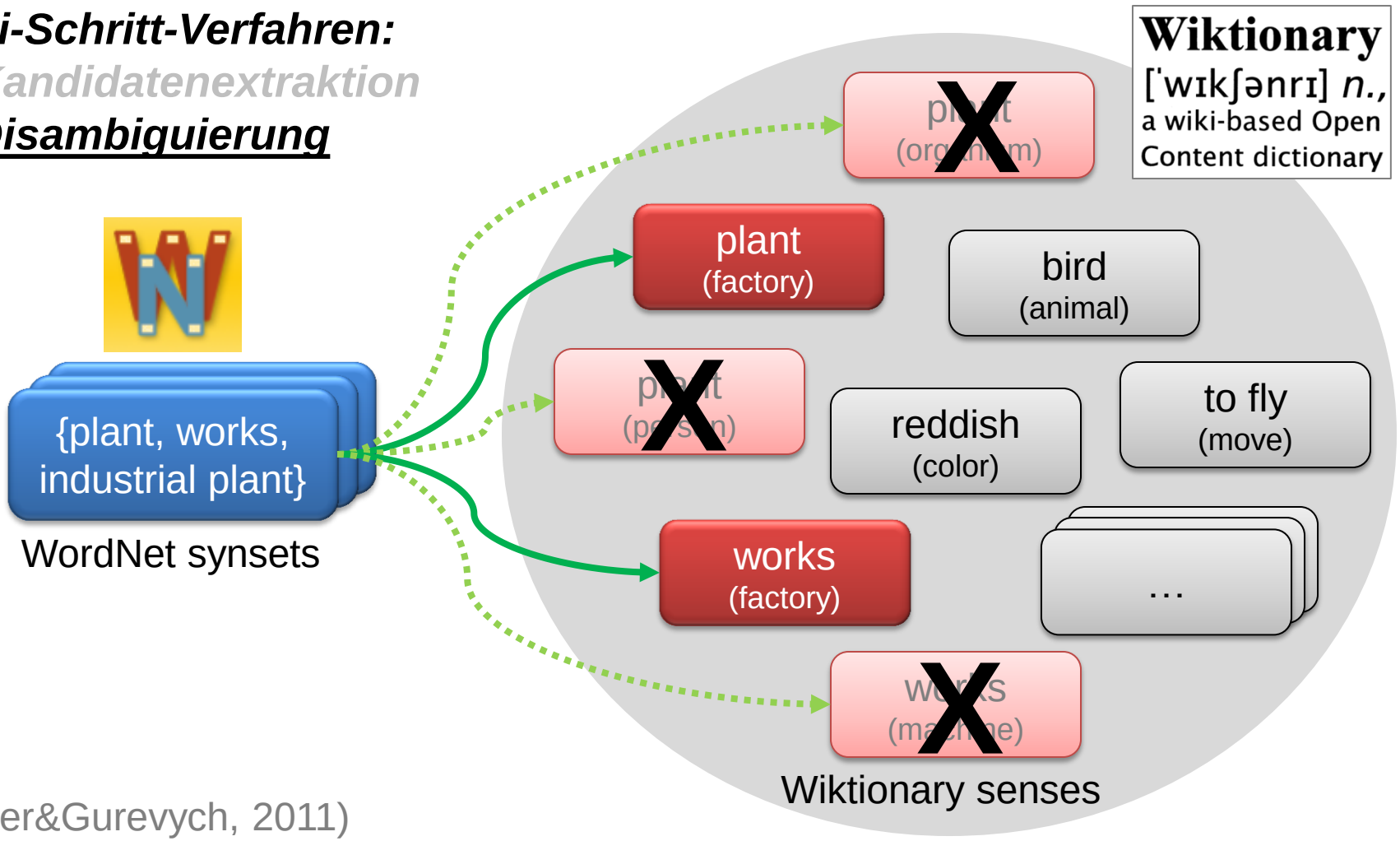
(Meyer&Gurevych, 2011)

Beispiel Wiktionary–WordNet

TECHNISCHE UNIVERSITÄT DARMSTADT

Zwei-Schritt-Verfahren:

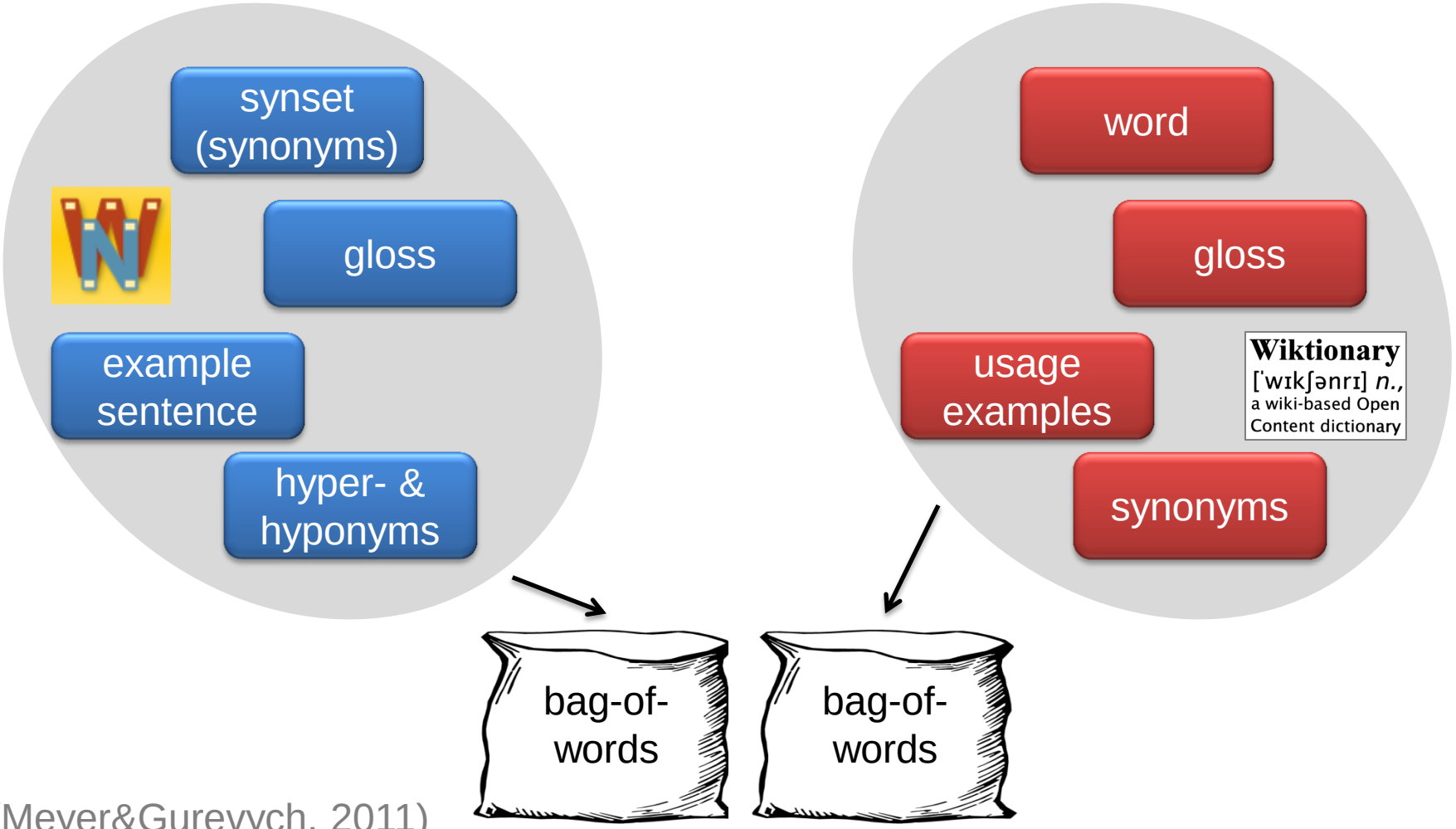
1. *Kandidatenextraktion*
2. ***Disambiguierung***



(Meyer&Gurevych, 2011)

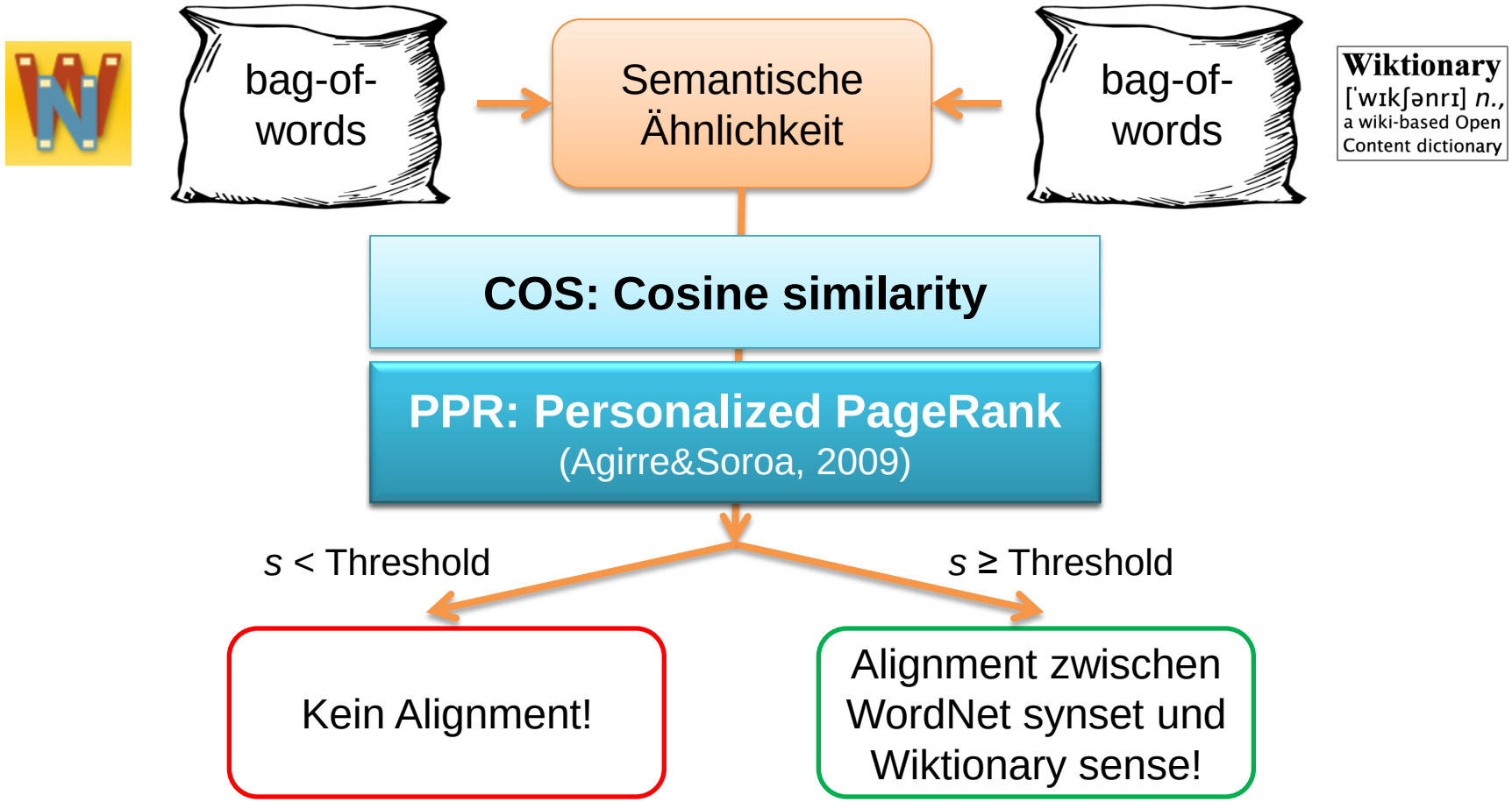
UKP

Disambiguierung: BoW-Repräsentation



(Meyer&Gurevych, 2011)

Disambiguierung: Bewertung



(Meyer&Gurevych, 2011)

Abdeckung: Wortarten

TECHNISCHE UNIVERSITÄT DARMSTADT

- 56.970 alignierte Bedeutungen
- Kombination der beiden Lexika: 488.988 Bedeutungen

	Wiktionary UND WordNet	Nur Wiktionary	Nur WordNet
Substantive	34.464	158.085	47.651
Verben	8.252	29.119	5.515
Adjektive/Adverbien	14.236	60.977	7.541
Andere Wortarten	0	16.778	0
Flektierte Formen	0	106.328	0

- Wiktionary ist nicht auf Substantive, Verben, Adjektive beschränkt: Phrasen, Sprichworte, Idiome,…

Abdeckung: Domänenwissen

TECHNISCHE UNIVERSITÄT DARMSTADT

	Wiktionary UND WordNet	Nur Wiktionary	Nur WordNet
Biology	4,465	4,067	12,869
Chemistry	2,561	8,260	2,268
Engineering	1,108	940	1,080
Geology	2,287	2,898	2,479
Humanities	4,949	2,700	5,060
IT	439	3,032	557
Linguistics	1,249	1,011	1,576
Math	615	2,747	483
Medicine	3,613	3,728	3,058
Military	574	426	585
Physics	1,246	2,835	1,252
Religion	733	1,154	781
Social Sciences	3,745	2,907	4,458
Sport	905	2,821	807

Bedeutungsrepräsentation

TECHNISCHE UNIVERSITÄT DARMSTADT

„Das beste aus zwei Welten“:

Synonyme

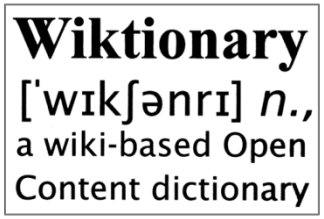
Bedeutungsparaphrase

Beispielsätze

Hyperonym-Taxonomie

Synset-Organisation

…



Aussprache

Etymologie

Syntaktische Angaben

Zitate

Verwandte Begriffe

Übersetzungen

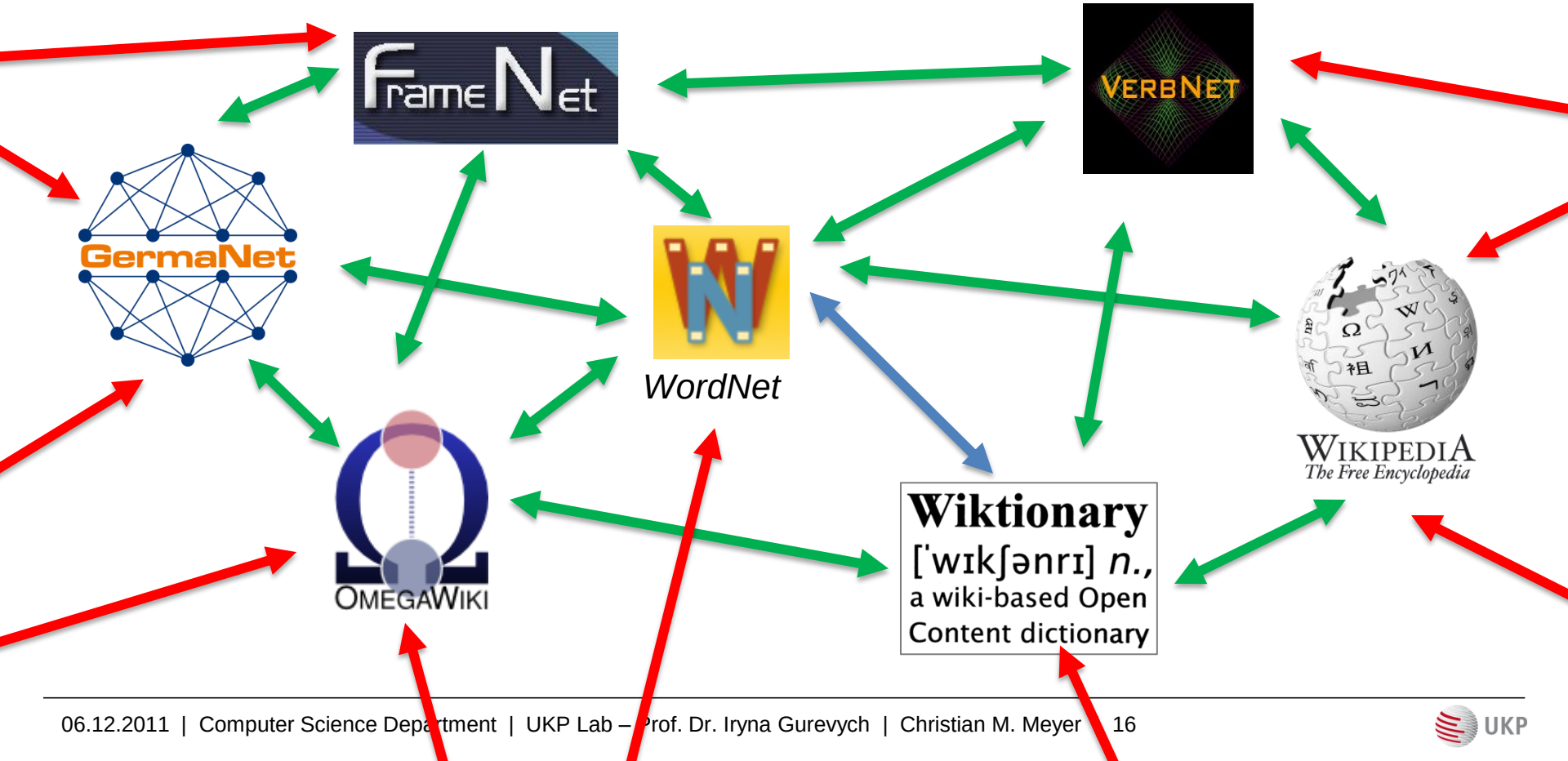
…

Ausblick Ressourcen-Alignierung

TECHNISCHE UNIVERSITÄT DARMSTADT

Die Alignierung von Wiktionary und WordNet ist ein Schritt…

…aber wir wollen mehr!



Teil II

ZUGRIFFSSTRUKTUREN

Standardisierung von Ressourcen mittels LMF

Warum Standardisierung?

TECHNISCHE UNIVERSITÄT DARMSTADT

Sprachressourcen sind nicht kompatibel:

Konzeptionelle Unterschiede

- Synset-Organisation (WordNet) vs. alphabetische Liste der headwords (Wörterbuch)

Inhaltliche Unterschiede

- Abdeckung, Granularität, Fokus, Informationstypen

Technische Unterschiede

- Datenbank (WordNet) vs. XML-Dump (Wiktionary)

Terminologische Unterschiede

- Lexical Unit (FrameNet) vs. word sense (Wiktionary)

Unterschiedlicher Zugriff

- Native Software-Bibliothek des Herstellers vs. Text Mining auf semi-strukturierten Daten

Lexical Markup Framework (LMF)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Definiert in ISO 24613:2008

- **Strukturelle Interoperabilität**: Terminologie, Lexikonstruktur in UML
  - Obligatorisch: LMF Core Package
  - Optional: LMF Extensions (z.B. Syntax, Semantics, Multilingual Extension)

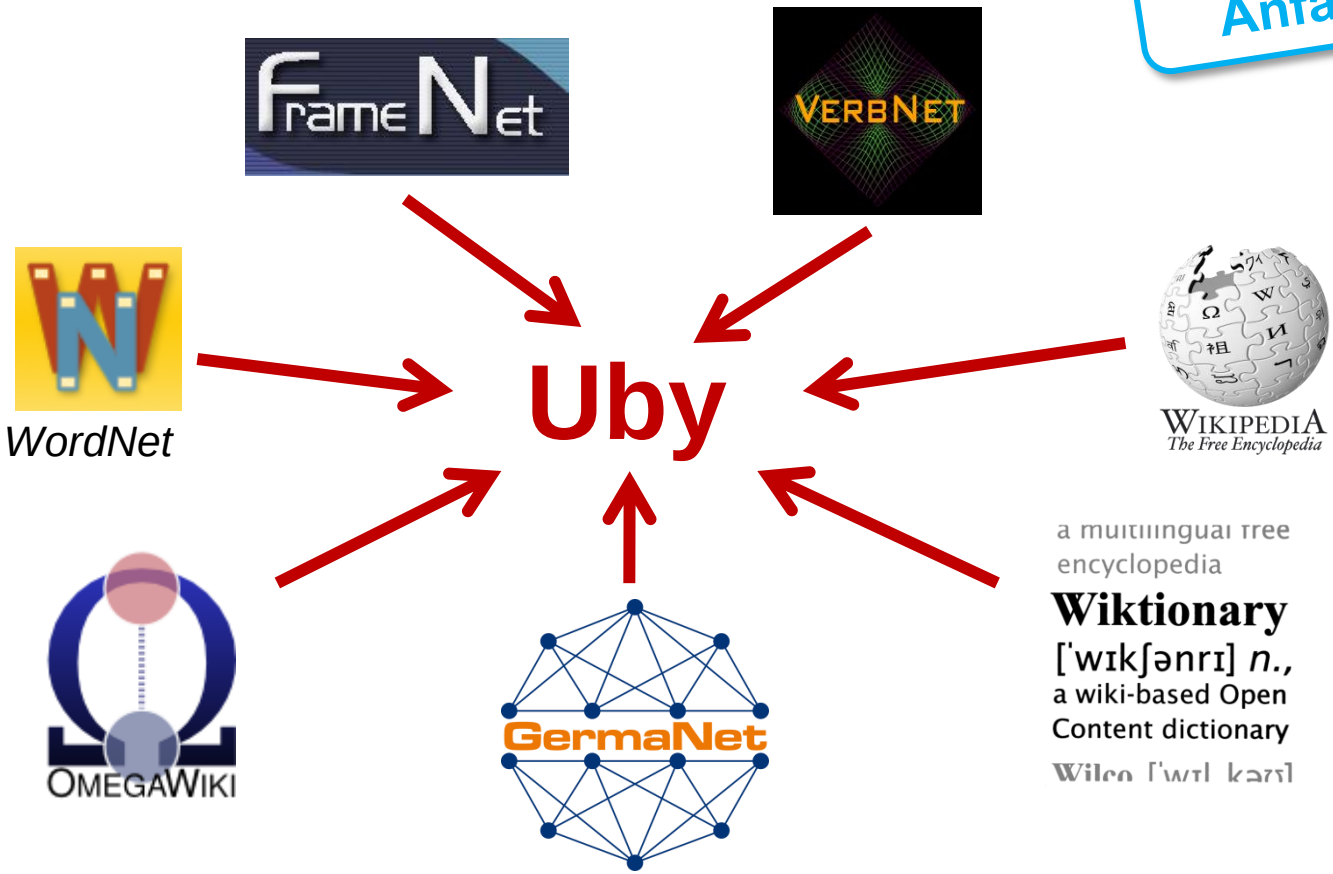
- **Semantische Interoperabilität**: Data Categories
  - Standardisiertes Vokabular für Attribute, z.B. extensionalDefinition (171), etymology (221), sampleSentence (455)
  - Verwaltet in ISOCat (Data Category Registry, ISO 12620)

Uby

TECHNISCHE UNIVERSITÄT DARMSTADT

Umfangreiches Integrationsprojekt:
einheitliche LMF-Repräsentation von Sprachressourcen

Veröffentlichung Anfang 2012



UKP

Uby

TECHNISCHE UNIVERSITÄT DARMSTADT

Umfangreiches Integrationsprojekt:
einheitliche LMF-Repräsentation von Sprachressourcen

Veröffentlichung Anfang 2012

Praktische Umsetzung des Standards mit vielen Ressourcen

Erstmals mit kollaborativen Ressourcen

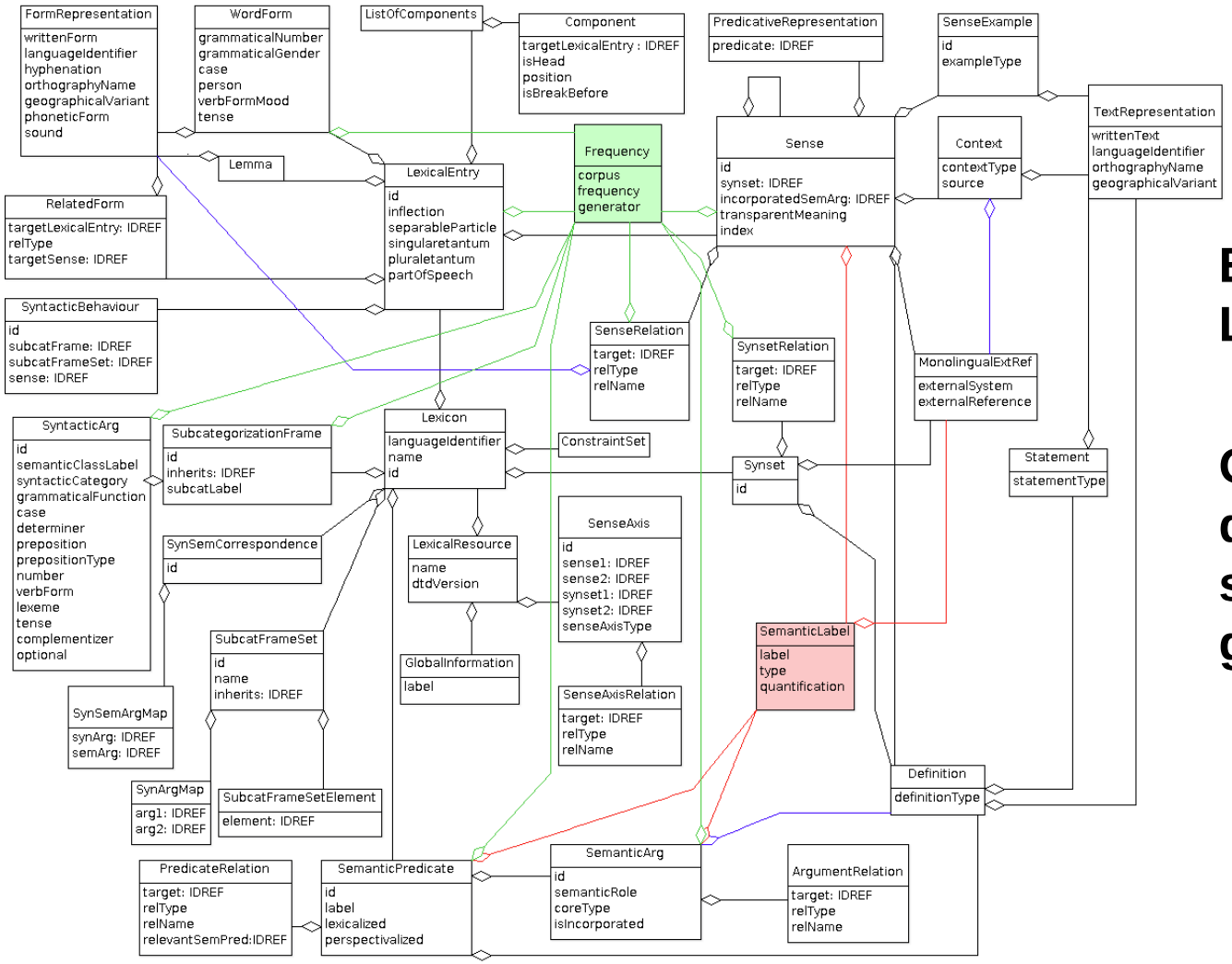
Multilingual (aktuell: Deutsch und Englisch)

Enge Ressourcen-Vernetzung durch Alignierung

UKP

Uby-LMF: Klassenstruktur

TECHNISCHE UNIVERSITÄT DARMSTADT



Erweiterung des LMF-Standards

Originäres Wissen der Ressourcen soll nicht verloren gehen!

UKP

Uby-LMF: Data Categories

TECHNISCHE UNIVERSITÄT DARMSTADT

Key	3723
PID	http://www.isocat.org/datcat/DC-3723
Type	complex/open
Owner	Nevskaya, Irina
Scope	public

1. Administration Information Section

1.1 Administration Record	
Identifier	lexeme_vernacular
Version	1:0
Registration Status	private
Administration Status	private
Justification	used in the MDF-set
Origin	Coward, David F. & Grimes, Charles E. (2000). Making Dictionaries: A guide to lexicography and the Multi-Dictionary Formatter. Waxhaw, North Carolina: SI /shoebox/MDF_2000.pdf http://www.sil.org/computing/shoebox/MDF_Updates.html
Explanatory Comment	in the vernacular language
Effective Date	2010-08-30
1.1.1 Creation	
Creation Date	2010-08-24
Change Description	creation of the category
1.1.2 Last Change	
Last Change Date	2010-12-07
Change Description	Added MDF marker as a valid data element name.

2. Description Section

Profile	Private
Profile	Lexicography
Profile	Language Codes
Profile	Lexical Resources
2.1 Data Element Name Section	
Data Element Name	lx
Source	MDF
2.2 Data Element Name Section	
Data Element Name	lexeme (vernacular)
Source	Coward, David F. & Grimes, Charles E. (2000). Making Dictionaries: A guide to lexicography and the Multi-Dictionary Formatter. Waxhaw, North Carolina: SI /shoebox/MDF_2000.pdf http://www.sil.org/computing/shoebox/MDF_Updates.html
[-] **2.3 English Language Section**	
Language	English (en)
2.3.1 Name Section	
Name	lexeme (vernacular)
Name Status	admitted name
2.3.2 Definition Section	
Definition	The Record marker for each record in a lexical entry. It contains the lexeme or headword (which is commonly mono-morphemic). Since such a lexeme form lc field to provide a more readable form for vernacular speakers.
Source	Coward, David F. & Grimes, Charles E. (2000). Making Dictionaries: A guide to lexicography and the Multi-Dictionary Formatter. Waxhaw, North Carolina: SI /shoebox/MDF_2000.pdf http://www.sil.org/computing/shoebox/MDF_Updates.html

3. Conceptual Domain

Data Type	string

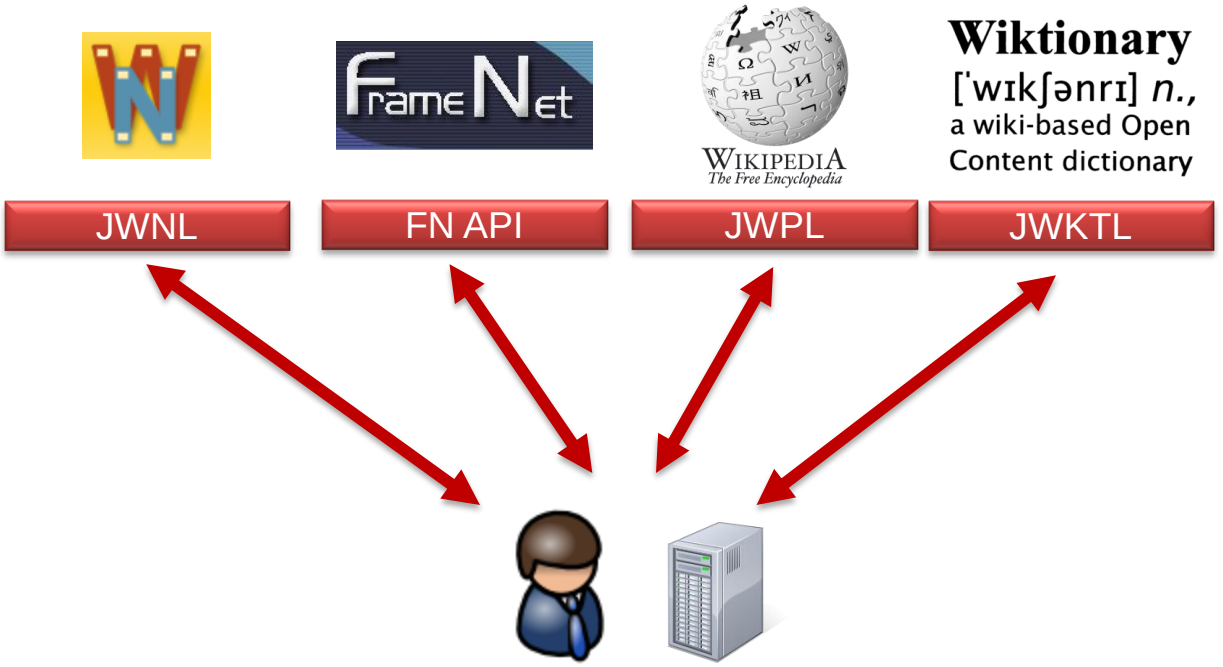
Auswahl von Data Categories aus ISOCat

Neudefinitionen wo nötig, z.B. bei semantischen Labels

Uby-API

- Früher: Abfrage von Informationen aus einer Ressource per manueller Extraktion, Web-Oberfläche, Software-Bibliothek (eigene od. von Dritten)

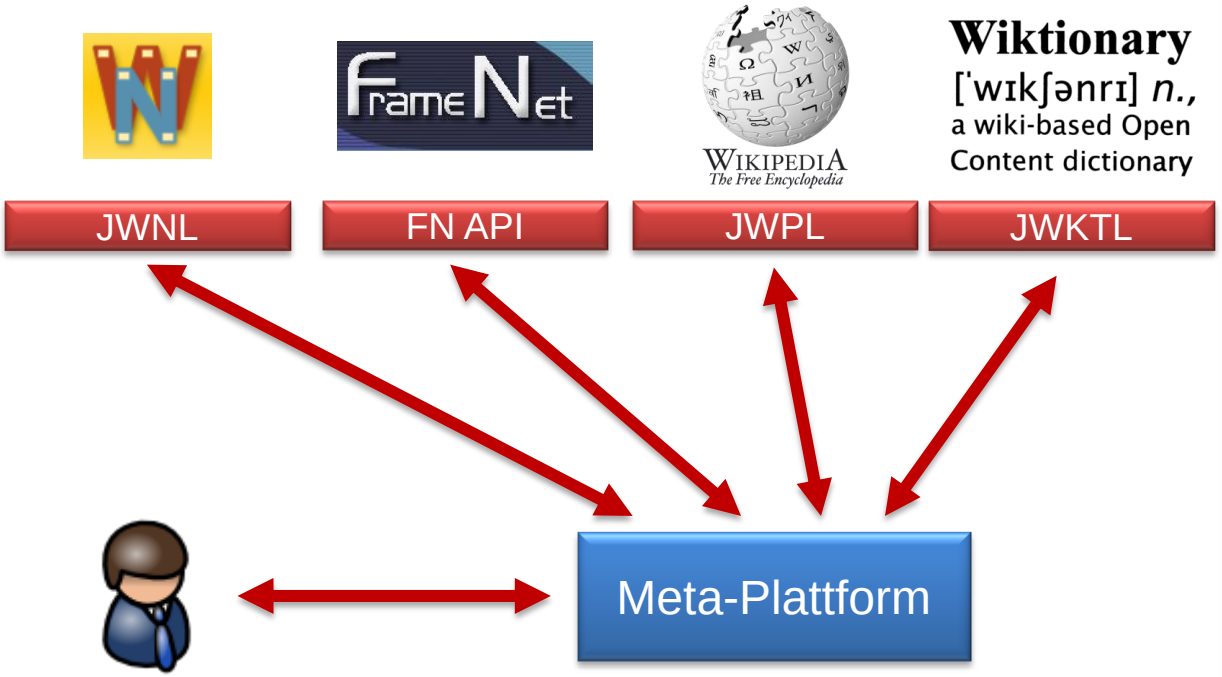
➡ Informationen müssen „mühevoll“ zusammengetragen werden



Uby-API

TECHNISCHE UNIVERSITÄT DARMSTADT

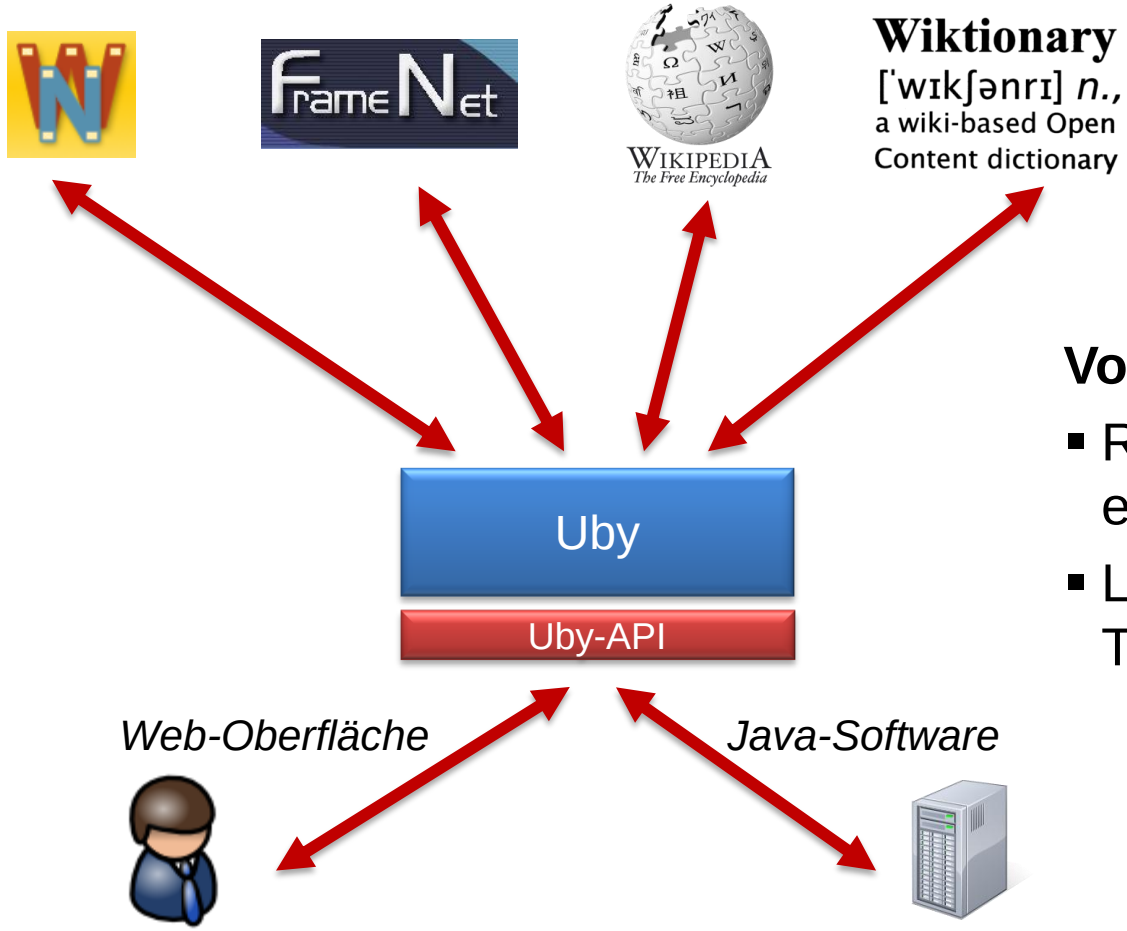
- Meta-Plattformen (z.B. DWDS) schaffen Abhilfe:
- Parallele Recherche mit einzelnen „Panels“



➡ **Aber: keine Alignierung/Kombination; kein programmat. Zugriff**

Uby-API

- Uby-API: Standardisierung erlaubt einheitlichen Zugriff auf Ressourcen:



Vorteile:

- Ressourcen leicht erweiter-/austauschbar
- LMF als einheitliche Terminologie

Fazit

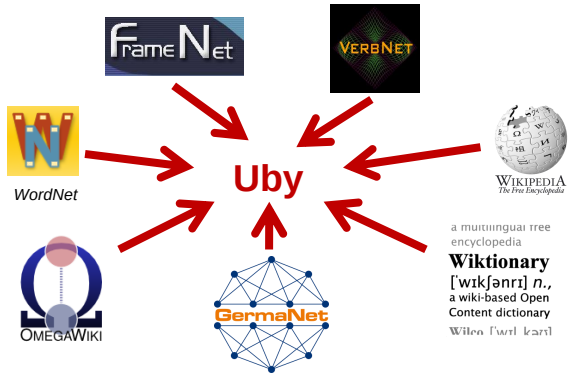
TECHNISCHE UNIVERSITÄT DARMSTADT

Vernetzungsstrategien:

- Automatische Alignierung von Wortbedeutungen
- Wiktionary–WordNet-Alignment als Beispiel
- Erste Einblicke in die Analyse alignierter Ressourcen

Zugriffsstrukturen:

- Uby: umfangreiches Integrationsprojekt für Sprachressourcen
- Lexical Markup Framework (LMF) als einheitliche Struktur
- Uby-API als einheitliche Zugriffsschnittstelle mit LMF-Terminologie



Frei verfügbare Ressourcen:
http://www.ukp.tu-darmstadt.de

Ausblick

TECHNISCHE UNIVERSITÄT DARMSTADT

Aktuelle Arbeit zum Thema „Collaborative Lexicogrpahy“:

Christian M. Meyer and Iryna Gurevych: **Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography**, in S. Granger & M. Paquot (Eds.): Electronic Lexicography, Oxford: Oxford University Press (to appear).

Weitere Referenzen:

Christian M. Meyer and Iryna Gurevych: **What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage**, in: Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), pp. 883-892, November 2011. Chiang Mai, Thailand.

Michael Matuschek and Iryna Gurevych: **Where the journey is headed: Collaboratively constructed multilingual Wiki-based resources**, in SFB 538: Mehrsprachigkeit (= Hamburger Arbeiten zur Mehrsprachigkeit), September 2011. Hamburg, Germany.

Elisabeth Niemann (geb. Wolf) and Iryna Gurevych: **The People’s Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet**, in: Proceedings of the 9th International Conference on Computational Semantics, p. 205-214, January 2011. Oxford, UK.

Christian M. Meyer and Iryna Gurevych: **How Web Communities Analyze Human Language: Word Senses in Wiktionary**, in: Proceedings of the Second Web Science Conference (WebSci), April 2010. Raleigh, NC, USA.

Ausblick

TECHNISCHE UNIVERSITÄT DARMSTADT

Aktuelle Arbeit zum Thema „Collaborative Lexicogrpahy“:

Christian M. Meyer and Iryna Gurevych: **Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography**, in S. Granger & M. Paquot (Eds.): Electronic Lexicography, Oxford: Oxford University Press (to appear).

Vielen Dank für die Aufmerksamkeit!

Christian M. Meyer and Iryna Gurevych: **What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage**, in: Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), pp. 883-892, November 2011. Chiang Mai, Thailand.

Michael Matuschek and Iryna Gurevych: **Where the journey is headed: Collaboratively constructed multilingual Wiki-based resources**, in SFB 538: Mehrsprachigkeit (= Hamburger Arbeiten zur Mehrsprachigkeit), September 2011. Hamburg, Germany.

Elisabeth Niemann (geb. Wolf) and Iryna Gurevych: **The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet**, in: Proceedings of the 9th International Conference on Computational Semantics, p. 205-214, January 2011. Oxford, UK.

Christian M. Meyer and Iryna Gurevych: **How Web Communities Analyze Human Language: Word Senses in Wiktionary**, in: Proceedings of the Second Web Science Conference (WebSci), April 2010. Raleigh, NC, USA.

Kontakt / Contact

Christian M. Meyer
Technische Universität Darmstadt
Ubiquitous Knowledge Processing Lab

Hochschulstr. 10, 64289 Darmstadt, Germany
+49 (0)6151 16–7477
+49 (0)6151 16–5455
meyer (at) ukp.informatik.tu-darmstadt.de