The Practitioner's Cookbook for Linked Lexical Resources

Iryna Gurevych & Judith Eckle-Kohler

Iryna Gurevych & Judith Eckle-Kohler

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

Outline

Part 1: Ingredients and Techniques

Part 2: Recipes

Part 1: Ingredients and Techniques

TECHNISCHE UNIVERSITÄT DARMSTADT

Ingredients: Lexical Resources

Integration Technique: Automatic Sense Linking

Interoperability Technique: Standardizing

Part 1: Ingredients and Techniques

TECHNISCHE UNIVERSITÄT DARMSTADT

Ingredients: Lexical Resources

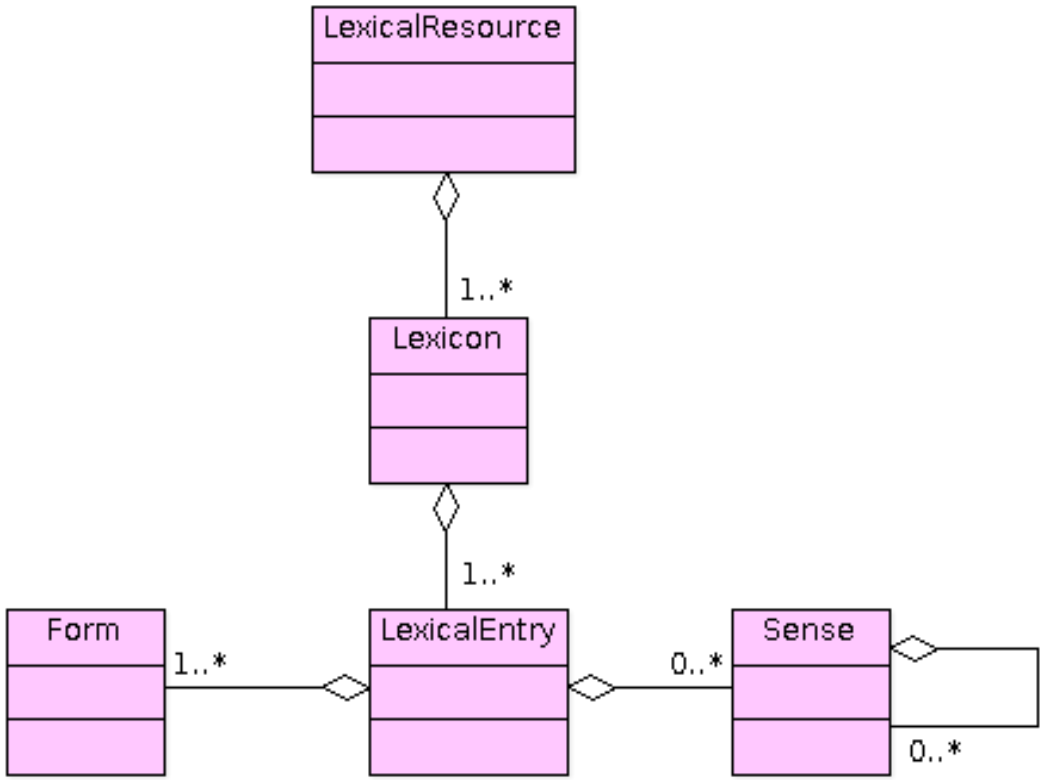
- **Elements of Lexical Resources**
- **Classic Lexical Resources**
- **Collaborative Lexical Resources**

Integration Technique: Automatic Sense Linking

Interoperability Technique: Standardizing

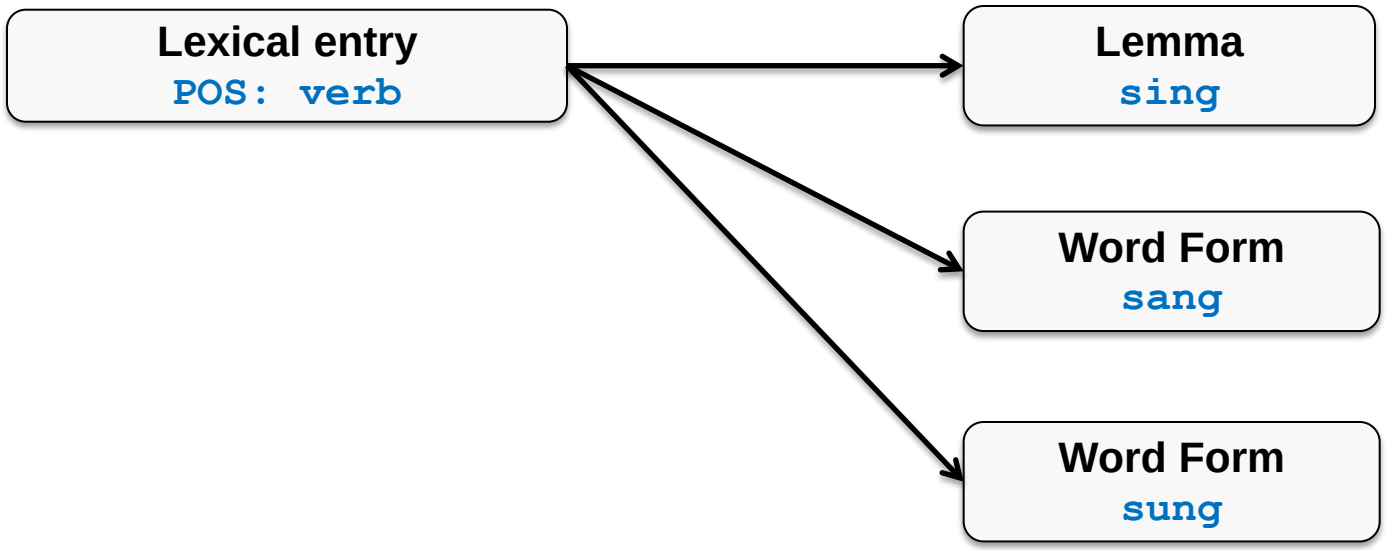
Elements of Lexical Resources

Lexical Markup Framework
(LMF, ISO 24613:2008) – Core Package



- A lexical resource consists of lexicons.
- A lexicon belongs to a particular language and consists of lexical entries.
- LexicalEntry is a class representing a lexeme in a given language.
- A lexeme is an abstract pairing of meaning and form (Jurafsky & Martin, 2008)

The Form Part of a Lexical Entry: Lemma and Word Form



The Meaning Part of a Lexical Entry: Sense

- Lemmas can have several senses (**lexical ambiguity**)
- Colloquial: „*words can have several meanings*“
- Many lemmas are associated with more than one sense

to sing

1. Produce tones with the voice

2. divulge confidential information or secrets

Lemma
`to sing`

→ **Sense #1**
`Sing / produce tones`

→ **Sense #2**
`Sing / divulge secrets`

Lexical Information about Senses I

- **Definitions**: short summaries of the meaning of a sense, also called **glosses**. They are meant to define a meaning.
  - The lemma *sing* has a sense that can be defined as “Produce tones with the voice“.
- **Sense examples**: Senses can also be described by example sentences. They illustrate a meaning.
  - Sense example for lemma *sing* in the sense “Produce tones with the voice“: "*She was singing while she was cooking*."

Lexical Information about Senses II

- **Translations**
  - German translation of *to sing* : *singen*

- **Usage information,** e.g. register (informal, slang, …)
  - *sing / divulge secrets* has register slang

- **Semantic field**
  - e.g., in WordNet: animate, food, location, communication …
  - semantic field of *sing / produce tones:* creation

- **Domain**:
  - *sing / produce tones* has domain music

Lexical Information about Senses III

TECHNISCHE UNIVERSITÄT DARMSTADT

Morphologically Related Senses

- E.g. *sing / produce tones with the voice* (verb) – *singer* (noun)

Sense Relations

- Synonymy: equivalent senses are related by the synonymy relation
  - Colloquial: „*the same meaning can be expressed by different words*“

- Hypernymy / Hyponymy (noun senses): Also called the IS-A relation
  - singer is-a musician

- Many more sense relations …

Lexical Information about Senses IV

- **Syntactic behavior**: Subcategorization frames (SCFs)
- SCFs specifiy
  - syntactic categories (NP-nominative, NP-accusative, PP …) and
  - grammatical functions (subject, object, ...) of the arguments of a verbal, nominal or adjectival predicate.
- partly language-specific

She *is singing.*
[subject, nominative]

She *is singing* *Christmas carols.*
[subject, nominative] **[object,accusative]**

Relation of Sense and Subcategorization Frame

Example:

to sing

1. Produce tones with the voice

2. divulge confidential information or secrets

- *Sing / produce tones with the voice* can be used with NP-accusative:
  - *They sing Christmas carols.*

- *Sing / divulge secrets* can not be used with NP-accusative:
  - *Mob Informant Joe Vollaro sings again.*

Grouping Related Senses – Different Ways to Organize a Lexicon

In dictionaries and in the ISO standard LMF, senses are grouped into lexical entries that share the same lemma and part-of-speech.

However, senses can also be **grouped differently**:

- Grouping of senses that are related by some sense relation
  - e.g., synonymy
- Grouping of senses that share the same syntactic behavior
  - e.g., subcategorization frame

Part 1: Ingredients and Techniques

TECHNISCHE UNIVERSITÄT DARMSTADT

Ingredients: Lexical Resources

- **Elements of Lexical Resources**
- **Classic Lexical Resources**
- **Collaborative Lexical Resources**

Integration Technique: Automatic Sense Linking

Interoperability Technique: Standardizing

WordNet

- Domain-independent, broad-coverage lexical-semantic network of English nouns, verbs and adjectives
- Realized at Princeton University by George Miller's team (started in 1985)
- Widely used for many NLP tasks and applications
- http://wordnet.princeton.edu/wordnet/

WordNet is Organized in Synsets

Synonymous senses are grouped into synsets.

spill the beans#1
let the cat out of the bag#1
talk#5
tattle#2
blab#1
peach#1
babble#4
sing#5
babble out#1
blab out#1

Synset ,
Synset ID: (1){00939238}

WordNet Synsets

Synonymous senses are grouped into synsets.

spill the beans#1
let the cat out of the bag#1
talk#5
tattle#2
blab#1
peach#1
babble#4
sing#5
babble out#1
blab out#1

Synset ,
Synset ID: (1){00939238}

Members of a synset

- are senses,
- are represented by their lemma.

All synset members belong to the same word class (e.g., noun, verb).

Synset Definitions (Glosses)

Synsets have glosses, i.e. definitions/short summaries of their meaning.

- The meaning of a synset is expressed in its gloss.
- The meaning of a synset can alternatively be captured by the list of its member synonyms.

Synset , Synset ID: (1){00939238}

Gloss: "divulge confidential information or secrets**"**

Members:
sing (Sense ID: sing%2:32:01::)
spill the beans (Sense ID: spill_the_beans%2:32:00::)

Sense Examples

Senses, i.e., members of a synset, are illustrated by sense examples – sentences illustrate a meaning.

Synset , Synset ID: (1){00939238}
Gloss: “divulge confidential information or secrets“

Members:
sing (Sense ID: sing%2:32:01::)
Sense Example: *"Mob Informant Joe Vollaro sings again"*

spill the beans (Sense ID: spill_the_beans%2:32:00::)
Sense Example: *"They had planned it as a surprise party, but somebody spilled the beans."*

VerbNet

- Domain-independent, broad-coverage lexicon of English verbs
  - lexical-syntactic and lexical-semantic information for verbs
  - 3962 verb lemmas in VerbNet 3.1

- http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

VerbNet is Organized in Verb Classes

Verbs are grouped into **verb classes**, based on Levin‘s classification of English verbs:

- 470 verb classes, hierarchically structured
- **VerbNet sense** = pair of verb lemma and verb class

chant chatter chirp chortle
chuckle cluck coo croak … scream screech shout shriek sibilate
sigh simper
sing smatter smile snap
snarl snivel
snuffle …

manner_speaking-37.3
Members: 103, Frames: 14

VerbNet Verb Classes

VerbNet verb classes group verbs that share the same predicate-argument structure, i.e.

- subcategorization frame,
- semantic roles and selectional preferences,
- semantic predicate based on the event decomposition of Moens and Steedman (1988).

Although the resulting verb classes are semantically coherent, the semantic relatedness of verb senses in a VerbNet class is **distant** compared to WordNet synsets.

- e.g., the verbs *believe, swear and doubt* are in the same verb class.

VerbNet – Information types

- subcategorization frame, semantic roles and selectional preferences
- semantic predicate
- single-verb sense example in the lexical resource VerbNet
- SemLink provides links to real sense examples in PropBank

Example:

```
VERB: sing
EXAMPLE: "Susan whispered about the party."
SYNTAX:
Agent[+animate|+organization] V {about} Topic[+communication]
SEMANTIC PRED:
transfer_info(during(E), Agent, ?Recipient, Topic)
cause(Agent, E)
```

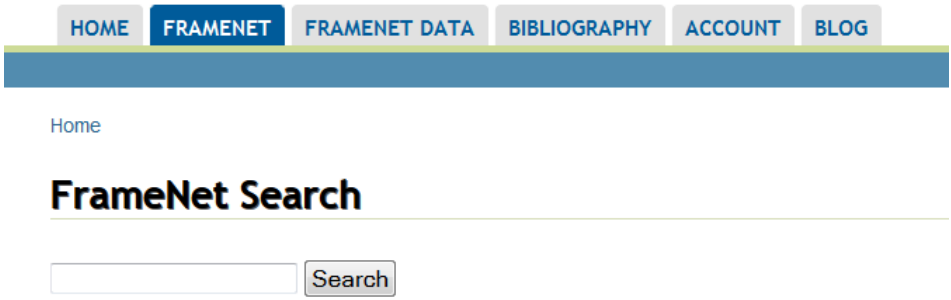
FrameNet

FrameNet is a lexical resource of English verbs, nouns, adjectives.

FrameNet 1.5 contains:

- 1,019 frames,
- 9,423 lemmas,
- 11,942 lexical units

http://framenet.icsi.berkeley.edu/



FrameNet is Organized in Frames

- Senses are grouped into **frames** based on Fillmore's Frame Semantics (Fillmore 1976, Fillmore et al. 2003).
- In FrameNet, senses are called lexical units.

babble.v, bluster.v, chant.v, chatter.v, drawl.v, gabble.v, gibber.v, jabber.v, lisp.v, mouth.v, mumble.v, mutter.v, natter.v, prattle.v, rant.v, rave.v, shout.v, simper.v, **sing.v**, slur.v, stammer.v, stutter.v, whisper.v

Communication_manner

FrameNet Frames

- A **frame** represents a conceptual structure, or a prototypical situation with a (frame-specific) set of **roles** that identify the participants involved in the situation.

- Frames group senses which **evoke** the same kind of situation with participants taking over particular roles.

- Senses in a FrameNet frame are semantically related, but not synonymous; e.g., the verbs *love* and *hate* are both in the same FrameNet frame.

Frame Evoking Word Classes

Frame evoking word classes are

- **verbs**, they are the prototypical frame-evoking word classes
- **predicate-like** nouns and adjectives
  - e.g. nouns denoting events (*development*), relations (*brother*), states (*height*)

Example:

COMMUNICATION_MANNER frame:

[They]$_{\text{Speaker}}$ all *sang* [Happy Birthday]$_{\text{Message}}$

Coverage Issues in FrameNet

Low lexical coverage compared to WordNet

- 9,423 lemmas in FrameNet 1.5
- 156,584 lemmas in WordNet 3.0

- The focus of FrameNet is on verbs and predicate-like nouns and adjectives.
  - Many nouns and adjectives evoke uninteresting frames, e.g., nouns denoting artifacts and natural kinds; adjectives denoting colors.
  - Hence, few of them have been included [Baker and Fellbaum 2009].

- Senses encountered in a corpus may have no corresponding sense in FrameNet, e.g.

  Sing / produce tones with the voice → Frame Communication_manner

  Sing / divulge secrets → No frame available

Part 1: Ingredients and Techniques

TECHNISCHE UNIVERSITÄT DARMSTADT

Ingredients: Lexical Resources

- **Elements of Lexical Resources**
- **Classic Lexical Resources**
- **Collaborative Lexical Resources**

Integration Technique: Automatic Sense Linking

Interoperability Technique: Standardizing

Wikipedia, a Multilingual Encyclopedia

- Wikipedia is a freely licensed encyclopedia written by thousands of volunteers in many languages.
- Free license allows others to freely copy, redistribute, and modify the work commercially or non-commercially.
- Founded January 15, 2001
- 6th most visited site according to Alexa (Feb. 2010)

http://www.wikipedia.org

WIKIPEDIA
The Free Encyclopedia

(Jimmy Wales)

Wikipedia – Information Types

Title

Redirects

Introduction (first paragraph)

Infoboxes

Headers

Hyperlinks

Categories

Benzoic acid

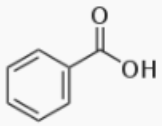
From Wikipedia, the free encyclopedia

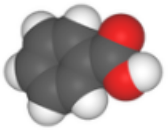
(Redirected from Benzenecarboxylic acid)

Benzoic acid, $C_7H_6O_2$ (or C_6H_5COOH), is a colorless crystalline solid and the simplest aromatic carboxylic acid. The name derived from gum benzoin, which was for a long time the only source for benzoic acid. This weak acid and its salts are used as a food preservative. Benzoic acid is an important precursor for the synthesis of many other organic substances.

Contents [hide]

Benzoic acid





IUPAC name	[show]
Other names	Benzenecarboxylic acid, Carboxybenzene, E210, Dracylic acid
Identifiers	
CAS number	65-85-0 ✓
PubChem	243
EC number	200-618-2
KEGG	C00180
MeSH	benzoic+acid
ChEBI	30746
RTECS number	DG0875000
SMILES	[show]

Carboxyl group

[edit]

All the reactions mentioned for carboxylic acids are also possible for benzoic acid.

- Benzoic acid esters are the product of the acid-catalysed reaction with alcohols.
- Benzoic acid amides are more easily available by using activated acid derivatives (such as benzoyl chloride) or by coupling reagents used in peptide synthesis like DCC and DMAP.

Categories: Organic acids | Benzoic acids | Aromatic compounds | Excipients

Wikipedia – Disambiguation pages

- Sense inventory, including domain specific senses

Forest (disambiguation)

From Wikipedia, the free encyclopedia

A forest is a large area covered by trees.

Forest can also mean:

- Royal forest, an area set aside for hunting

Forest may also be:

- In Windows networking, the collection of every object, their attributes and rules in an Activ
- In graph theory, a disjoint union of trees
- *Forest* (album), an album by George Winston
- "Forest" (song), a song by the band System of a Down

The Forest may refer to:

- *The Forest*, a video game
- *The Forest*, a 2002 film

JWPL
Java-based Wikipedia Programming Library

TECHNISCHE UNIVERSITÄT DARMSTADT

Highlights

- High performance access to Wikipedia content
- Parser for the WikiMedia syntax
- Articles, discussion pages, categories as Java objects
- Access to information nuggets
- Redirects, links, link anchors, interlanguage links, sections, first paragraph, etc.
- Supports all Wikipedia language editions

Available open source: http://code.google.com/p/jwpl/

Wiktionary, a Multilingual Machine Readable Dictionary

TECHNISCHE UNIVERSITÄT DARMSTADT

Français
Le dictionnaire libre
856 000 + articles

English
The free dictionary
841 000+ articles

Tiếng Việt
Từ điển mở
227 000+ mục từ

Türkçe
Özgür sözlük
208 000+ madde

a multilingual free encyclopedia

Wiktionary

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

Wiko [ˈwɪ] kəɹɪ]

Русский
Свободный словарь
137 000+ статей

Ido
La libera vortaro
137 000+ artikli

中文
自由的多语言词典
116 000+ 条词条

Ελληνικά
Το Ελεύθερο Λεξικό
107 000+ λήμματα

0B A4 0B AE 0B BF 0B B4 0B CD
0B 95 0B 9F 0B CD 0B 9F 0B B1 0B CD 0B B1 0B 85 0B 95 0B B0 0B AE 0B C1 0B A4 0B B2 0B BF
102 000+ 0B 95 0B 9F 0B C0 0B 9F 0B C1 0B B0 0B C8 0B 95 0B B3 0B C0

Polski
Wolny słownik
93 000+ stron

http://www.wiktionary.org/

UKP

Wiktionary – Information Types

acid

See also **ACID** *and* **àcid**

Wikipedia has an article on: **Acid**

Contents [show]

English [edit]

Etymology [edit]

From French *acide*, from Latin *acidus* ("sour, acid"), from *aceō* ("I am sour").

Pronunciation [edit]

- IPA: /ˈæs.ɪd/, SAMPA: /"{s.Id/
  - Audio (US) help, file

Adjective [edit]

acid (*not comparable*)

1. Sour, sharp, or biting to the taste; tart; having the taste of vinegar.
   acid *fruits or liquors*
2. (*figuratively*) Sour-tempered.
3. Of or pertaining to an acid; acidic.
4. (*music*) Denoting a musical genre that is a distortion (as if hallucinogenic) of an existing genre, as in acid house, acid jazz, acid rock.

Quotations [edit]

- *For examples of the usage of this term see the citations page.*

Synonyms [edit]

- acidic

Antonyms [edit]

- alkaline
- base

- Language
- Etymology
- Pronunciation
- Part of speech
- Derived terms, related terms
- Abbreviations
- Collocations
- Word senses
- Glosses
- Examples
- Synonyms, antonyms, hypernyms, hyponyms
- Translations
- Morphology
- Quotations
- …

JWKTL
Java-based Wiktionary Library

TECHNISCHE UNIVERSITÄT DARMSTADT

Highlights:

- efficient and structured access to the information encoded in the English, the German, and the Russian Wiktionary language editions
- sense definitions
- part of speech tags
- etymology
- example sentences
- translations
- semantic relations
- … and many other lexical information types.

Available open source: http://code.google.com/p/jwktl/

OmegaWiki, a Multilingual Lexical-Semantic Resource

- A free, **multilingual** resource
- Over 420.000 expressions in 255 languages
- Over 40.000 **language-independent** concepts
- Around 3,000 users
- **Goals**
  - Overcome Wiktionary's structural inconsistencies
  - Create a resource for translations/synonyms which is easily accessible and maintainable
- **Consequence:** a fixed database schema
  - Users can only contribute if they stick to the predefined structure
    - …but the price is a loss in expressiveness

http://www.omegawiki.org/

JOWKL
Java-based OmegaWiki Library

TECHNISCHE UNIVERSITÄT DARMSTADT

Highlights:

- Fast and efficient access to OmegaWiki
- Direct access to OmegaWiki database dumps, no preprocessing necessary
- Language independent

Available open source: http://code.google.com/p/jowkl/

Part 1: Ingredients and Techniques

TECHNISCHE UNIVERSITÄT DARMSTADT

Ingredients: Lexical Resources

- **Elements of Lexical Resources**
- **Classic Lexical Resources**
- **Collaborative Lexical Resources**

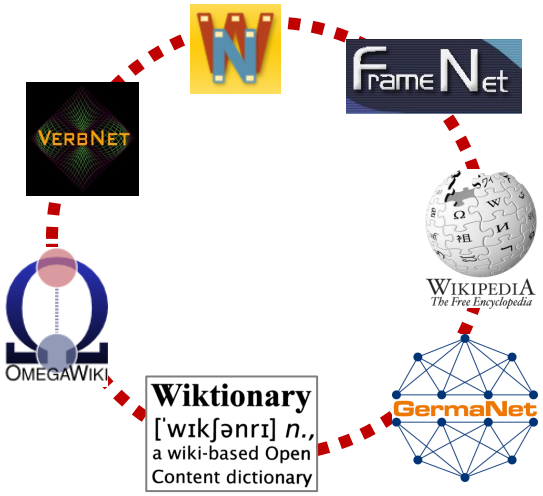
Integration Technique: Automatic Sense Linking

Interoperability Technique: Standardizing

Lexical Resource Integration – Motivation

Lexical resources are largely different

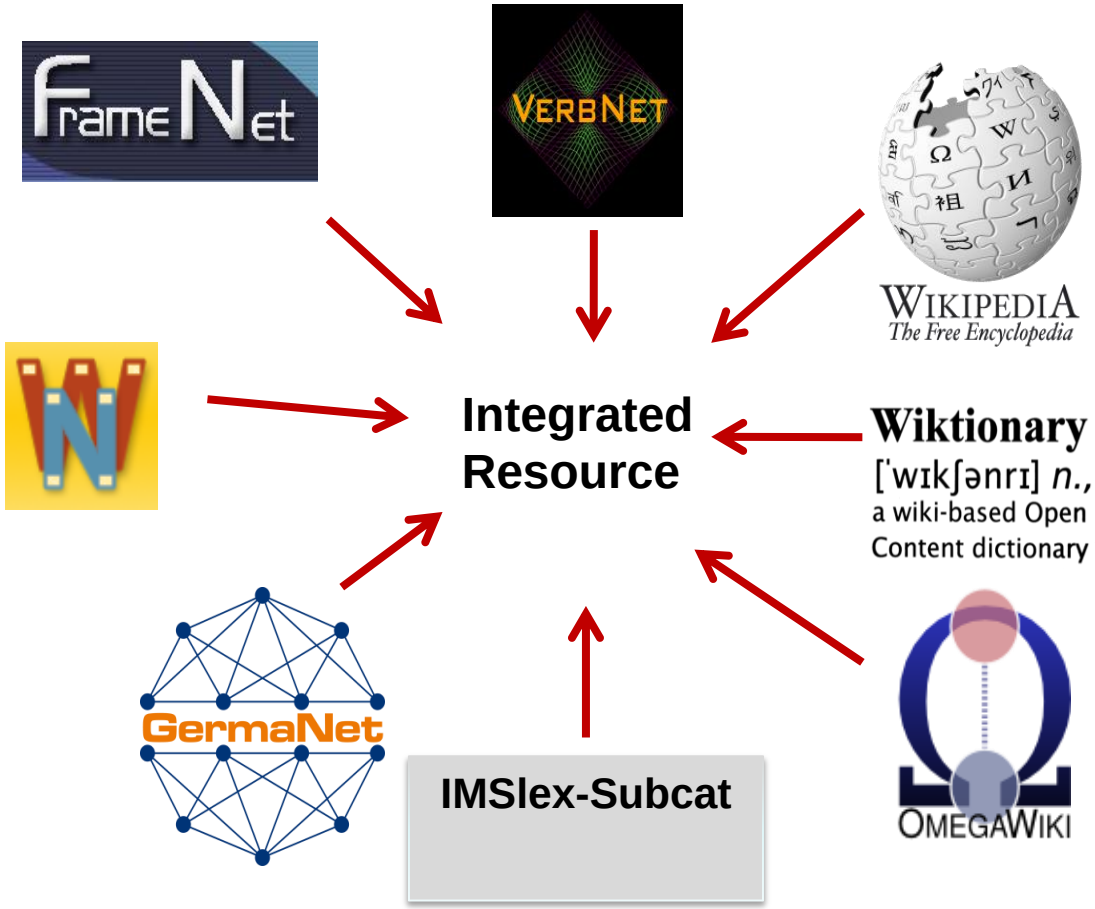
- Different coverage of words/word senses
- Different information types
  - Encyclopedic vs. linguistic knowledge
  - Syntactic vs. semantic knowledge
  - …



This can significantly influence the performance of an NLP system – Instead of choosing only one (best performing):

Why not combine multiple resources and benefit from all their knowledge?

Linking Lexical Resources at the Word Sense Level



Linking at the word sense level: **sense alignment of resources**
= linking of equivalent senses

Previous Work on Linking Lexical Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

Linked Lexical Resources:

- Meaning Multilingual Central Repository, Atserias et al. (2004)
- Yago, Suchanek et al. (2007)
- SemLink (Palmer, 2009)
- Universal Wordnet (UWN), Gerard de Melo and Gerhard Weikum (2009)
- eXtended WordFrameNet, Laparra and Rigau (2010)
- BabelNet, Navigli and Ponzetto (2010)
- NULEX, McFate and Forbus (2011)
- UBY, Gurevych et al. (2012)
- … many more, e.g., on the Semantic Web

Linking at the Word Sense Level: Example

TECHNISCHE UNIVERSITÄT DARMSTADT

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

1. To produce musical or harmonious sounds with one's voice.
2. To express audibly by means of a harmonious vocalization.
3. To confess under interrogation.

to sing

1. Produce tones with the voice

2. divulge confidential information or secrets

OmegaWiki — **to sing**

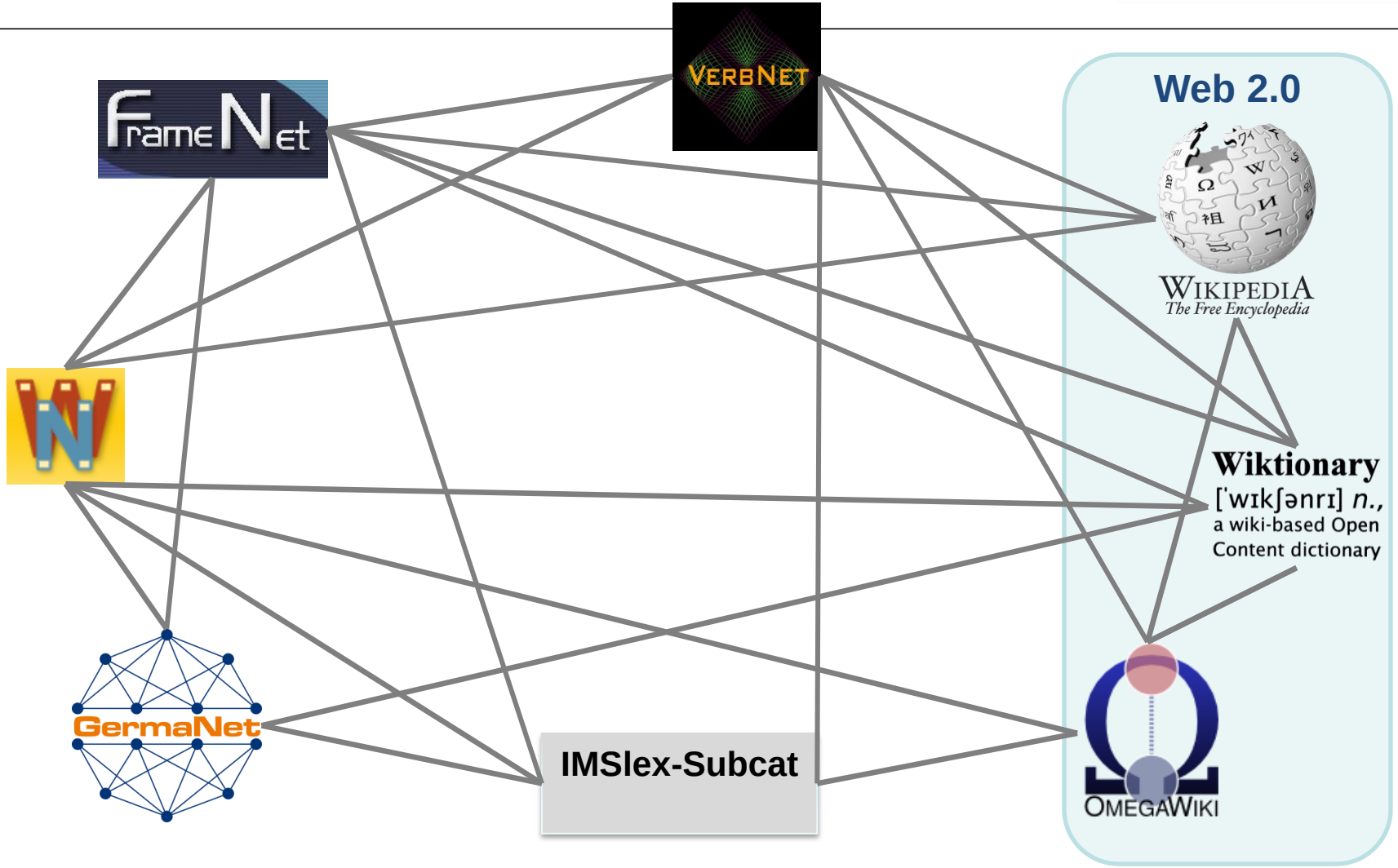
1. To produce harmonious sounds with one's voice.

OmegaWiki — **singen**

1. Mit der Stimme harmonische Töne erzeugen.

UKP

Sense Alignment of Multiple Resources



Case Study: Automatic Word Sense Alignment of Wiktionary and WordNet

Aims:

Create a word sense alignment between Wiktionary and WordNet that comes with

(1) Increased coverage

(2) Enriched sense representations

Meyer, Gurevych (2011)

Wiktionary vs. WordNet: Differently Developed

TECHNISCHE UNIVERSITÄT DARMSTADT

Wiktionary: Online lexicon that is collaboratively constructed by a community of Web users

a multilingual free encyclopedia
Wiktionary
['wɪkʃənrɪ] *n.*, a wiki-based Open Content dictionary

Entry | Discussion | Citations | Read | Edit | History | Search

Main Page
Community portal
Preferences
Requested entries
Recent changes
Random entry
(by language)
Help
Donations
Contact us
Toolbox
What links here
Related changes
Upload file

plant

English

Pronunciation

- (*RP*) IPA: /plɑːnt/, /plænt/
- (*US*) IPA: /plænt/
- Audio (US) (file)

Etymology 1

From Latin *planta*, later influenced by French *plante*.

Noun

plant (*plural* **plants** *or* **plantæ**)

1. An organism that is not an animal, especially an organism capable of photosynthesis. Typically a small or herbaceous organism of this kind, rather than a tree.

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: plant | Search WordNet

Display Options: (Select option to change) | Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relat

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **plant**, works, industrial plant (buildings for carrying on industrial labor) *"they built a large plant to manufacture automobiles"*
- S: (n) **plant**, flora, plant life ((botany) a living organism lacking the power of locomotion)
- S: (n) **plant** (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- S: (n) **plant** (something planted secretly for discovery by another) *"the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"*

WordNet: Semantic network created by psycholinguists at Princeton University (Fellbaum, 1998)

Wiktionary vs. WordNet: Different Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

a multilingual free encyclopedia

Wiktionary

[ˈwɪkʃənrɪ] *n.*,

a wiki-based Open Content dictionary

Wikto [ˈwɪk kəɪs]

WordNet

Noun

- S: (n) **plant**, works, industrial plant (buildings for carrying on industrial labor) *"they built a large plant to manufacture automobiles"*
- S: (n) **plant**, flora, plant life ((botany) a living organism lacking the power of locomotion)
- S: (n) **plant** (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- S: (n) **plant** (something planted secretly for discovery by another) *"the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"*

Noun

plant (*plural* **plants**)

1. An organism that is not an animal, especially an organi
   The garden had a couple of trees, and a cluster of
2. (*botany*) An organism of the kingdom *Plantae*; now spe
   b, or any organism closely related to such an organism.
3. (*ecology*) Now specifically, a multicellular eukaryote that includes chloroplasts in its cells, which have a cell wall.
4. A factory or other industrial or institutional building or facility.
5. An object placed surreptitiously in order to cause suspicion to fall upon a person.
   *That gun's not mine! It's a **plant**! I've never seen it before!*
6. Anyone assigned to behave as a member of the public during a covert operation (as in a police investigation).
7. A person, placed amongst an audience, whose role is to cause confusion, laughter etc.
8. (*snooker*) A play in which the cue ball knocks one (usually red) ball onto another, in order to pot the second; a set.

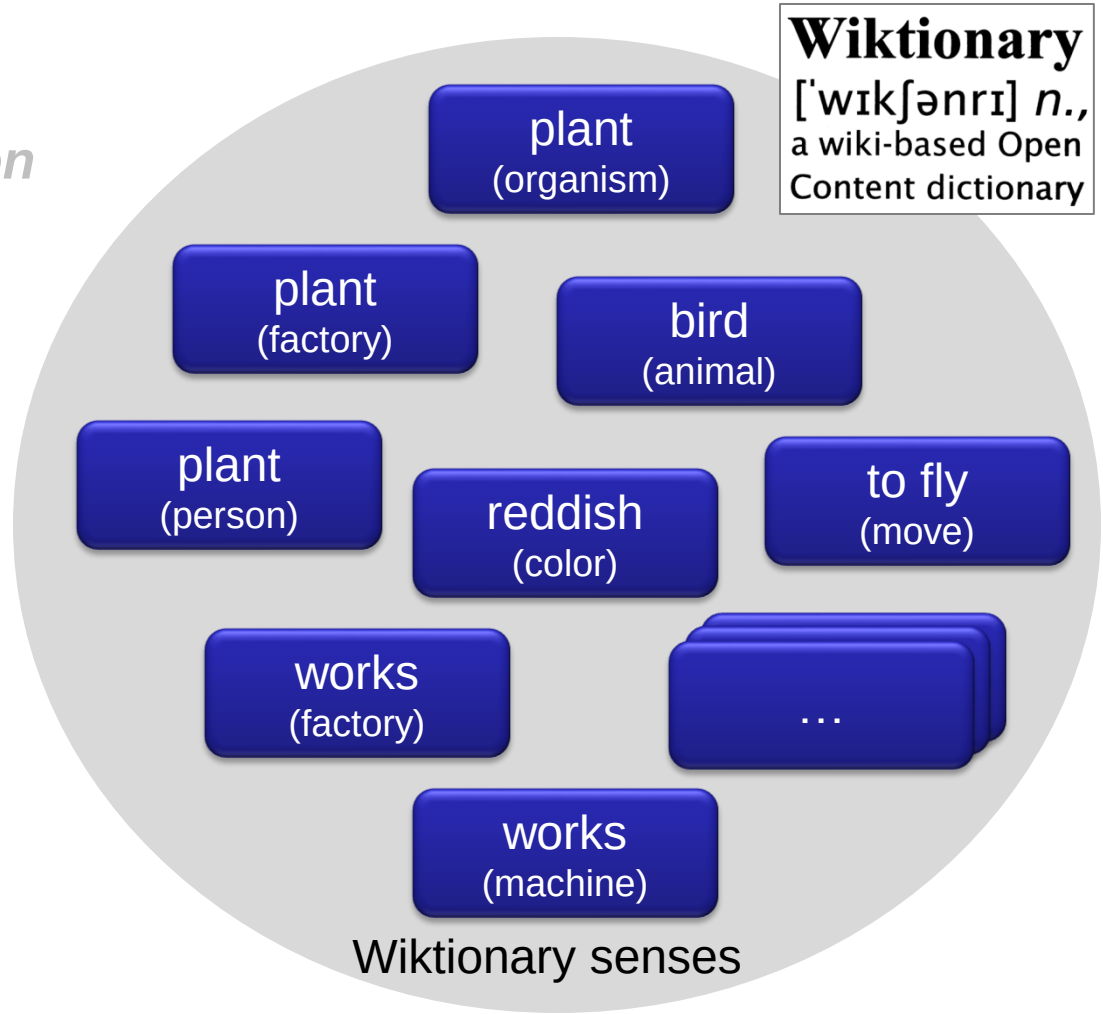
Linking Wiktionary and WordNet at the Sense Level

TECHNISCHE UNIVERSITÄT DARMSTADT

A two-step approach:

1. *Candidate extraction*
2. *Candidate disambiguation*

WordNet synsets



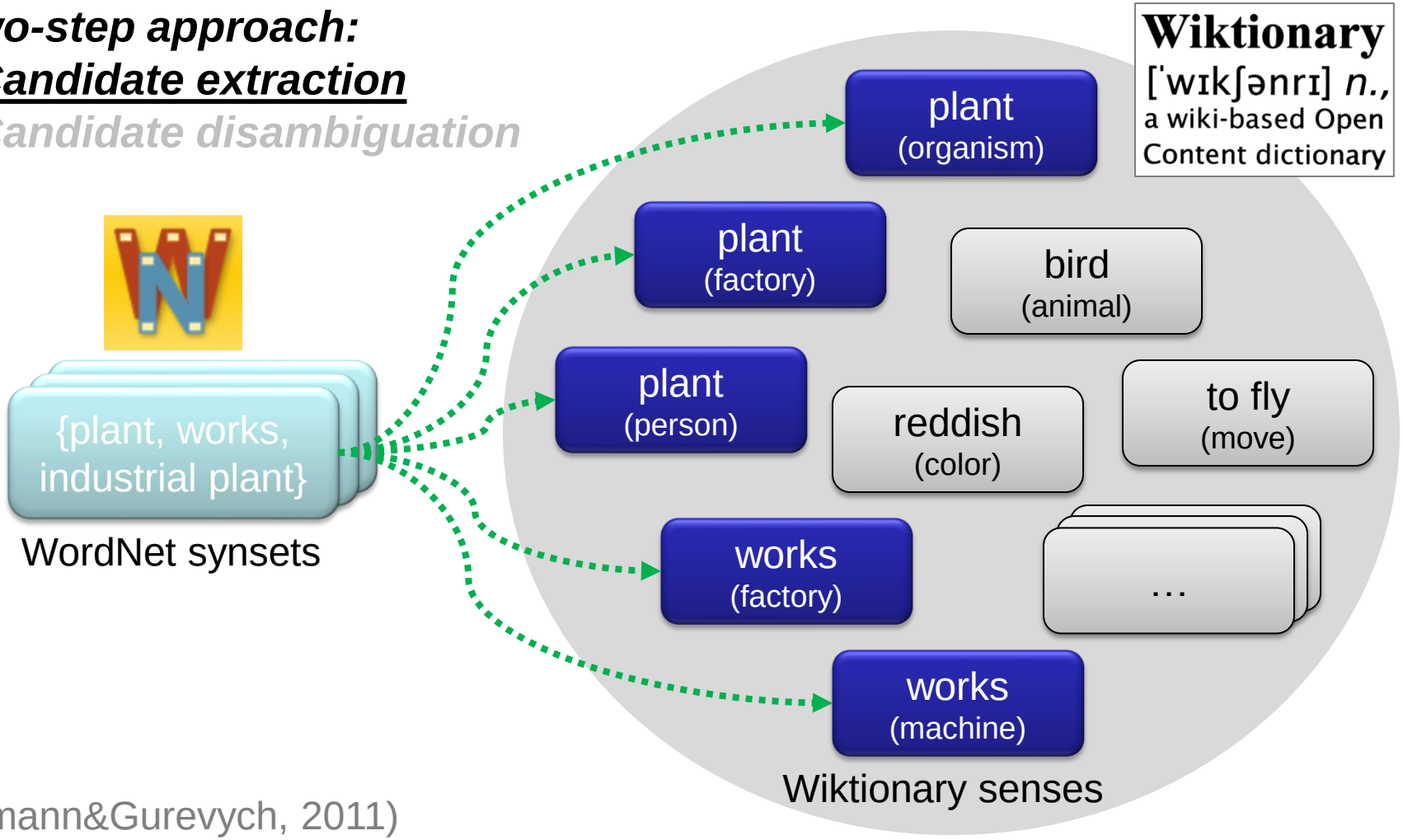
Wiktionary senses

Linking Wiktionary and WordNet at the Sense Level

TECHNISCHE UNIVERSITÄT DARMSTADT

A two-step approach:

1. ***Candidate extraction***
2. ***Candidate disambiguation***



(Niemann&Gurevych, 2011)

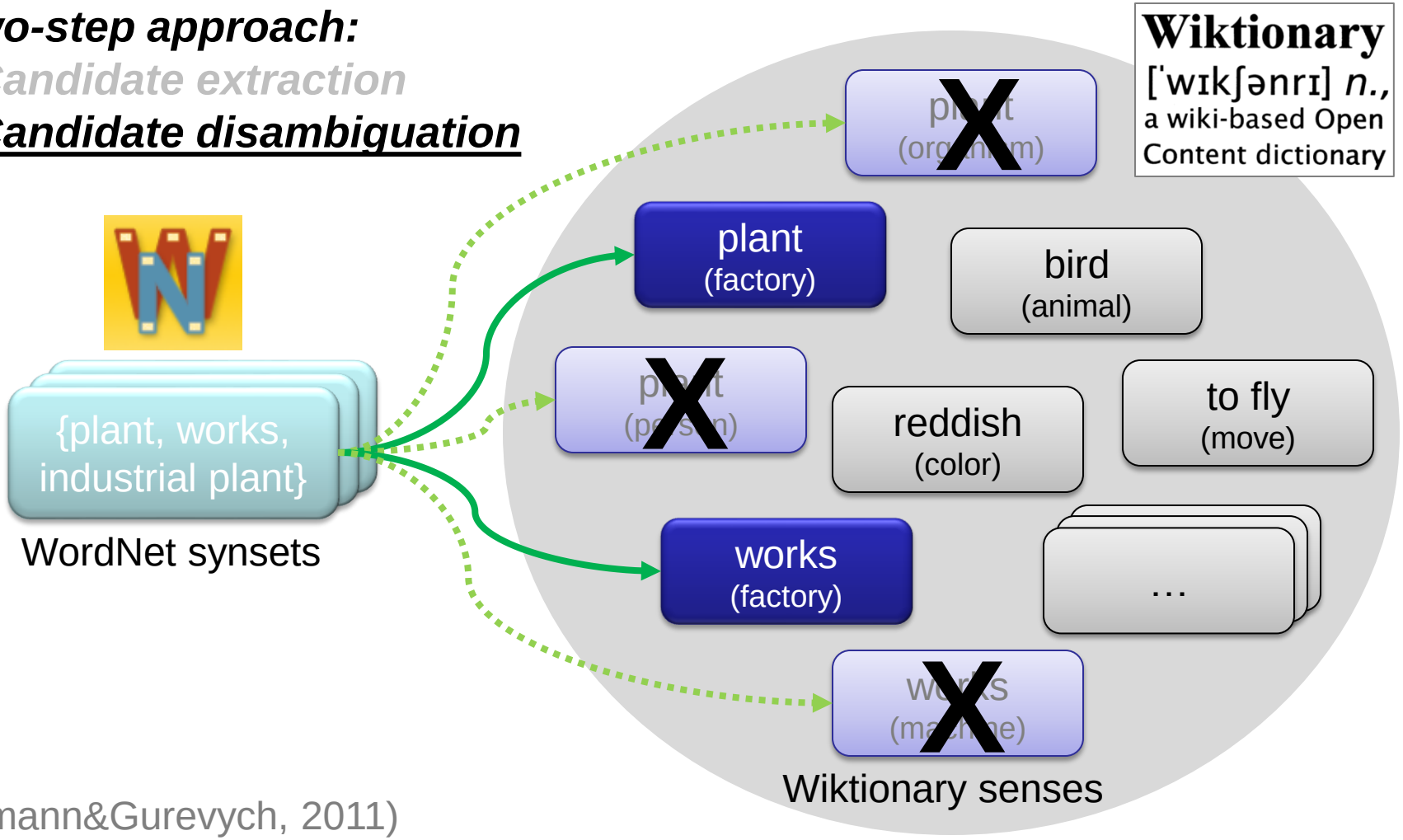
UKP

Linking Wiktionary and WordNet at the Sense Level

TECHNISCHE UNIVERSITÄT DARMSTADT

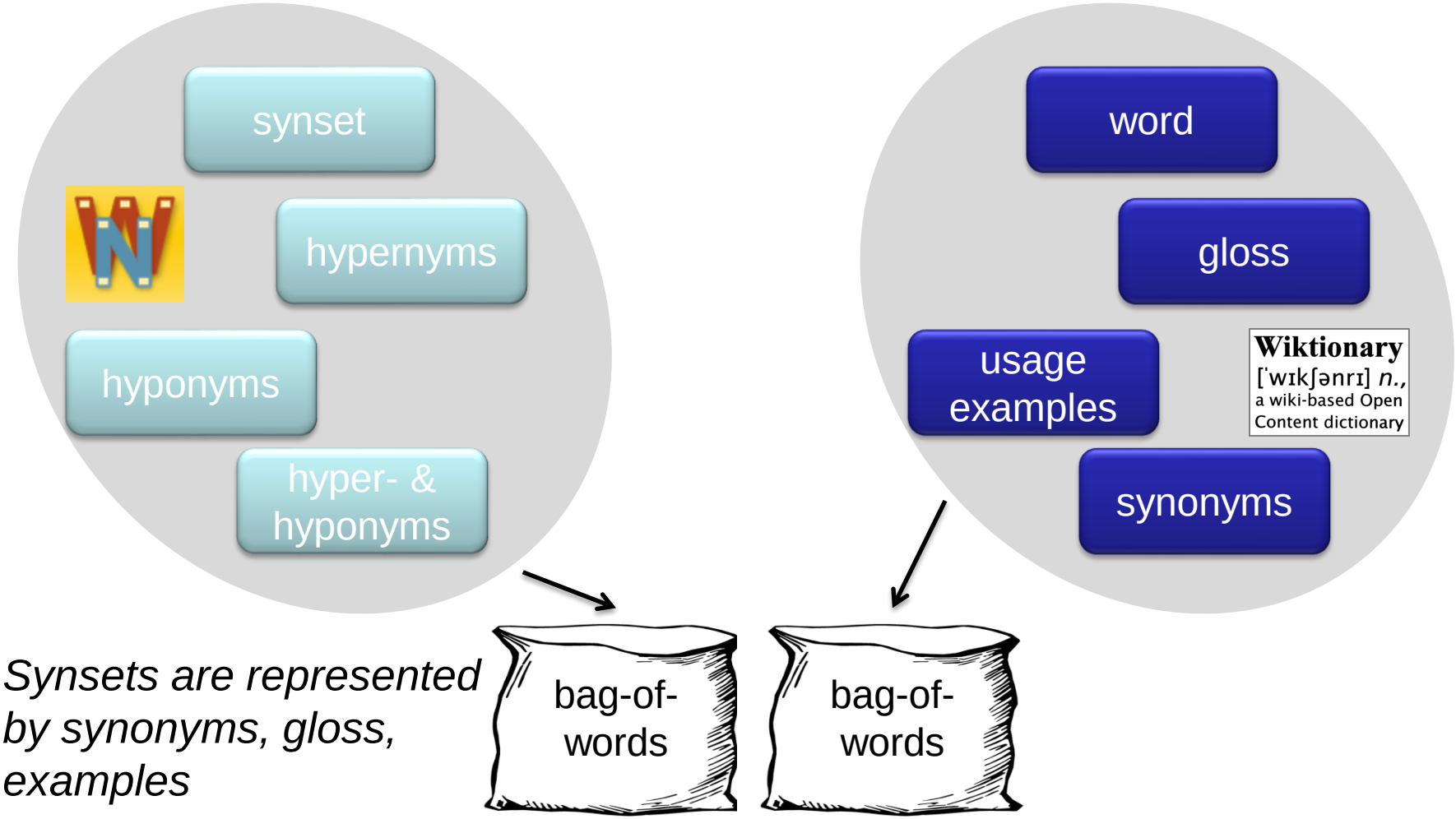
A two-step approach:

1. ***Candidate extraction***
2. ***<u>Candidate disambiguation</u>***



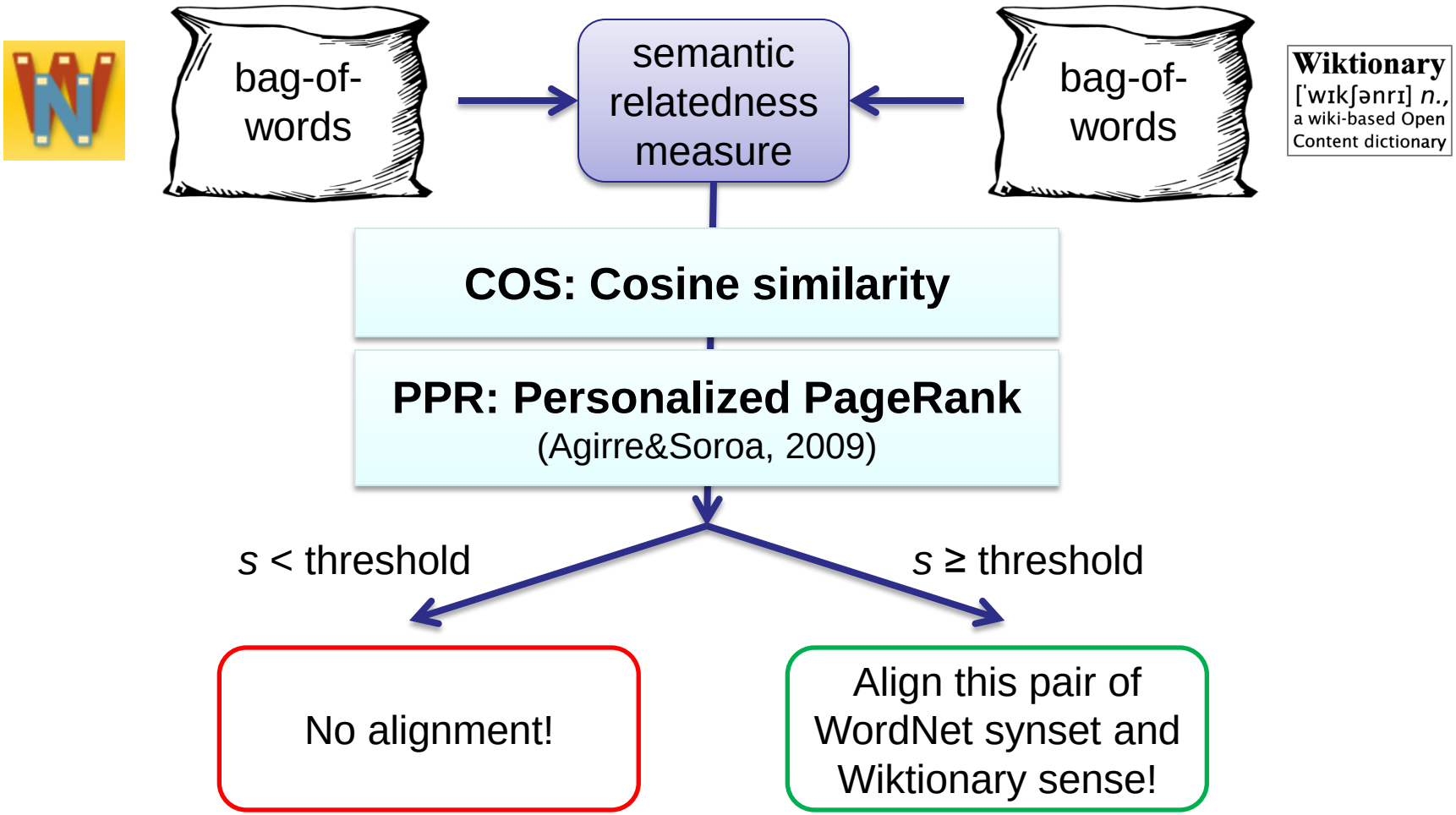
(Niemann&Gurevych, 2011)

Disambiguation: BoW Representation



Synsets are represented by synonyms, gloss, examples

Disambiguation: Alignment Classification



Evaluation

Performance of automatic sense alignment framework:

- F1 ~ 0.66, Precision ~ 0.67, Recall ~ 0.65

Open issues to be addressed in future work:

- false negatives *"same meaning, but was not aligned"*
- Very different wording
- Similar senses but slightly below threshold
- Pointing to another entry rather than a content-based gloss

- false positives *"different meaning, but have been aligned"*
- Similar wording, but refer to different concepts
- Generic- versus domain-specific vocabulary

Increased Coverage: Domains

	Wiktionary AND WordNet	Only Wiktionary	Only 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

Enriched Sense Representation

TECHNISCHE UNIVERSITÄT DARMSTADT

Access to complementary information

Synonyms

Gloss

Example sentence

Subsumption hierarchy

Synset organization

…

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

Pronunciation

Etymology

Usage

Quotations

Related terms

Translations

…

UKP

NLP Applications of Linked Lexical Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

The increased coverage and the enriched sense representation yield synergies.

Previously shown:

- Linking FrameNet, VerbNet, and WordNet for semantic parsing (Shi and Mihalcea, 2005)
- Linking VerbNet, FrameNet and PropBank for semantic role labeling (Palmer, 2009)
- Linking WordNet and Wikipedia for word sense disambiguation (Navigli and Ponzetto, 2010)
- Linking WordNet and Wiktionary for measuring verb similarity (Meyer and Gurevych, 2012)

Future work:

- Semantic relatedness, information retrieval, information extraction,…
- Your application?

Part 1: Ingredients and Techniques

TECHNISCHE UNIVERSITÄT DARMSTADT

Ingredients: Lexical Resources

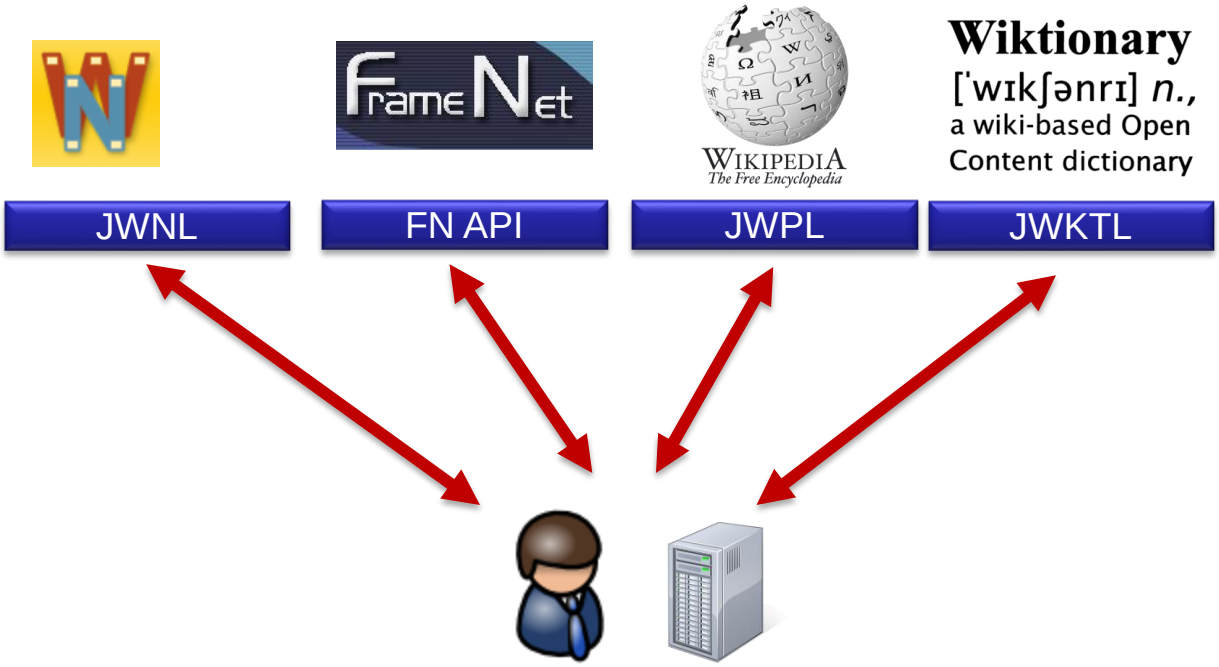
- **Elements of Lexical Resources**
- **Classic Lexical Resources**
- **Collaborative Lexical Resources**

Integration Technique: Automatic Sense Linking

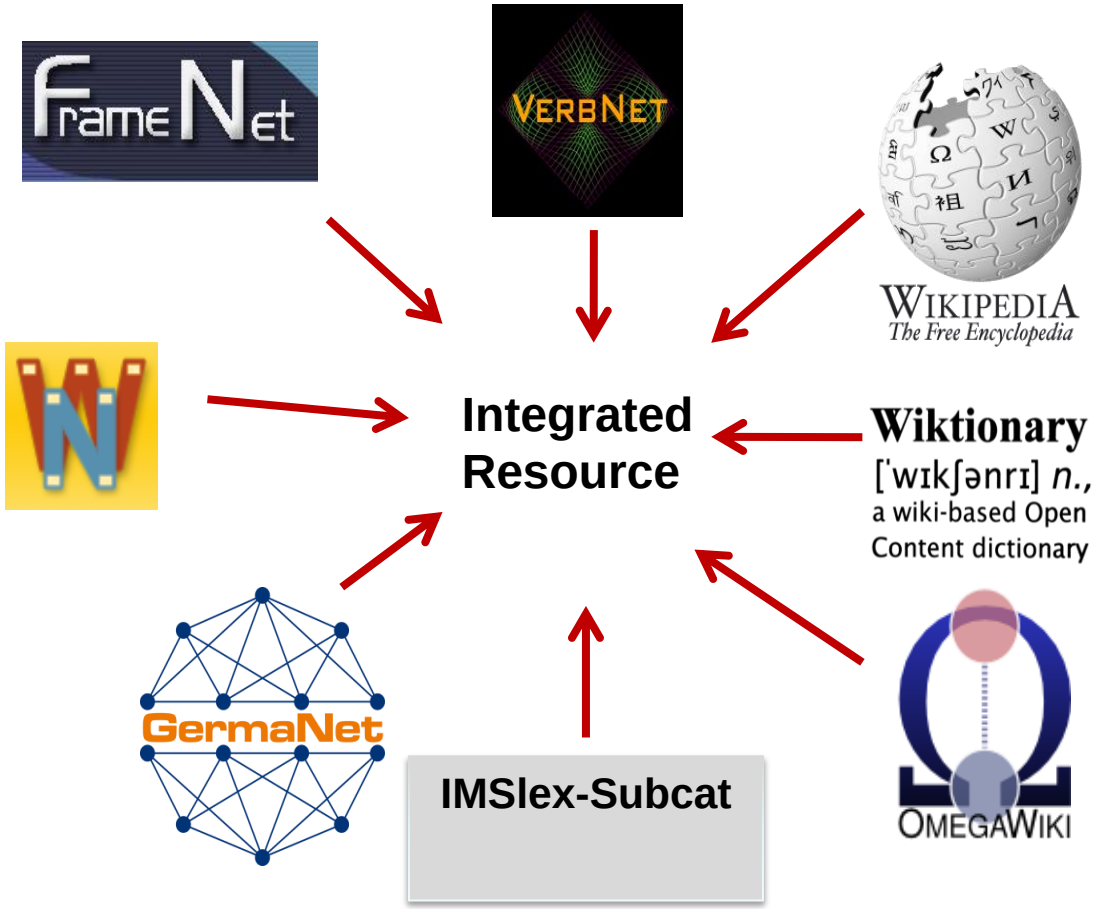
Interoperability Technique: Standardizing

Interoperability – Motivation

- Different APIs …

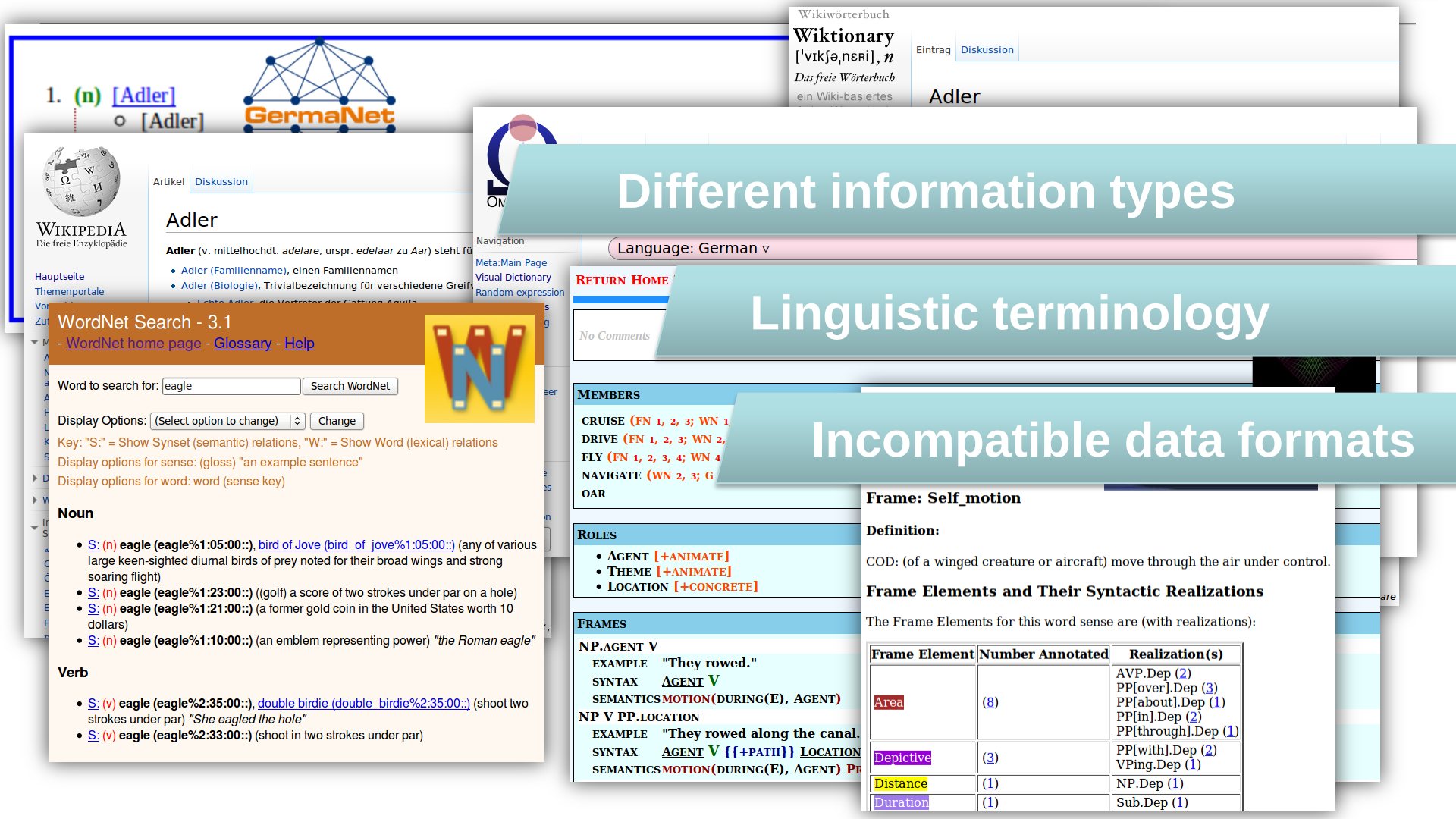


Integration of Lexical Resources at the Representation Level



Integration at the representation level: **standardization of resources**

Heterogeneous Lexical Resources



Different information types

Linguistic terminology

Incompatible data formats

Dimensions of Interoperability

Structural Interoperability:

Uniform lexicon structure – given by the main organizational units

- e.g., harmonizing the *synset-based* organization (WordNet), the *headword-based* organization (Wiktionary) and the *frame-based* organization (FrameNet)

Interoperability at the level of linguistic terminology:

Uniform Data Categories – these are descriptions of the meaning of linguistic terms

- e.g., harmonizing the terms *lexical unit* (FrameNet) and *word sense* (Wiktionary)

Standardizing Lexical Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

ISO Standards for lexical resources:

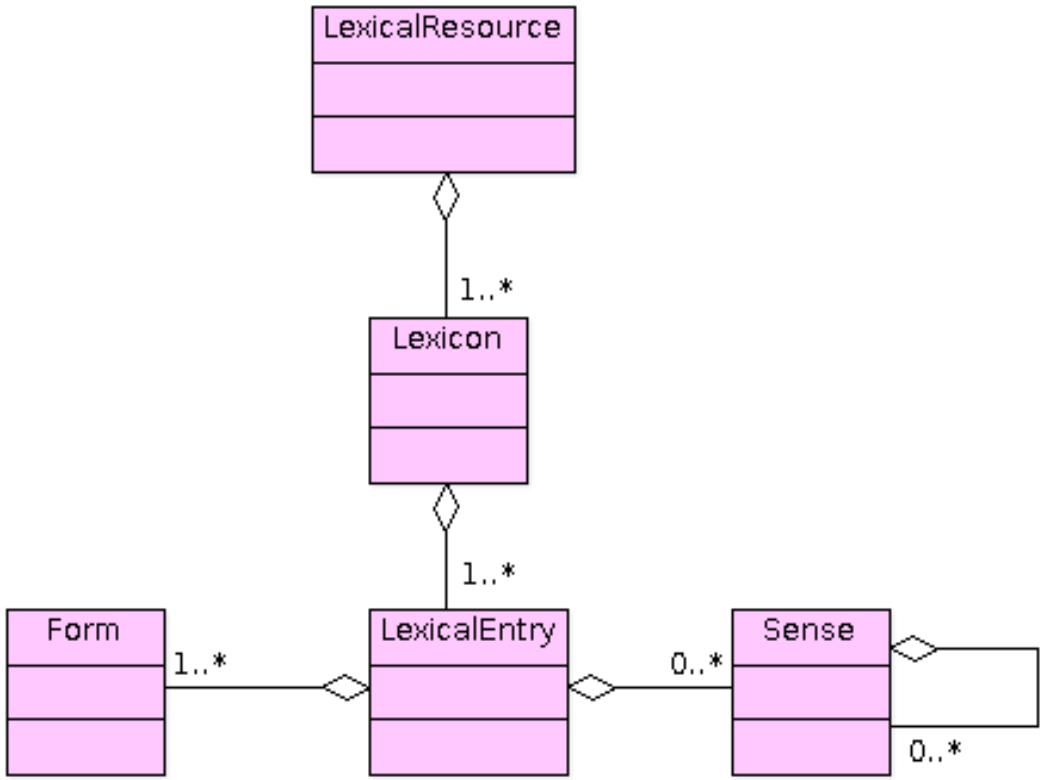
- ISO 24613:2008 **Lexical Markup Framework (LMF)**
- ISO 12620:2009 Data Category Registry: **ISOcat**

Standardization of lexical resources according to LMF makes them interoperable:

- at the structural level **LMF classes**
- at the level of linguistic terminology **ISOcat Data Categories**

UKP

ISO-LMF is an Abstract Standard …



- Classes are given in LMF, but they have no attributes.
- Attributes and their values = linguistic terminology
- In order to make use of ISO-LMF, attributes and their values have to be defined for each class.
- The result is **an instantiation of the abstract LMF standard – a lexicon model** that can be populated by lexicon data.

An Instantiation of LMF for NLP: UBY-LMF

TECHNISCHE UNIVERSITÄT DARMSTADT

UBY-LMF

- is an instantiation of LMF that is directly usable
- specifies attributes and their values for all LMF classes used
- extends the LMF standard by two classes: SemanticLabel and Frequency

Many different **types of lexical resources** in UBY-LMF:

- expert built vs. collaboratively constructed
- differently organized:
  - headword-based
  - wordnets
  - lexical resources based on frame semantics
  - subcategorization lexicons

UBY-LMF – Design Principles

TECHNISCHE UNIVERSITÄT DARMSTADT

Extensibility

- further languages
- further lexicons
- automatically mined information types (e.g. domain labels)

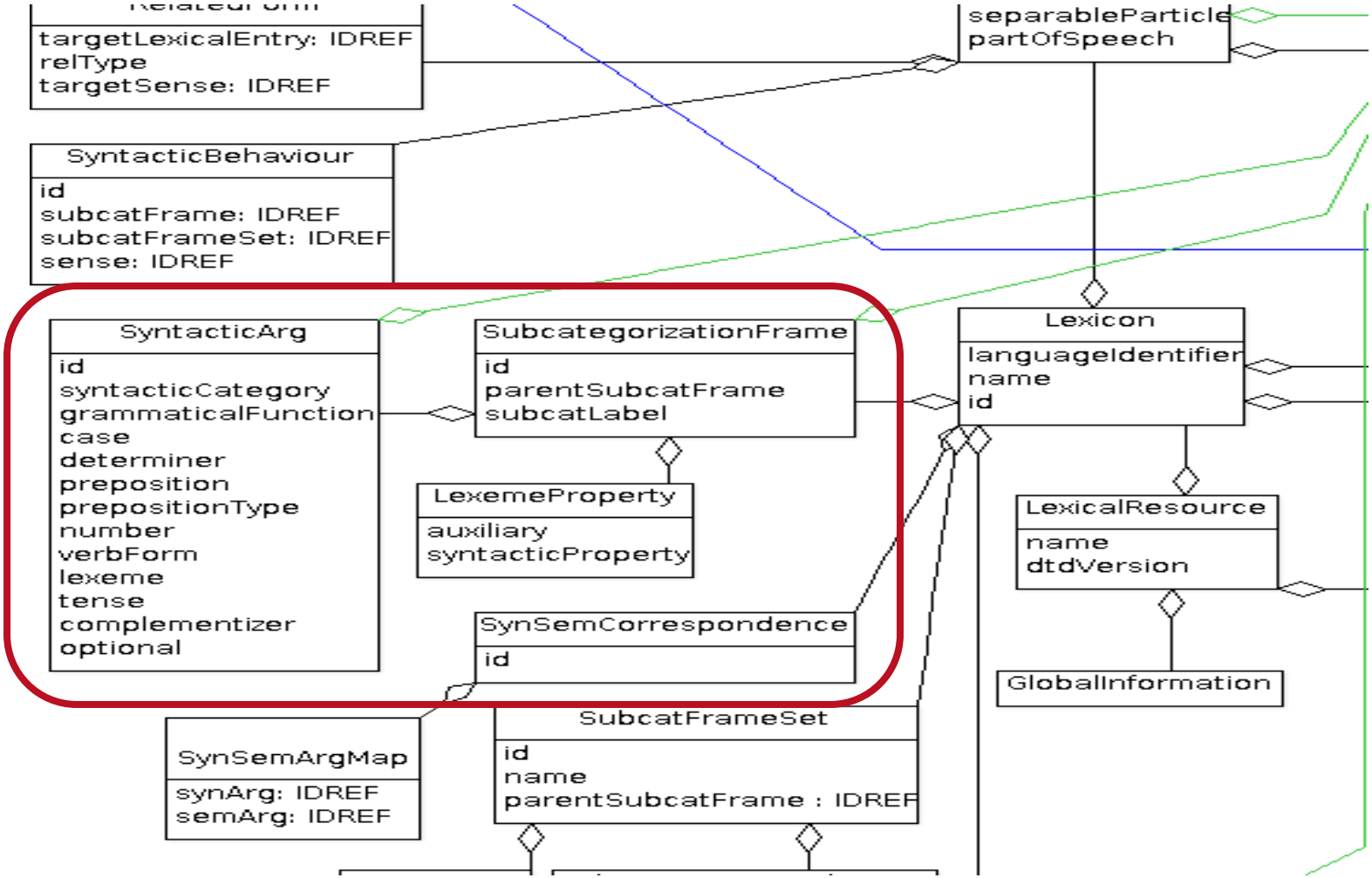
No information loss

- including conflicting information
- e.g. keeping the original sense key in the `MonolingualExtRef` class

Implementation of UBY-LMF

- Definition of a lexicon model in UML (Unified Modeling Language)
- Implementation of the lexicon model in Java
- Conversion of lexical resources to UBY-LMF according to the lexicon model
- Import of converted resources into an SQL database based on an Object-Relational Mapping using Hibernate
- Export format: XML

Example: SCFs in UBY-LMF – Specification of Syntactic Arguments



SCFs in UBY-LMF – Example

lachen

NN.Pp

laugh

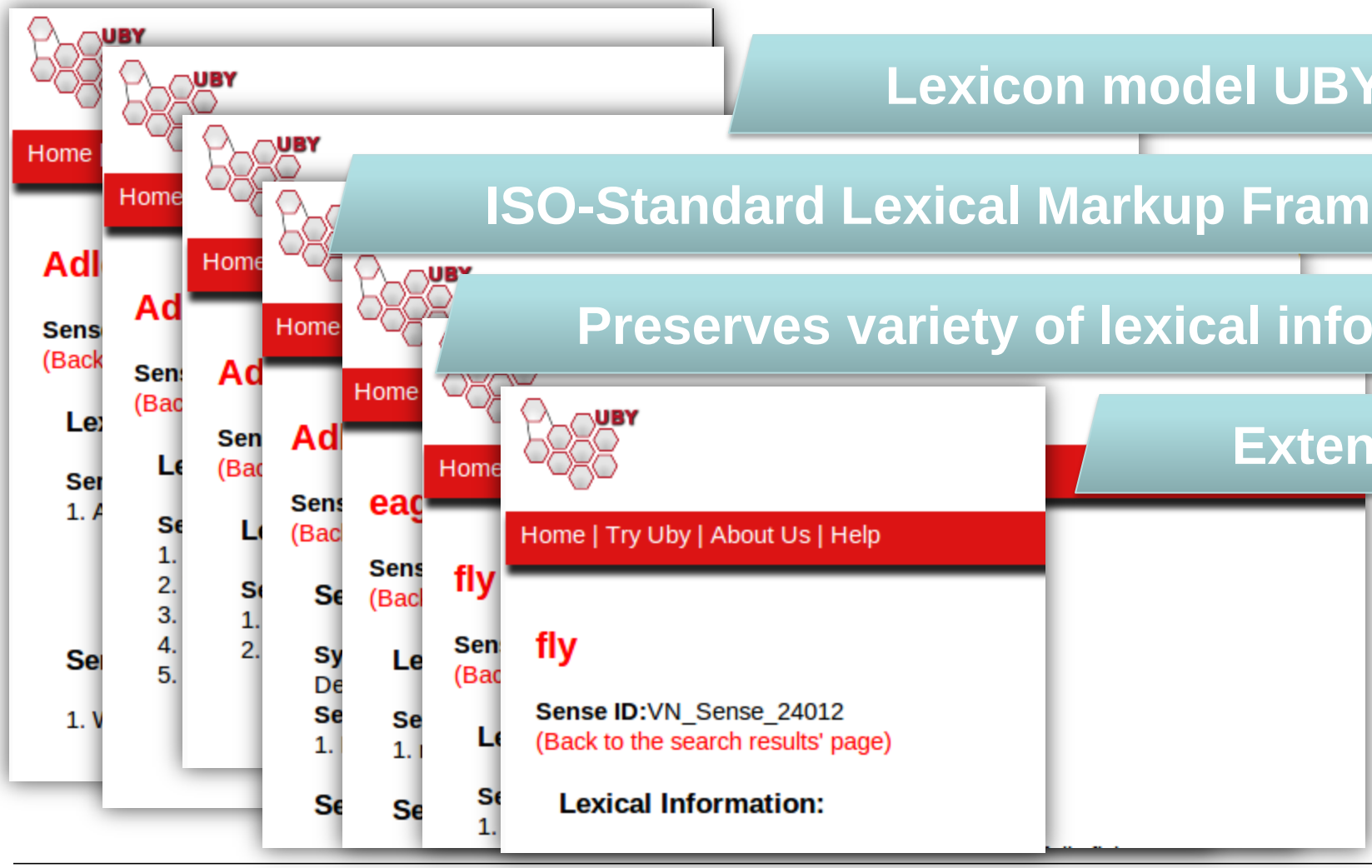
NP V PP {about}

UBY

GermaNet Syntactic Arguments	**VerbNet Syntactic Arguments**
syntacticCategory: nounPhrase grammaticalFunction: subject	syntacticCategory: nounPhrase grammaticalFunction: subject
syntacticCategory: prepositionalPhrase	syntacticCategory: prepositionalPhrase
grammaticalFunction: prepositionalComplement	grammaticalFunction: prepositionalComplement
Preposition: -	Preposition: about

Homogeneous Lexical Resources

TECHNISCHE UNIVERSITÄT DARMSTADT



Lexicon model UBY-LMF

ISO-Standard Lexical Markup Framework

Preserves variety of lexical information

Extensible

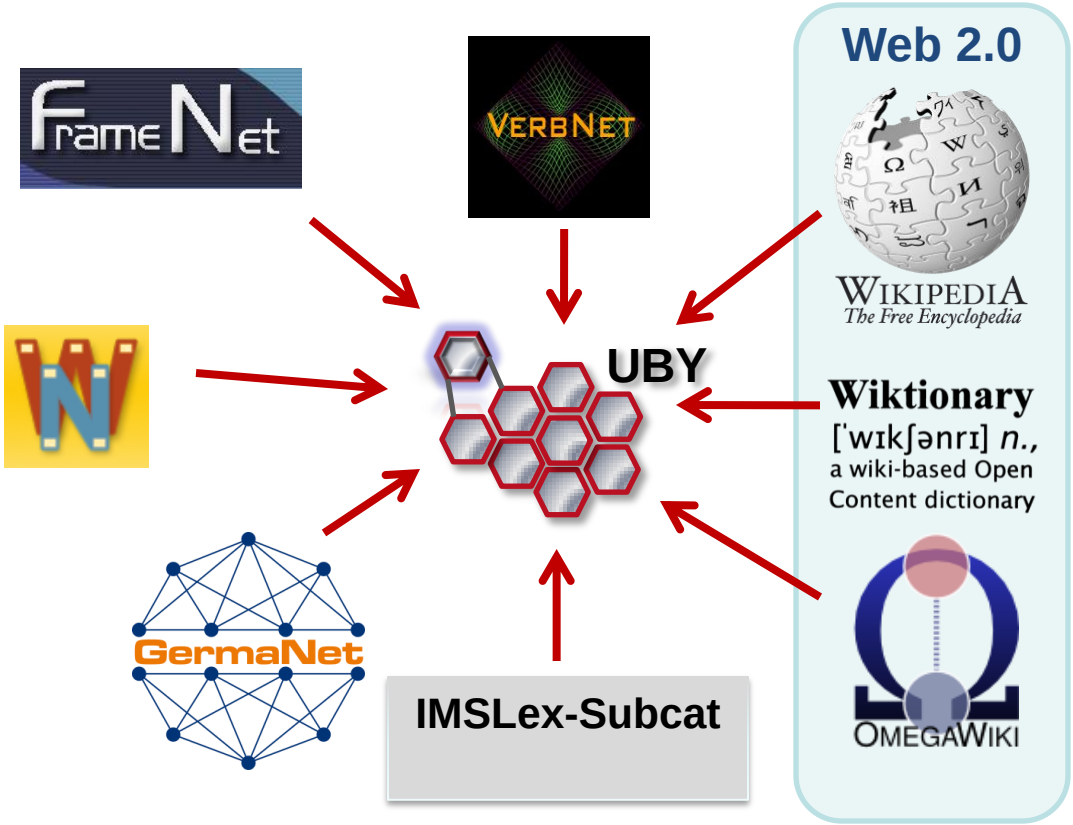
UKP

Outline

Part 1: Ingredients and Techniques

Part 2: Recipes

UBY – the Main Ingredient



Goal: to exploit wide variety of lexical knowledge

Joint work with
Iryna Gurevych, Kostadin Cholakov, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, Tri-Duc Nghiem

Part 2: Recipes

TECHNISCHE UNIVERSITÄT DARMSTADT

Tools and Planning Guide: What UBY has to offer

Mixed Starters: How to query UBY (for WSD)

Appetizer: UBY as UIMA resource

Main Dish: UBY for UIMA-based Semantic Tagging

Dessert: Cross-lingual verb sense linking

What UBY has to offer

Tools

UBY – Data and Tools

TECHNISCHE UNIVERSITÄT DARMSTADT

Web Interface

https://uby.ukp.informatik.tu-darmstadt.de/webui/

Database Dumps

UBY

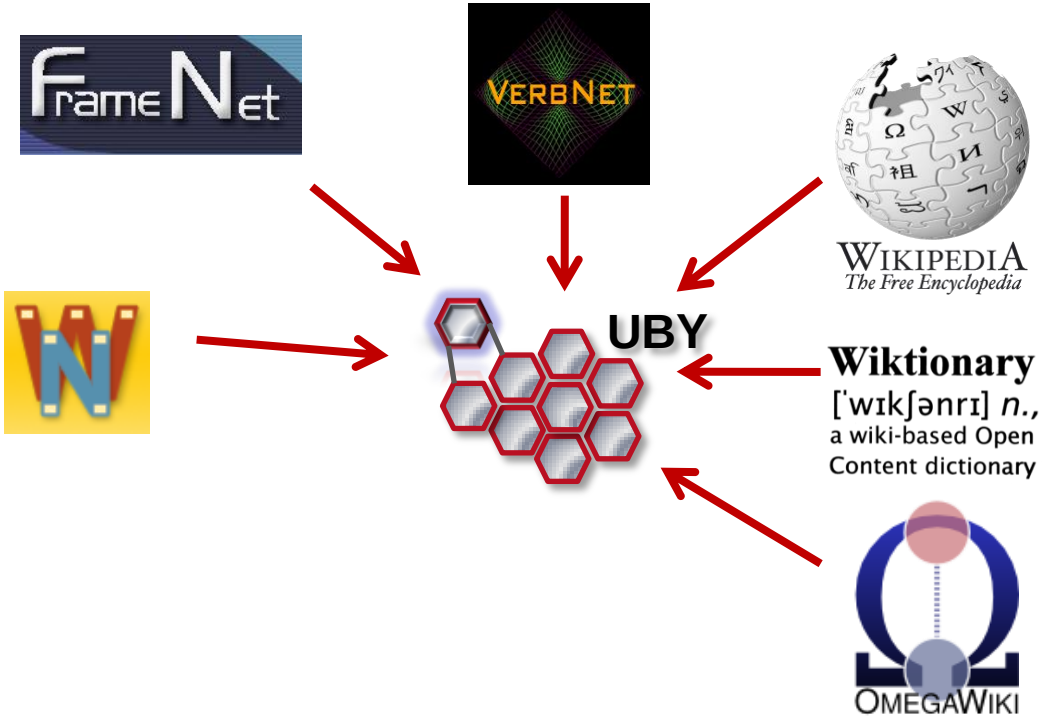
http://uby.ukp.informatik.tu-darmstadt.de/uby/

Open Source API (JAVA)

http://code.google.com/p/uby/

UBY Web UI – a Small Glimpse of UBY

- Resources with **open licenses**
- **Selected** information types are presented



UBY Web UI – Textual View

Textual View: allows to **list senses across all resources**, to display **sense details** and to perform **sense comparisons.**

Text Browser
Visual Browser

Resources (3)

All Resources
FrameNet
OmegaWikide
OmegaWikien
VerbNet
WikipediaDE
Wikipedia
WiktionaryDE
WiktionaryEN
WordNet

Sense Comparison View

Drag and drop two senses into this box to compare.

Compare (4)

OmegaWikien

align (verb)
To arrange in a straight line. (1)

align (verb)
To bring into cooperation or agreement with a particular group, party, cause, etc.

Wikipedia

Align (noun)
Align is a privately held IT solutions company with key competencies in "Technology Transformation Initiatives" – designing, deploying, moving and consolidating technologies and advanced network infrastructures. Specializing in IT Strategy consulting, technology relocation services, network infrastructure design and build-out, contact center technology and IT asset management (ITAM) solutions, the company also provides communications products and integration services for call centers and other facilities requiring large-scale computer telephony installations. Align operates from offices in New York; London; Chicago; Princeton, New Jersey; and Toronto. The company serves customers in such industries as finance, health care and energy. Clients have included Blackrock, BP and Dun and Bradstreet.

WiktionaryEN

align (verb)
To form in line; to fall into line.

align (verb)
To adjust or form to a line; to range or form in line; to bring into line.

Expand...

align
Sense ID:WN_Sense_142794

Lexical Information: (2)

Semantic Labels:
1. verb.stative

Semantic Information:

Semantic Predicate

- Semantic Argument
  - Semantic Role: something

Synset (Original SynsetID:2658867):
Definition: be or come into adjustment with

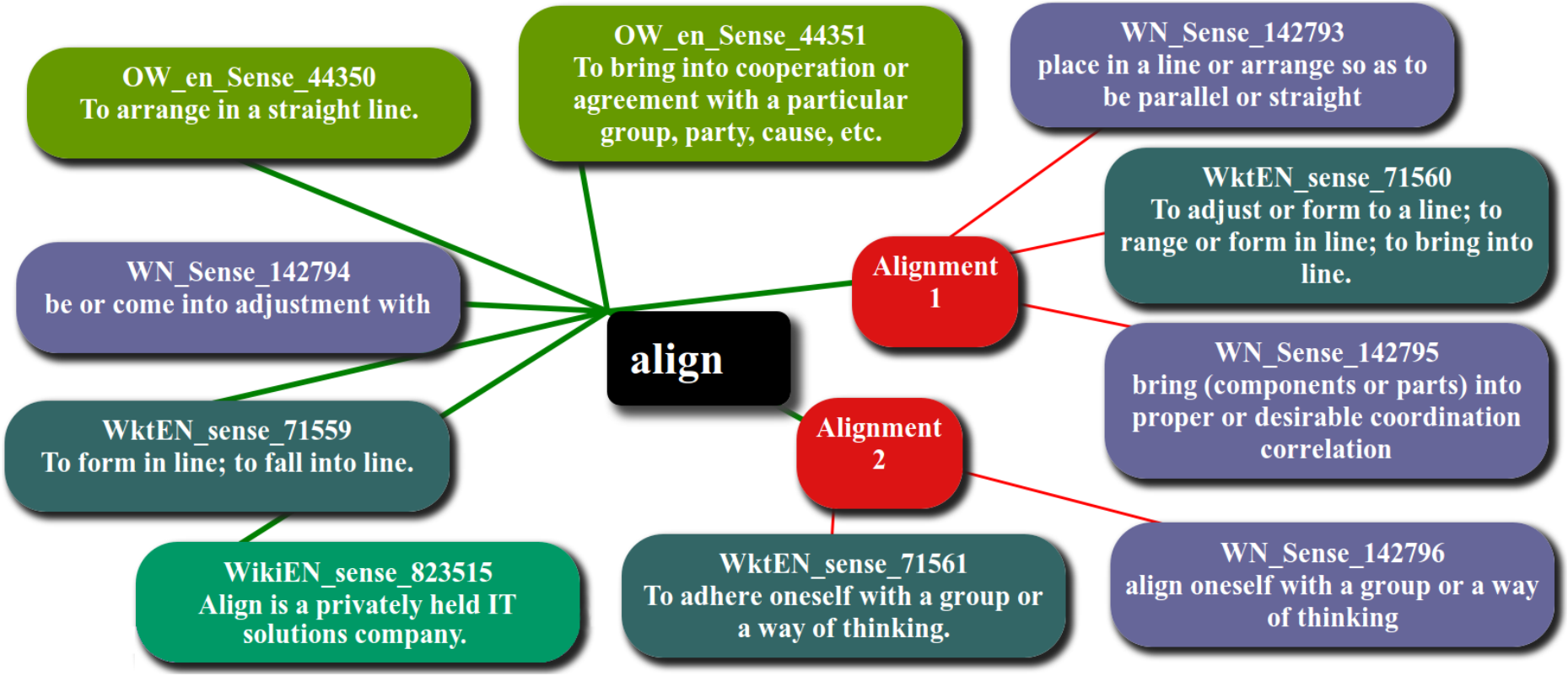
Syntactic Information:

1. subject_nounPhrase align complement_prepositionalPhrase

Click here for more details ...

UBY Web UI – Visual View

Visual view: allows to **explore the sense alignments.**



UBY Java API

The UBY API is **open source** at Google Code:
http://code.google.com/p/uby/

Getting Started:

1. Download a UBY database dump
2. Import the dump into a MySQL database
3. Start using the UBY API

The UBY API is work in progress!

Many API methods need to be added – consider contributing!

UBY Database Dumps

Downloads:

http://uby.ukp.informatik.tu-darmstadt.de/uby/

- Resources with **open licenses**
- database dumps in two different sizes:
  - all UBY resources along with pairwise alignments (without GermaNet, IMSLex), uby_open_<version>
  - all UBY resources along with pairwise alignments without Wikipedia, GermaNet, IMSLex, uby_medium_<version>

UBY API – Getting Started

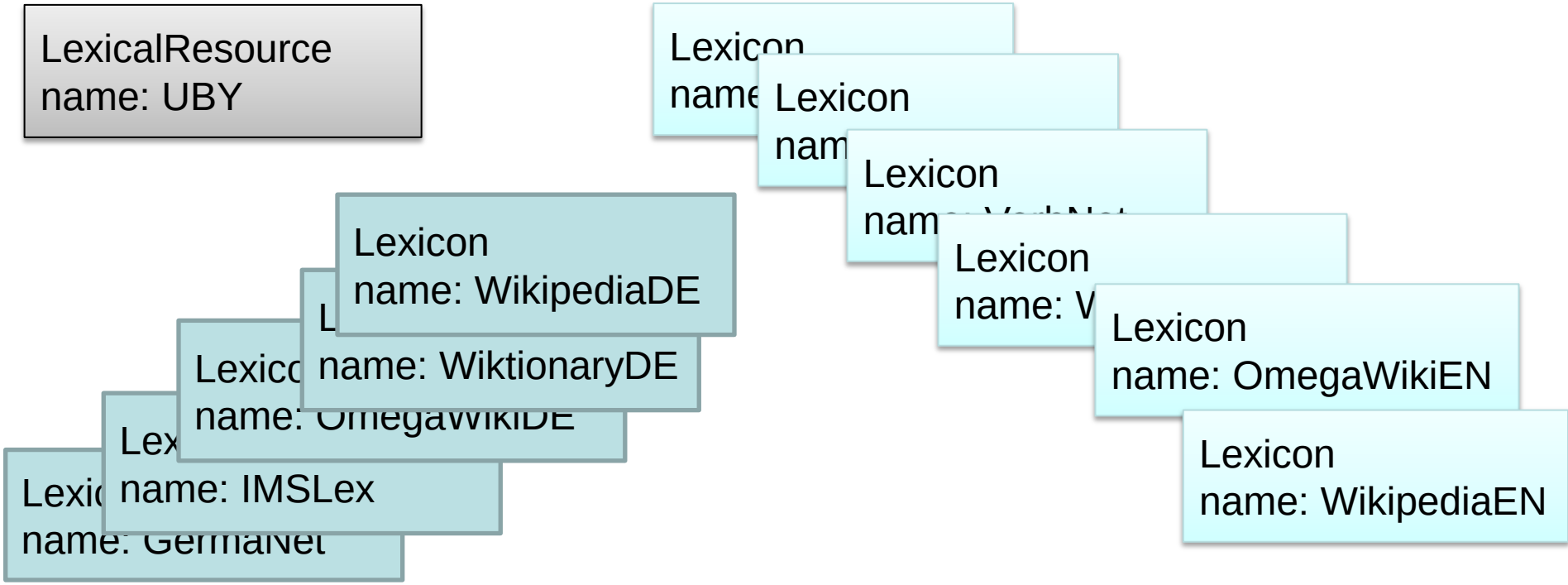
- Recommended: using the UBY API with Maven
- Currently, it is necessary to use a specific **settings.xml** that configures the access to the public Maven repository maintained by UKP Lab.
- We plan to deploy the UBY API on Maven Central soon!

```
<dependency>
  <groupId>de.tudarmstadt.ukp.uby</groupId>
  <artifactId>de.tudarmstadt.ukp.uby.lmf.api-asl</artifactId>
  <version>0.3.0</version>
  <type>jar</type>
  <scope>compile</scope>
</dependency>
```

UBY API – Based on UBY-LMF

UBY-LMF has been implemented in Java. An object-relational mapping by means of the Hibernate framework allows mapping any instance of UBY-LMF to a SQL database.

UBY is an instance of the UBY-LMF lexicon model:



UBY API – Based on UBY-LMF

UBY (as instance of the LexicalResource class) aggregates instances of the Lexicon class.

- Lexicon instances: WordNet, FrameNet, Wiktionary, Wikipedia, …
- Example: querying the names of UBY lexicons

```
Uby uby = new Uby(dbConfig);

List<String> lexiconNames = uby.getLexiconNames();
```

UBY API – Uniform Access to UBY Resources

The uniform representation of resources allows querying across all UBY lexicons.

- Option to filter by POS

- Example: querying lexical entries for a given lemma and POS **across all UBY lexicons** (here: noun “album”)

```
Uby uby = new Uby(dbConfig);

List<LexicalEntry> nounEntries =
uby.getLexicalEntries("album", EPartOfSpeech.noun, null);
```

UBY API – Uniform Access to UBY Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

Individual UBY lexicons are queried the same way:

- First, retrieve the lexicon to be queried
- Option to filter by POS

- Example: querying UBY WordNet for the lexical entry of the noun “album”

```
Uby uby = new Uby(dbConfig);

Lexicon wordNet = uby.getLexiconByName("WordNet");

List<LexicalEntry> lexEntries =
uby.getLexicalEntries("album", EPartOfSpeech.noun, wordNet);
```

What UBY has to offer

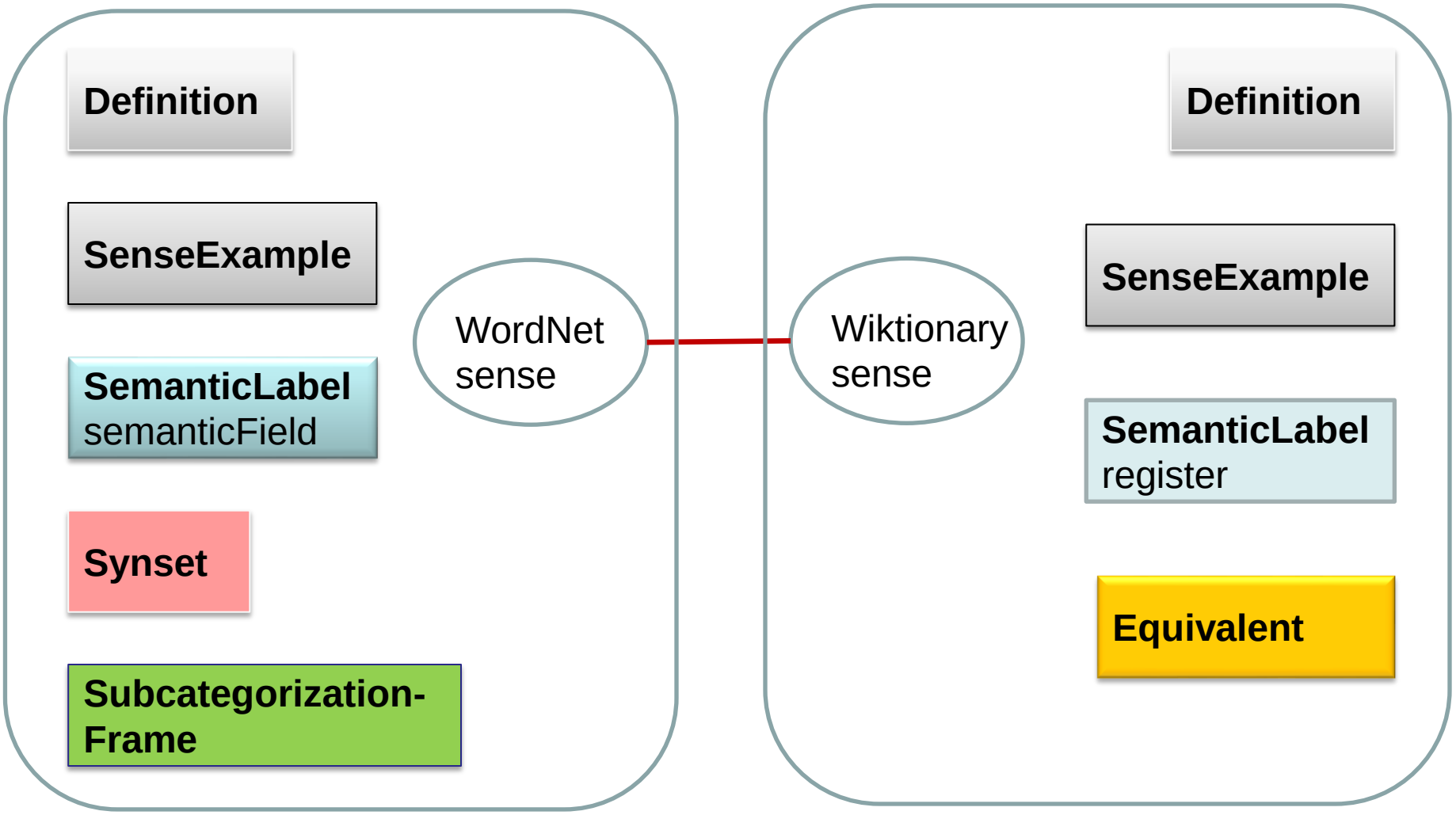
Planning Guide

Increased Coverage of Lemmas (and their Senses)

UBY lexicons contain different, complementary lemmas:
over 3.08 million unique lemma-POS combinations for English

EN Lexicons	noun	verb	adjective
5	1	699	-
4	1,630	1,888	430
3	8,439	1,948	2,271
2	53,856	4,727	12,290
1	2,900,652	50,209	41,731
Σ (unique EN)	3,080,771		

Enriched Senses Based on Sense Alignments



Enriched Senses Based on Sense Alignments

Lexicon pair	Languages	SenseAxis
WN–WP-en	EN–EN	50,351
WN–WKT-en	EN–EN	99,662
WN–VN	EN–EN	40,716
FN–VN	EN–EN	17,529
WP-en–OW-en	EN–EN	3,960
WP-de–OW-de	DE–DE	1,097
WN–OW-de	EN–DE	23,024
WP-en–WP-de	EN–DE	463,311
OW-en–OW-de	EN–DE	58,785
UBY	**All**	**758,435**

Access to all available, **partly complementary information types** attached to the aligned senses, e.g., semantic relations, subcategorization frames, encyclopedic or translation information.

For English, ca. **32,000** senses simultaneously take part in **at least two pairwise sense alignments**, i.e. information from 3 UBY lexicons is available.

UBY as Sense Inventory

UBY provides sense inventories at different levels of granularity.

- Example: VerbNet – WordNet alignment
- coarse-grained vs. fine-grained sense inventory

VerbNet:

sing
performance-26.7-1-1

WordNet:

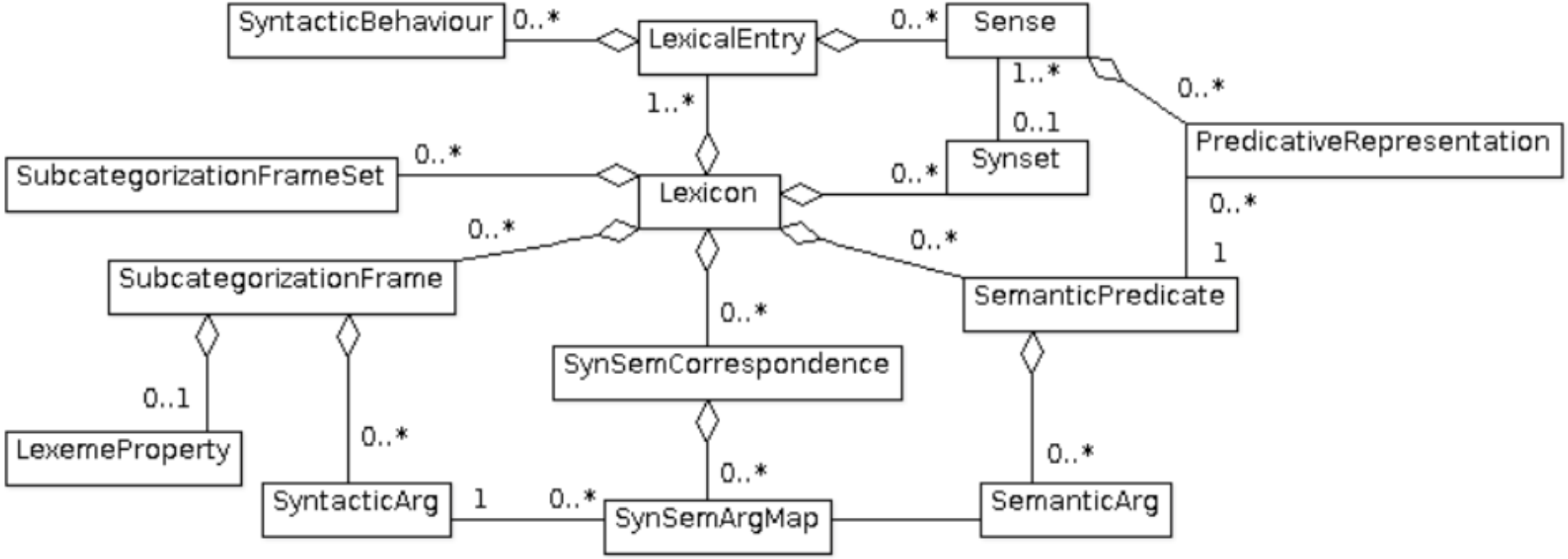
sing
sing%2:36:01
deliver by singing

sing
sing%2:36:00
produce tones with the voice

Information Types in UBY – Selection

TECHNISCHE UNIVERSITÄT DARMSTADT

Main organizational LMF classes



UKP

Information Types in UBY – Sense Related

TECHNISCHE UNIVERSITÄT DARMSTADT

Information types attached to senses

SenseExample

Definition

SemanticLabel

Sense

SenseRelation

Synset

Equivalent

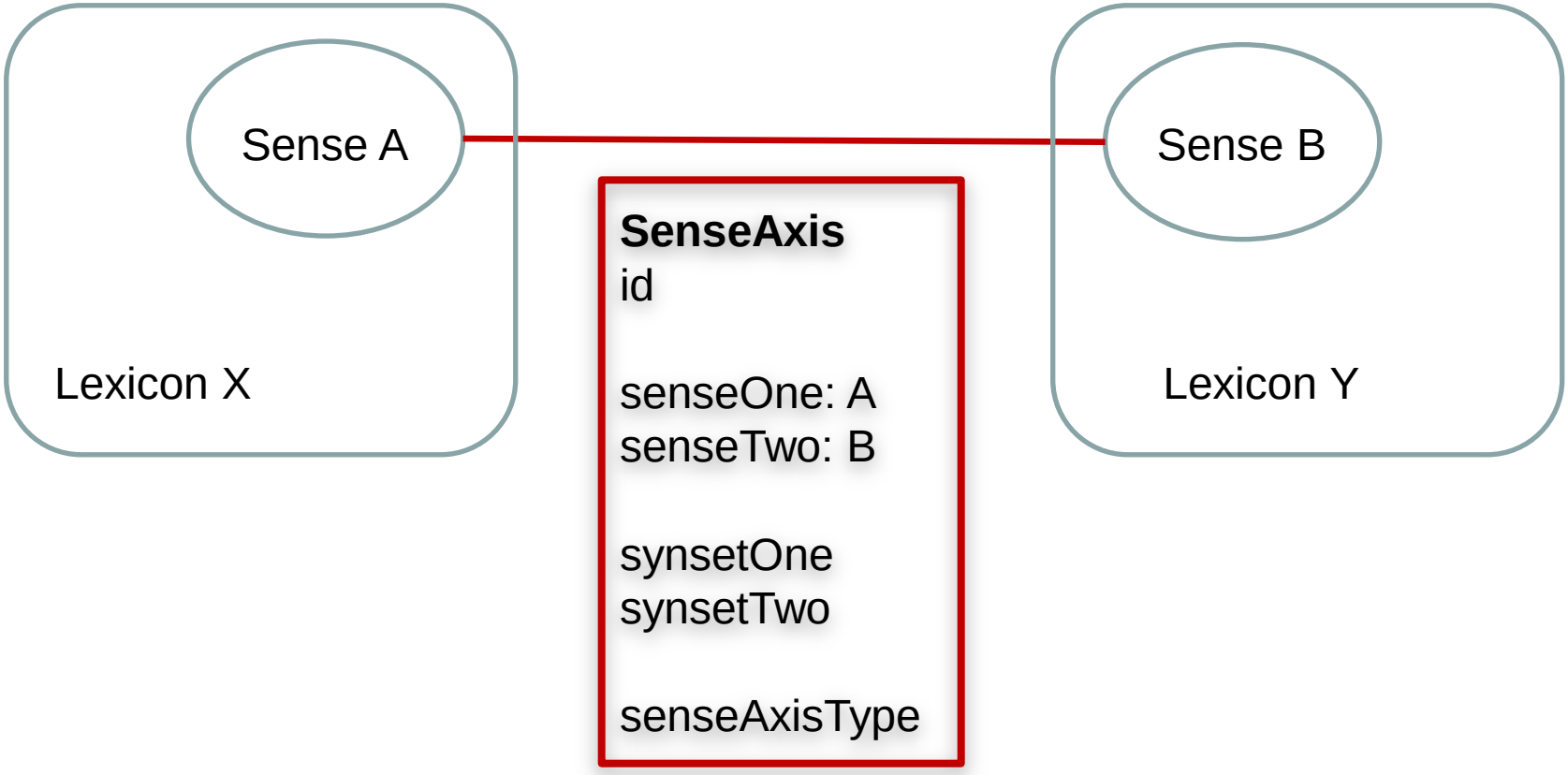
Subcategorization-Frame

Context

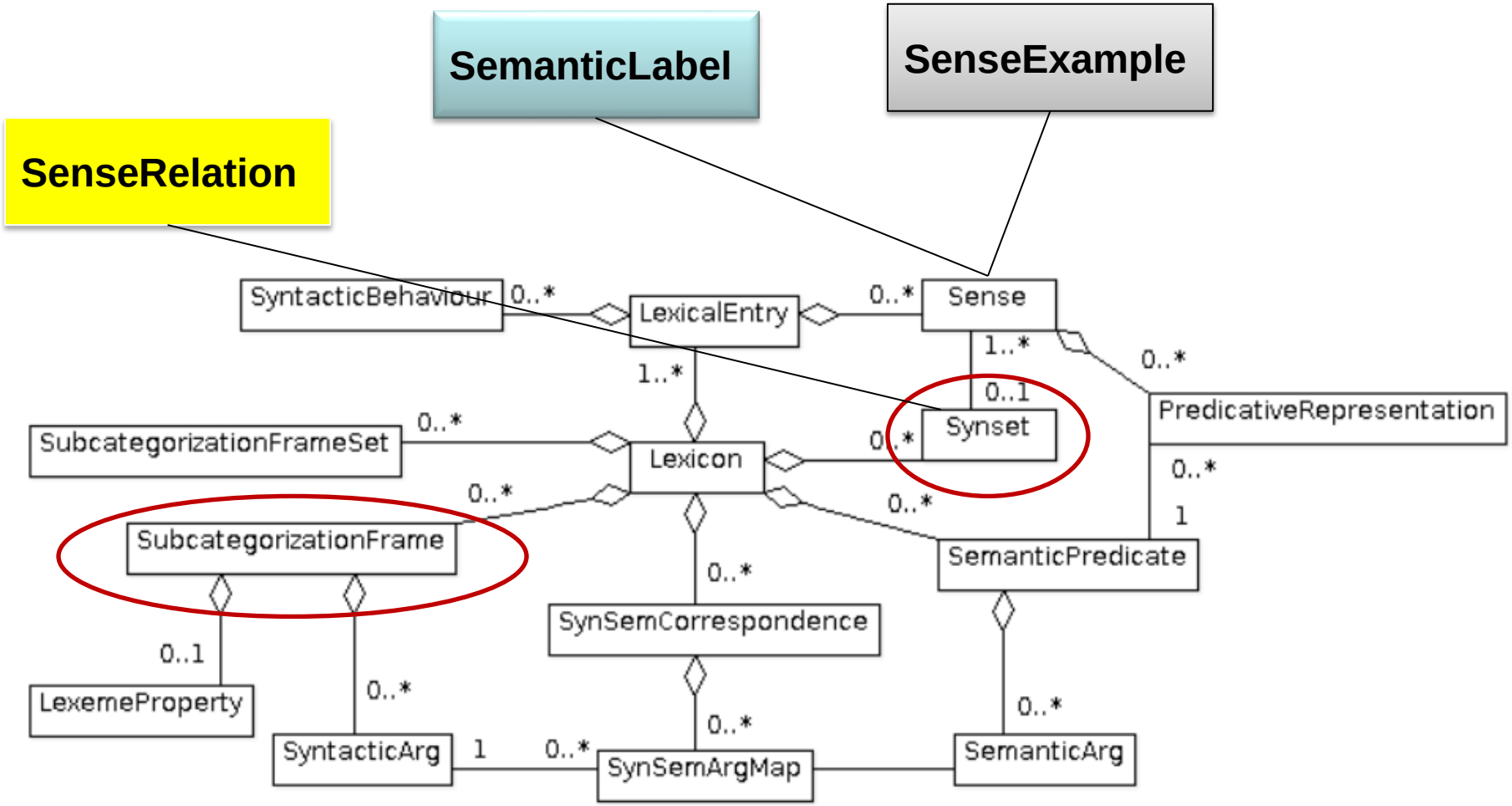
Semantic-Predicate

Information Types in UBY – Sense Links

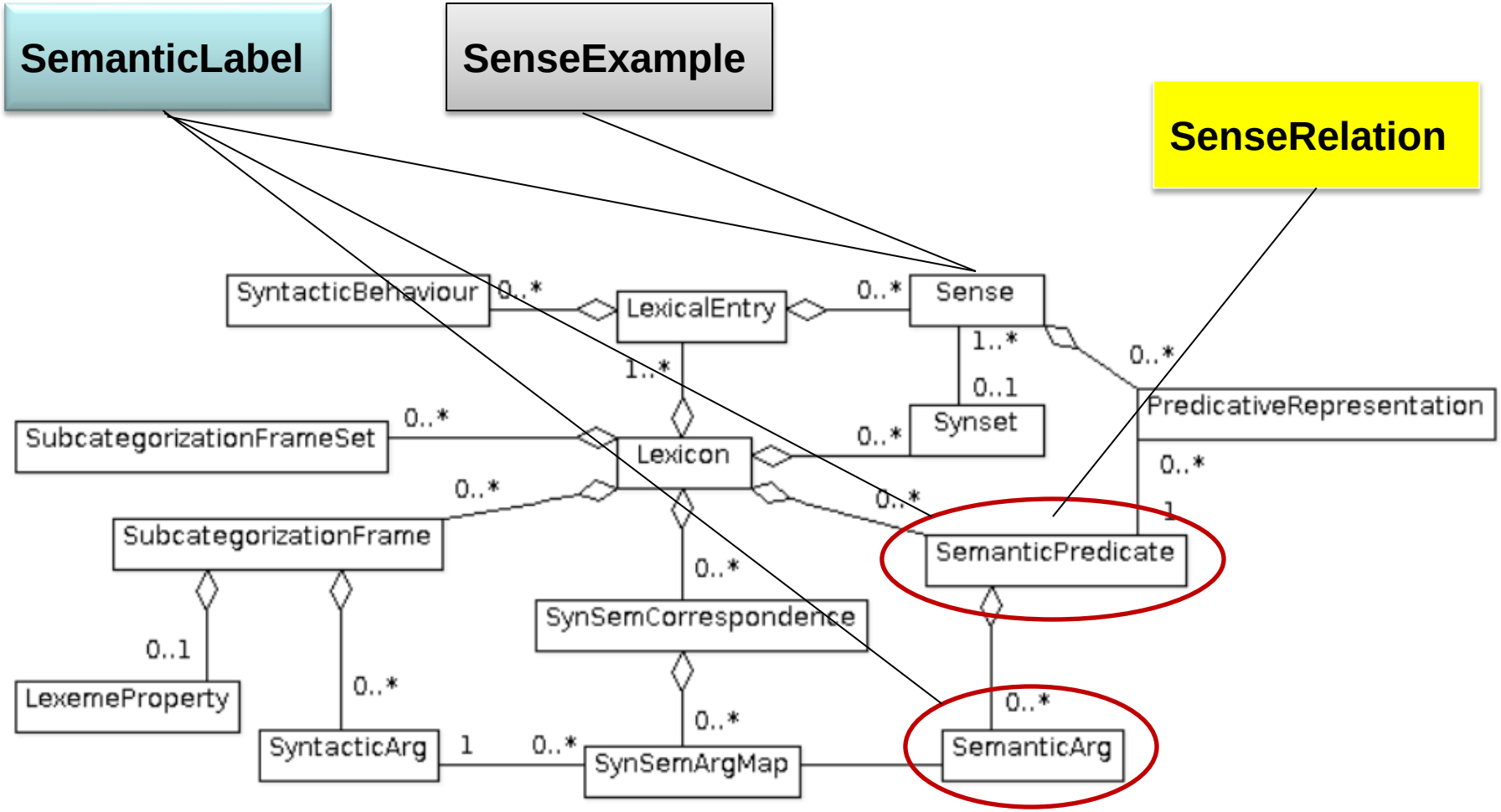
Sense alignments between resource pairs



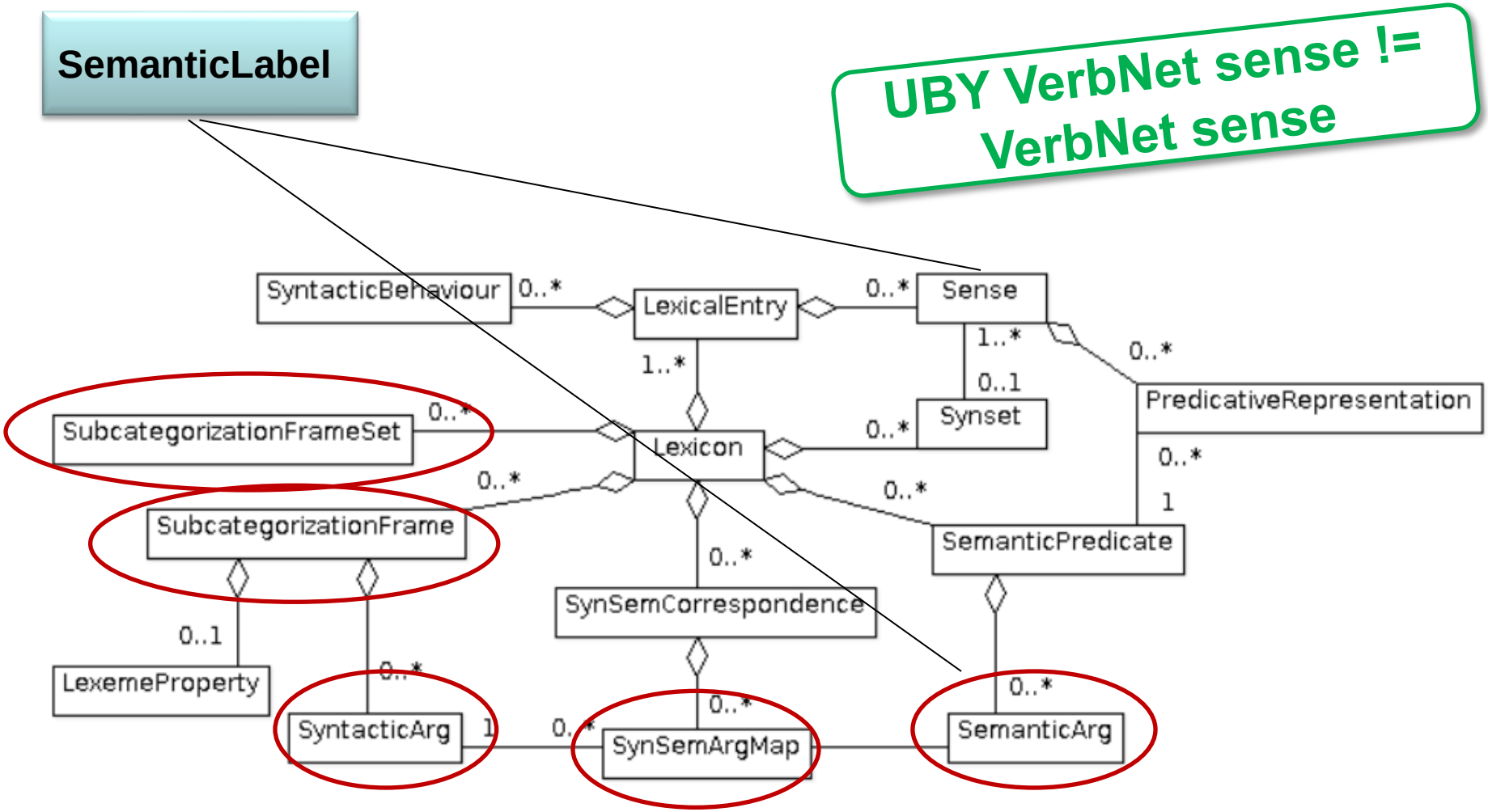
Wordnets in UBY (Main Information Types)



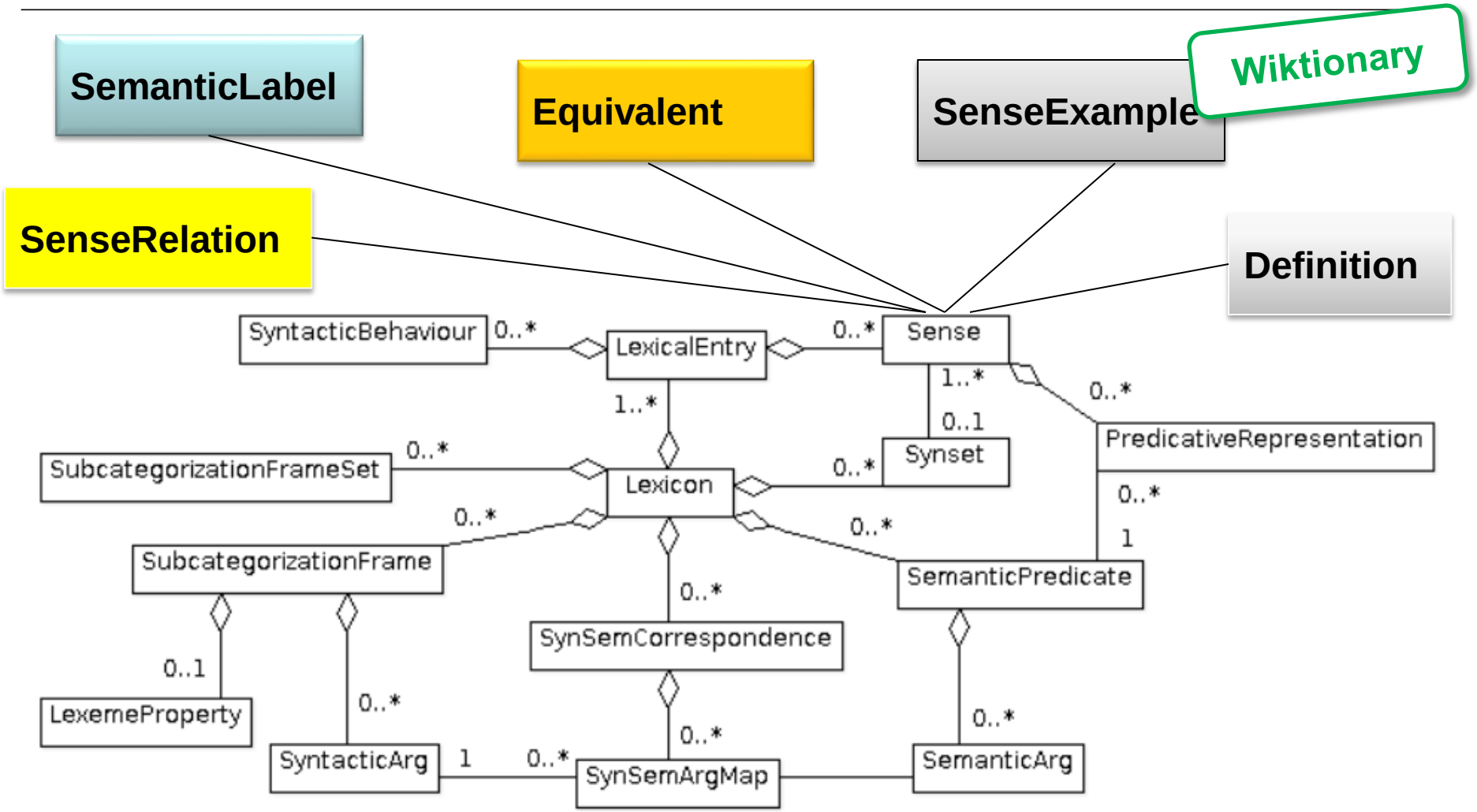
FrameNet in UBY (Main Information Types)



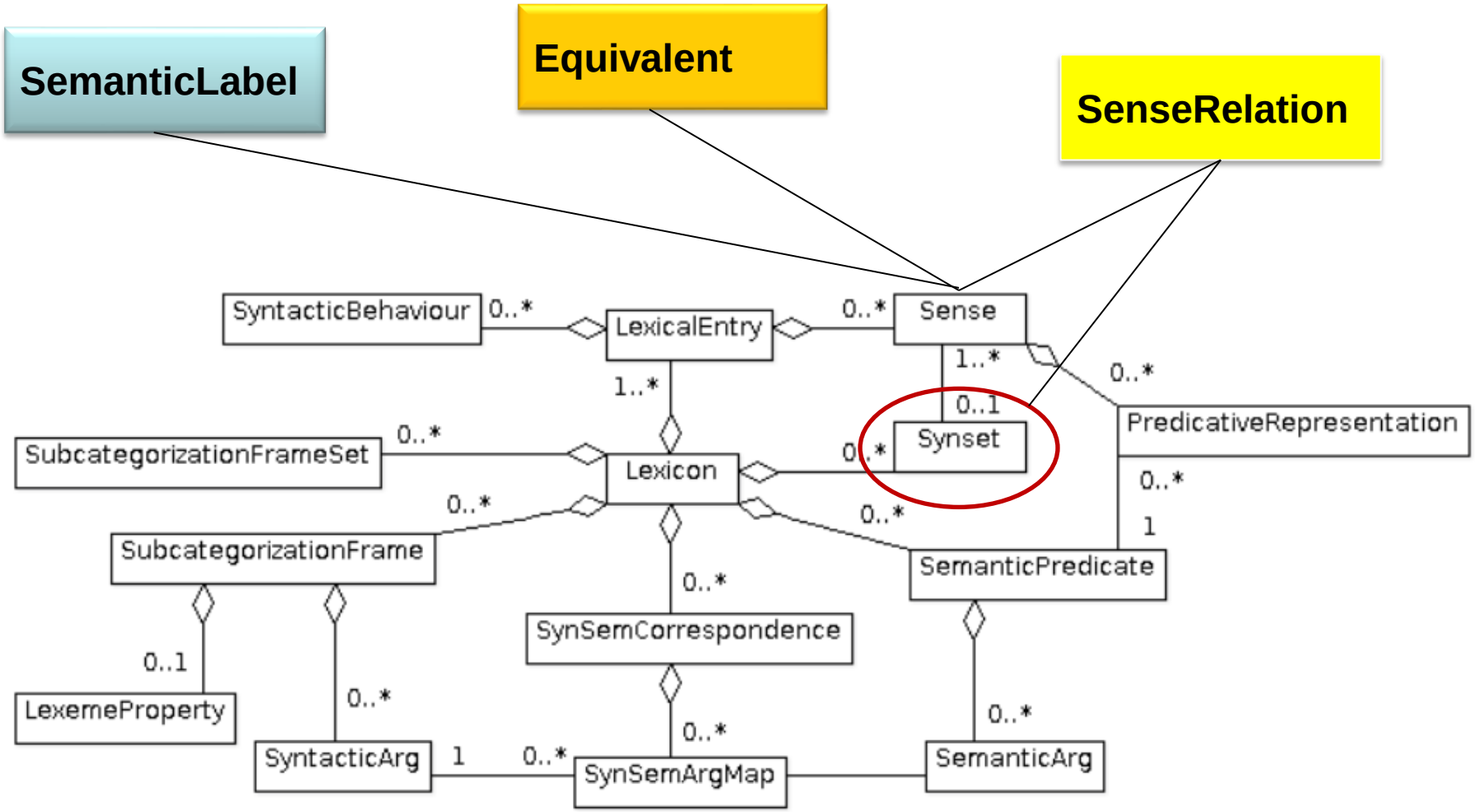
VerbNet in UBY (Main Information Types)



Wiktionary and Wikipedia in UBY (Main Information Types)



OmegaWiki in UBY (Main Information Types)



Mapping of POS for Main Word Classes

Resource	Original POS – UBY POS (`EPartOfSpeech`)
WordNet	noun – noun, verb – verb, adjective – adjective
GermaNet	common noun – nounCommon, proper noun – nounProper, verb – verb, adjective – adjective
FrameNet	noun – noun, verb – verb, adjective – adjective
VerbNet	verb – verb
IMSLex	verb – verbMain, noun – nounCommon, adjective - adjective
WikipediaEN / DE	common noun – noun, proper noun – noun
Wiktionary EN / DE	noun – noun, proper noun – nounProper, verb – verb, adjective – adjective
OmegaWiki	noun – noun, verb – verb, adjective – adjective

Part 2: Recipes

TECHNISCHE UNIVERSITÄT DARMSTADT

Tools and Planning Guide: What UBY has to offer

Mixed Starters: How to query UBY (for WSD)

Appetizer: UBY as UIMA resource

Main Dish: UBY for UIMA-based Semantic Tagging

Dessert: Cross-lingual verb sense linking

Example Code is Open Source:

http://code.google.com/p/dkpro-tutorials/

UBY for Knowledge-Based WSD

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

- The WordNet sense inventory, though standard, has issues, e.g. the **overly fine granularity** of the senses.

Solution

- UBY as a source for new sense inventories
- **richer (coarse-grained) sense inventories**
- sense inventories may depend on domains, tasks, etc.

Goal

- Perform sense tagging using UBY

Joint work with Kostadin Cholakov

UBY for Knowledge-Based WSD – Recipe

TECHNISCHE UNIVERSITÄT DARMSTADT

Occasion

- Sense tagging of text with UBY:
  - to determine which UBY sense occurs in a given text is a **WSD task**

- The sense alignments in UBY help to solve the WSD task, because they enlarge the context (or feature space) for a given UBY word sense
  - e.g., more input for the various types of the Lesk algorithm

Ingredients

- All UBY resources that are aligned

Techniques

- Based on sense linking

Three Ways to Query UBY for WSD

UBY-LMF has been implemented in Java. An object-relational mapping (using Hibernate) allows mapping any instance of UBY-LMF to a SQL database.

The UBY-API **`Uby.java`** provides methods …

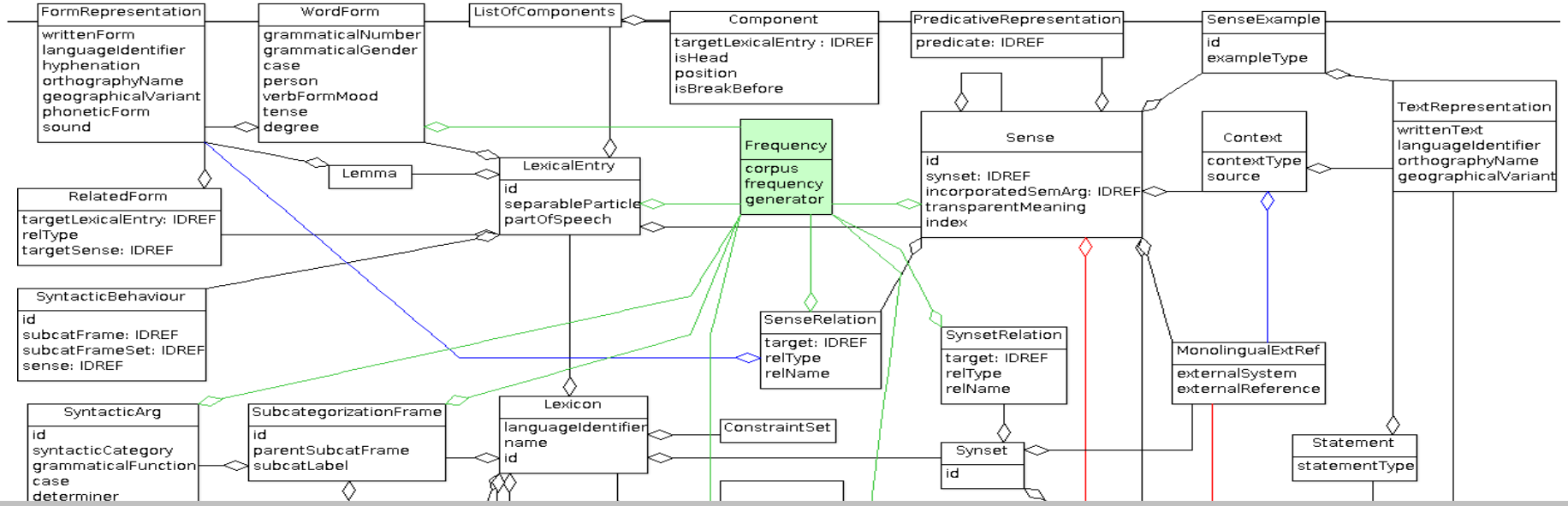
1. to retrieve UBY-LMF objects based on readily available queries
2. that allow you to specify queries yourself
3. Once an UBY-LMF object has been retrieved from the SQL database, the object-relational mapping allows to access all associated objects in the UBY-LMF graph structure.

Retrieving UBY-LMF Objects Based on Readily Available Queries

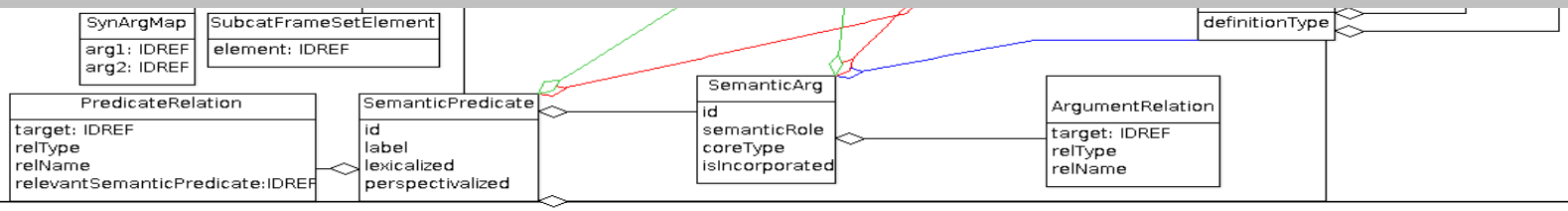
- Example: `getSenseAxisBySense`

```
for (Sense sense : lexEntry.getSenses())

  List<SenseAxis> sas = uby.getSenseAxisBySense(sense);
  for (SenseAxis sa : sas) {

       // DO SOMETHING, SEE CODE EXAMPLES
  }
}
```

UBY API with Hibernate – Based on UBY-LMF



Once an UBY-LMF object has been retrieved from the SQL database, the object-relational mapping allows to access all associated objects in the UBY-LMF graph structure.



Accessing Associated Objects in the UBY-LMF Graph Structure

- Example: translations (Equivalent class)

```
for (Sense sense : lexEntry.getSenses())

  List<Equivalent> eqs = sense.getEquivalents();

  for (Equivalent eq : eqs) {
       System.out.println("- Equivalent: "
       +eq.getLanguageIdentifier()
       +" : "+eq.getWrittenForm());
  }
}
```

Accessing Associated Objects in the UBY-LMF Graph Structure

- Example: sense alignments (SenseAxis class)

backlink – not in UBY-LMF

```
for (SenseAxis sa : sas) {

  if (sa.getSenseOne().getId() matches("WktEN.*")) {
    Sense wktSense = sa.getSenseOne();
    String wktVerb =
       wktSense.getLexicalEntry().getLemmaForm();

  } else if (sa.getSenseTwo().getId().matches("WktEN.*")) {
    Sense wktSense = sa.getSenseTwo();
    String wktVerb =
       wktSense.getLexicalEntry().getLemmaForm();

  }
}
```

Using the UBY API to Specify Queries Yourself

- Criteria API provided by Hibernate

- Example: retrieving the mapping of syntactic and semantic arguments from VerbNet

```
Uby uby = new Uby(dbConfig);
session = uby.getSession();

Criteria criteriaSynSem =
      session.createCriteria(SynSemArgMap.class);
List<SynSemArgMap> SynSemArgMaps =
      criteriaSynSem.list();
for (SynSemArgMap synSem : SynSemArgMaps) {
      SynargSemargMap.put(
            synSem.getSyntacticArgument(),
            synSem.getSemanticArgument());
}
```

Good to Know – Extraction of Bulk Data from UBY

To extract **all instances** of LexicalEntry, Sense, SenseAxis, always use the **`Iterator`** methods.

- Example: `lexicalEntryIterator`
  - Option to filter by POS and Lexicon

```
Iterator<LexicalEntry> lexicalEntryIterator =
      uby.getLexicalEntryIterator(null,lex);

while (lexicalEntryIterator.hasNext()) {
      LexicalEntry le = lexicalEntryIterator.next();
      for (Sense s: le.getSenses()) {
            System.out.println(
                  s.getLexicalEntry().getLemmaForm());
      }
}
```

Part 2: Recipes

TECHNISCHE UNIVERSITÄT DARMSTADT

Tools and Planning Guide: What UBY has to offer

Mixed Starters: How to query UBY (for WSD)

Appetizer: UBY as UIMA resource

Main Dish: UBY for UIMA-based Semantic Tagging

Dessert: Cross-lingual verb sense linking

UBY as UIMA Resource

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

- Using UBY (or a UBY lexicon) as interchangable resource in UIMA-based NLP pipelines

Solution

- Specifying UBY as UIMA resource based on uimaFIT

Goal

- Retrieving information from UBY in UIMA annotators

Joint work with Richard Eckart de Castilho

UBY as UIMA Resource – Recipe

TECHNISCHE UNIVERSITÄT DARMSTADT

Occasion

- UIMA annotator needs to retrieve information from UBY

Ingredients

- UBY
- uimaFIT

Techniques

- Based on standardization

UIMA – Unstructured Information Management Architecture

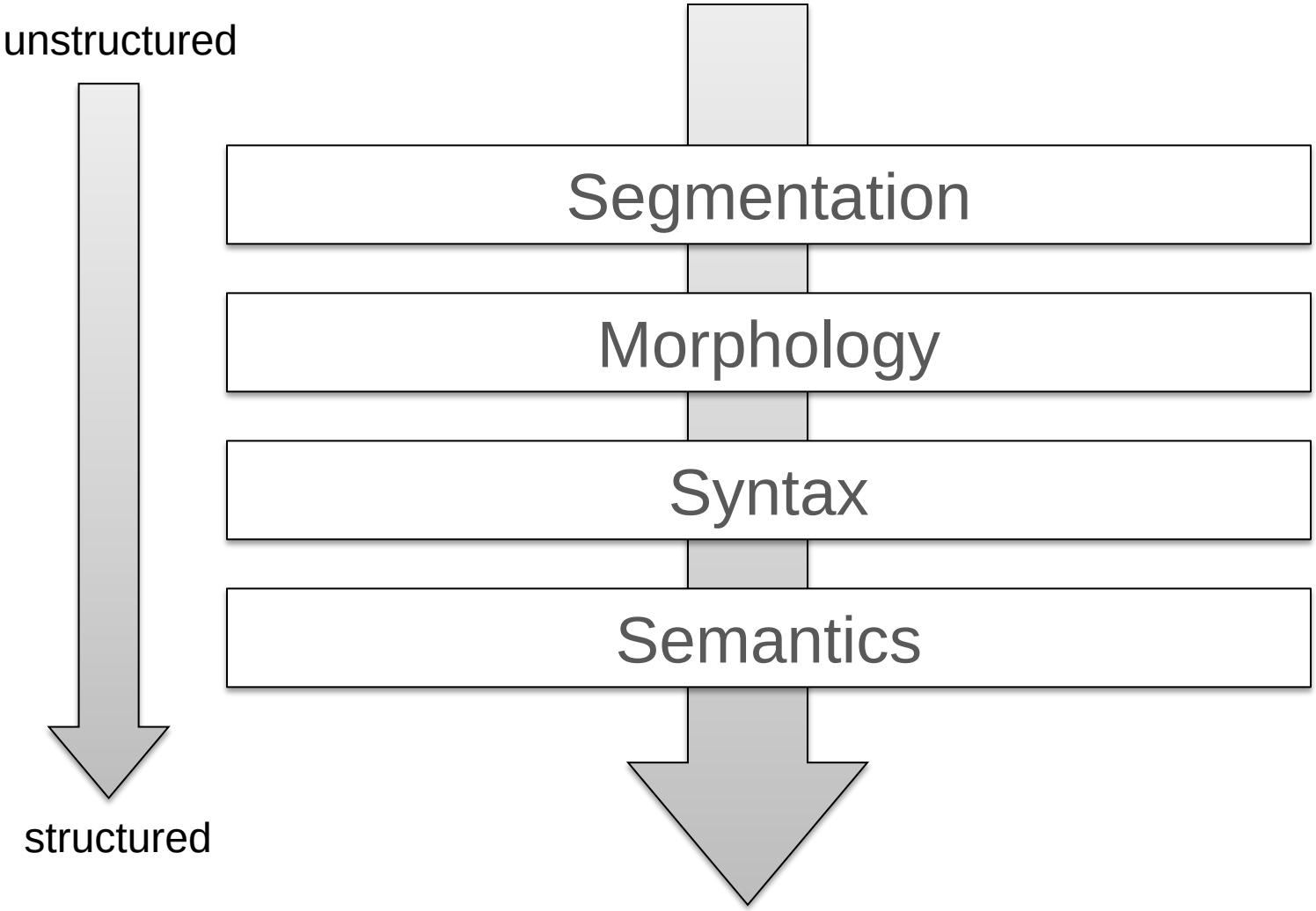
Major goal: transform **unstructured** information to **structured** information

… in order to discover knowledge that is relevant to an end user

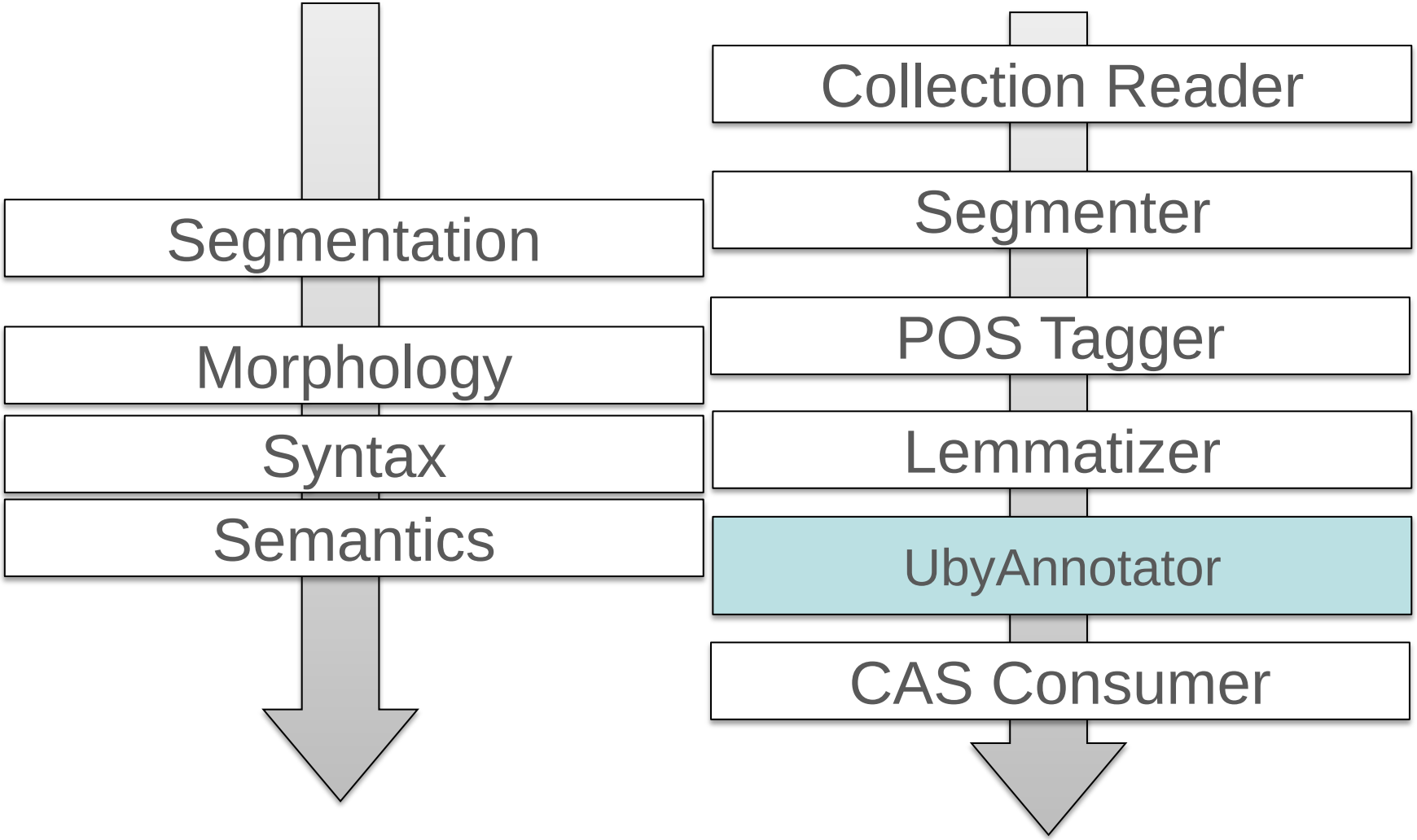
- **Component-based** architecture for analysis of unstructured content like text, video, audio
- Originally developed at IBM – today an Apache project
- Used in commercial as well as educational contexts
  - LanguageWare, Watson (IBM)
  - **uimaFIT** (University of Colorado, now: Apache UIMA uimaFIT™)
  - **DKPro Core** (TU Darmstadt)
  - many more...

How it works: think of UIMA components as machines in an assembly line

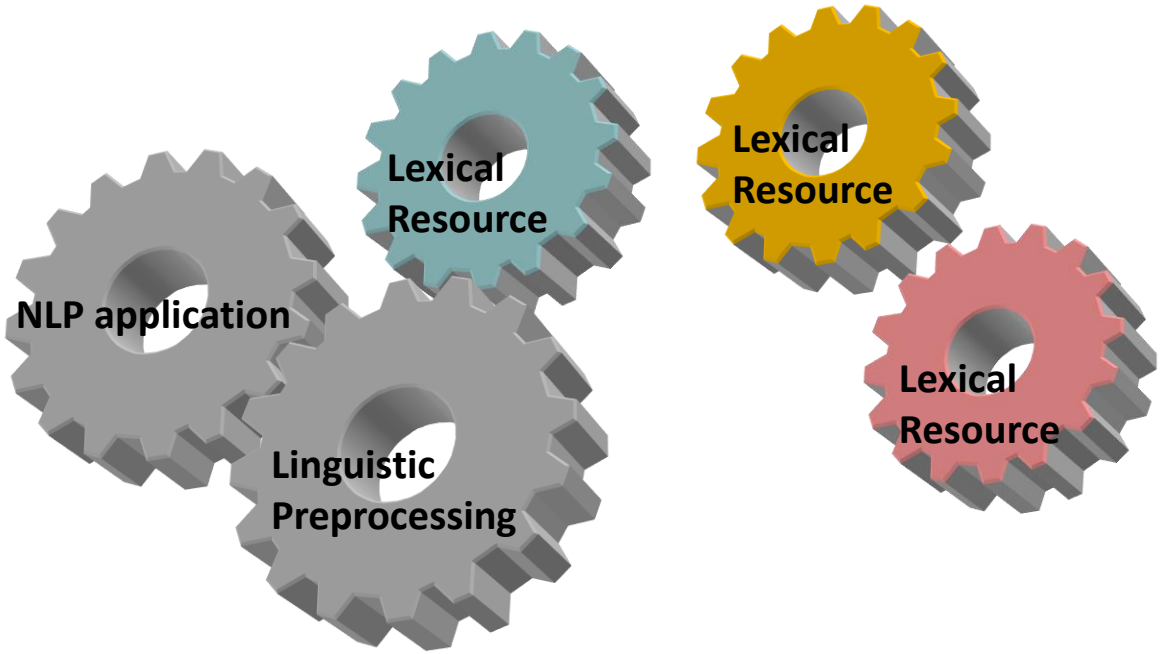
Analysis Levels in Text Processing



UIMA Example Pipeline



Interoperable Lexical Resources



Interoperable lexical resources can be used as UIMA resources that may be shared between UIMA components.

uimaFIT

uimaFIT is an „add-on“ for UIMA, **it simplifies typical development tasks** and allows for **rapid and easy development of NLP processing pipelines.**

http://uima.apache.org/d/uimafit-current/tools.uimafit.book.html

For instance, uimaFIT provides annotations for component configuration:

- `@ConfigurationParameter` annotation
- `@ExternalResource` annotation for (shared) resource management

To specify UBY as UIMA resource based on uimaFIT, we use

- the base class for external resources `Resource_ImplBase`
- the interface `ExternalResourceLocator`
- the `@ExternalResource` annotation

`@ExternalResource` Annotation

uimaFIT provides an annotation-based mechanism for the **configuration of UIMA resources** in annotator components.

Example: `@ExternalResource` annotation used in a UbyAnnotator component

```
public static final String PARAM_UBY_RESOURCE = "uby";
@ExternalResource(key = PARAM_UBY_RESOURCE)
Uby uby;
```

UBY Annotator – Example Use Cases

A UBY annotator can add information from UBY to the CAS in order to enrich the context of a token.

- What can be annotated at the token level?
  - Semantic label information, e.g., semantic field, domain, register
  - Lexical information attached to related senses, e.g., synonyms, hypernyms
  - Translations
- **Example use cases** include
  - processing of text corpora (document collections) in order to perform knowledge-based sense tagging, e.g., using Lesk-based WSD
  - processing of example sentences from lexical resources in order to acquire more fine-grained lexical information, e.g., subcategorization frames

UBY as Resource – Pipeline Configuration

```
AnalysisEngineDescription analysisEngine =
  createEngineDescription(
    UbyAnnotator.class,
    UbyAnnotator.PARAM_UBY_RESOURCE,
    createExternalResourceDescription(
    UbyResourceLocator.class,
    UbyResourceLocator.PARAM_URL, "localhost/uby_open",
    UbyResourceLocator.PARAM_DRIVER, "com.mysql.jdbc.Driver",
    UbyResourceLocator.PARAM_DRIVER_NAME, "mysql",
    UbyResourceLocator.PARAM_USERNAME, "user",
    UbyResourceLocator.PARAM_PASSWORD, "pass"
  )
);
```

Part 2: Recipes

TECHNISCHE UNIVERSITÄT DARMSTADT

Tools and Planning Guide: What UBY has to offer

Mixed Starters: How to query UBY (for WSD)

Appetizer: UBY as UIMA resource

Main Dish: UBY for UIMA-based Semantic Tagging

Dessert: Cross-lingual verb sense linking

Semantic Tagging with UBY

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

- WordNet provides semantic field information that is often used as coarse-grained semantic feature in NLP
- However, the coverage of WordNet is limited
  - This is especially true for the vocabulary (i.e., lemmas) covered: **WordNet lacks vocabulary** for particular domains, e.g., IT domain

Solution

- Exploiting the **increased lemma coverage** offered by UBY

Goal

- Broad-coverage semantic tagging of text with semantic field information

Joint work with Richard Eckart de Castilho

Semantic Tagging with UBY – Recipe

TECHNISCHE UNIVERSITÄT DARMSTADT

Occasion

- Broad-coverage semantic tagging of text with semantic field information

Ingredients

- UBY version of WordNet / GermaNet
- UBY
- DKPro Core and uimaFIT

Techniques

- Based on standardization

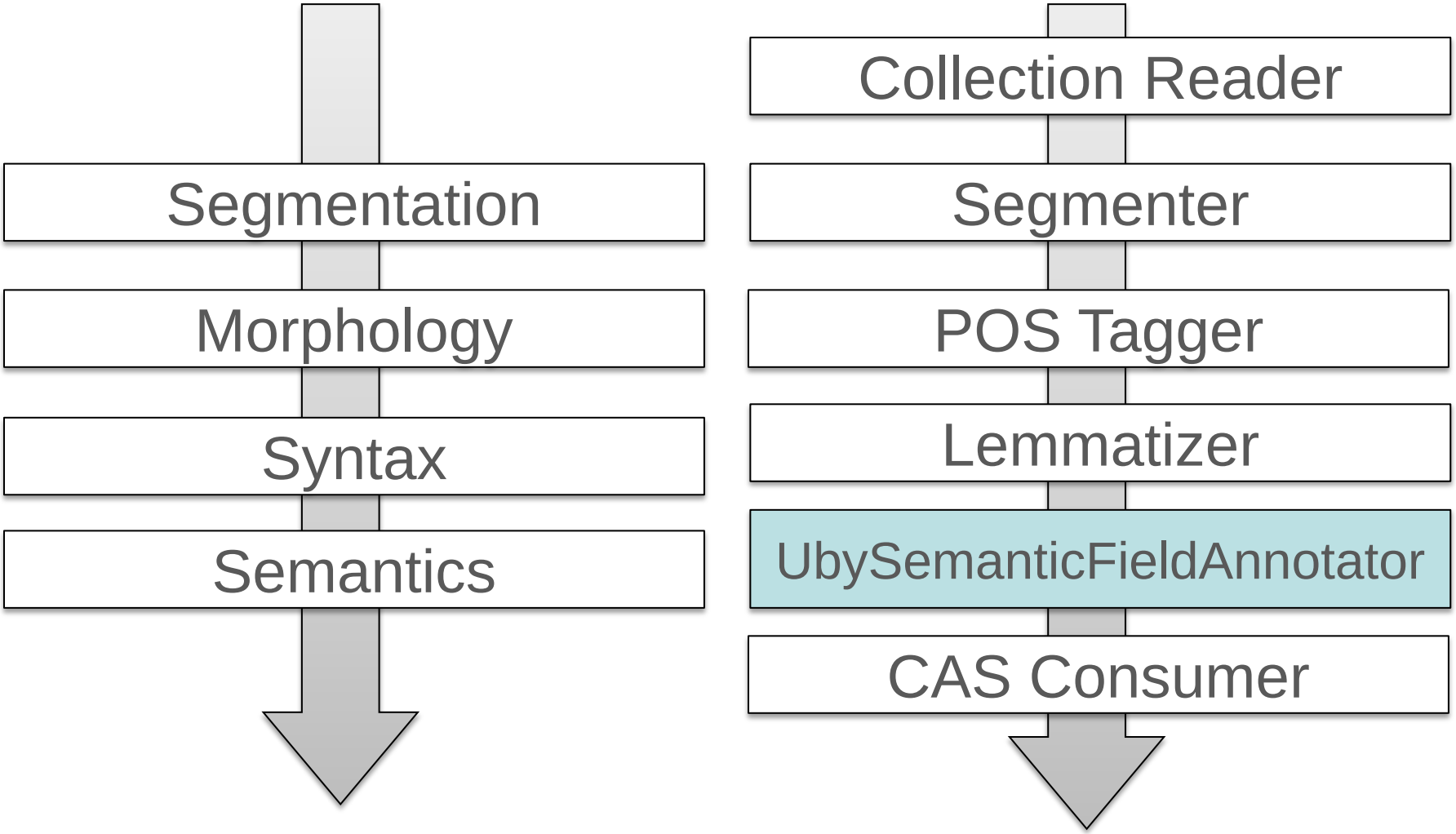
Usage of Semantic Field Information in NLP

WordNet provides coarse-grained **semantic field information** through its **lexicographer file** names.

WordNet semantic field information has widely been used in many NLP tasks and applications, including Information Extraction and text classification.

- Semantic field information from WordNet has also been called *supersenses* and has been applied for *supersense tagging*, e.g.
  - (Ciaramita and Johnson, 2003) and (Qiu et al., 2011) each present a framework for classifying words not in WordNet into the WordNet semantic fields
  - (Ciaramita and Altun, 2006) describe a supersense tagger based on sequence labeling

UIMA Example Pipeline for Semantic Tagging



DKPro Core – Linguistic Annotation of Text

DKPro Core is a collection of software **components for preprocessing and linguistic annotation** of text based on the Apache **UIMA** framework.

http://code.google.com/p/dkpro-core-asl/

DKPro Core builds on uimaFIT.

Integrated Tools

- TreeTagger
- OpenNLP
- Stanford NLP
- JWordSplitter
- Language Tool
- MaltParser
- …

Supported Formats

- Text
- PDF
- TEI XML, BNC XML
- SQL Databases
- Google web1t n-grams
- …

Interface `SemanticTagProvider`

- Interface to create various UIMA resources that provide semantic tags: `SemanticTagProvider`

- The interface `SemanticTagProvider` can be used to retrieve semantic tag information from
  - simple key value files
  - UBY
  - many other (linked) lexical resources, such as BabelNet, UWN, …

`SemanticTagProvider` offers the method `getSemanticTag`:

```
String getSemanticTag(Token token)
```

UIMA Resource `UbySemanticFieldResource`

`UbySemanticFieldResource`

- extends `Resource_ImplBase` and implements `SemanticTagProvider`
- follows the first sense (WordNet) / random sense (other UBY resources) heuristic

How to use `UbySemanticFieldResource` in annotator components:

```
public static final String PARAM_UBY_SEMANTIC_FIELD_RESOURCE =
       "ubySemanticFieldResource";

@ExternalResource(key = PARAM_UBY_SEMANTIC_FIELD_RESOURCE)
private UbySemanticFieldResource ubySemanticFieldResource;
```

UIMA Annotator
UbySemanticFieldAnnotator

How to use `UbySemanticFieldAnnotator` in a pipeline:

```
AnalysisEngineDescription processor = createEngineDescription(
 createEngineDescription(
  UbySemanticFieldAnnotator.class,
  UbySemanticFieldAnnotator.PARAM_UBY_SEMANTIC_FIELD_RESOURCE,
  createExternalResourceDescription(
  UbySemanticFieldResource.class,
  UbySemanticFieldResource.PARAM_URL, "localhost/uby_open",
  UbySemanticFieldResource.PARAM_DRIVER, "com.mysql.jdbc.Driver",
  UbySemanticFieldResource.PARAM_DRIVER_NAME, "mysql",
  UbySemanticFieldResource.PARAM_USERNAME, "root",
  UbySemanticFieldResource.PARAM_PASSWORD, "pass"
  )
 )
);
```

Part 2: Recipes

TECHNISCHE UNIVERSITÄT DARMSTADT

Tools and Planning Guide: What UBY has to offer

Mixed Starters: How to query UBY (for WSD)

Appetizer: UBY as UIMA resource

Main Dish: UBY for UIMA-based Semantic Tagging

Dessert: Cross-lingual verb sense linking

Cross-lingual Verb Sense Linking

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

- German resources that are publicly available for research **lack semantic role and selectional preference information** for verbs.

Solution

- Exploiting the **standardized format for subcategorization frames** (SCFs) in English (EN) and German (DE) provided by UBY-LMF in order to perform a cross-lingual linking of verb senses and their SCFs

Goal

- Enriching verb senses in the German resources GermaNet and IMSLex

Cross-lingual Verb Sense Linking – Recipe

TECHNISCHE UNIVERSITÄT DARMSTADT

Occasion

- Enriching verb senses in GermaNet and IMSLex with semantic role and selectional preference information
- Using this information, e.g., in WSD

Ingredients

- UBY versions of IMSLex, GermaNet, WordNet, VerbNet

Techniques

- Based on sense linking and standardization

Standardized Format for SCFs in English (EN) and German (DE)

TECHNISCHE UNIVERSITÄT DARMSTADT

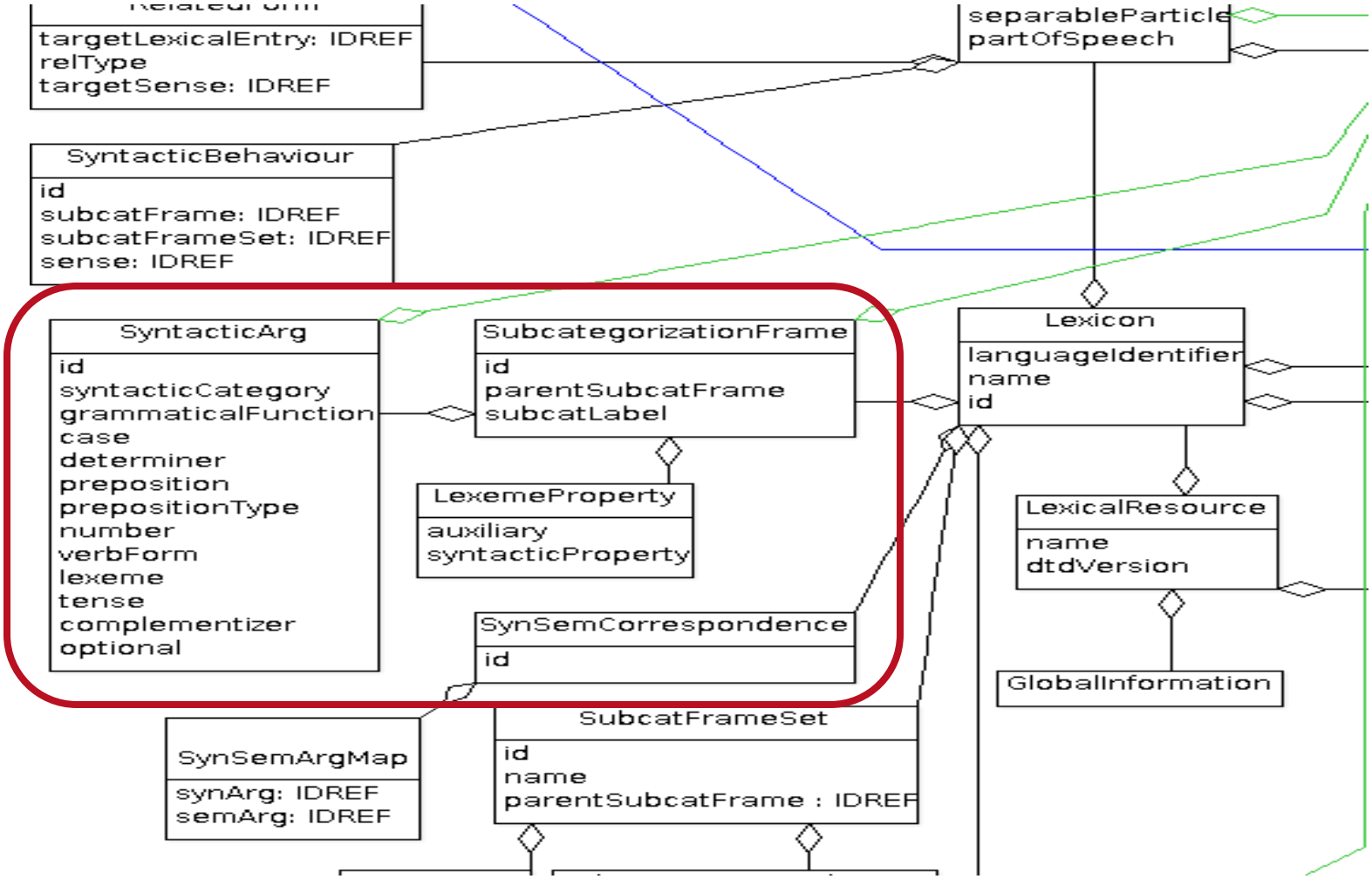
Requirement: flexible format for SCFs in EN and DE, applicable both across languages and at the language-specific level

- Across languages: in order to automatically detect correpondences
- Language-specific: preserve distinctions between fine-grained EN and DE SCFs in order to automatically identify them as well

UBY-LMF – based on ISO-Standards – meets this requirement

- ISO 24613:2008 Lexical Markup Framework (LMF)
- ISO 12620:2009 Data Category Registry: ISOcat

SCFs in UBY-LMF – Specification of Syntactic Arguments



SCFs in UBY-LMF – Example

lachen

NN.Pp

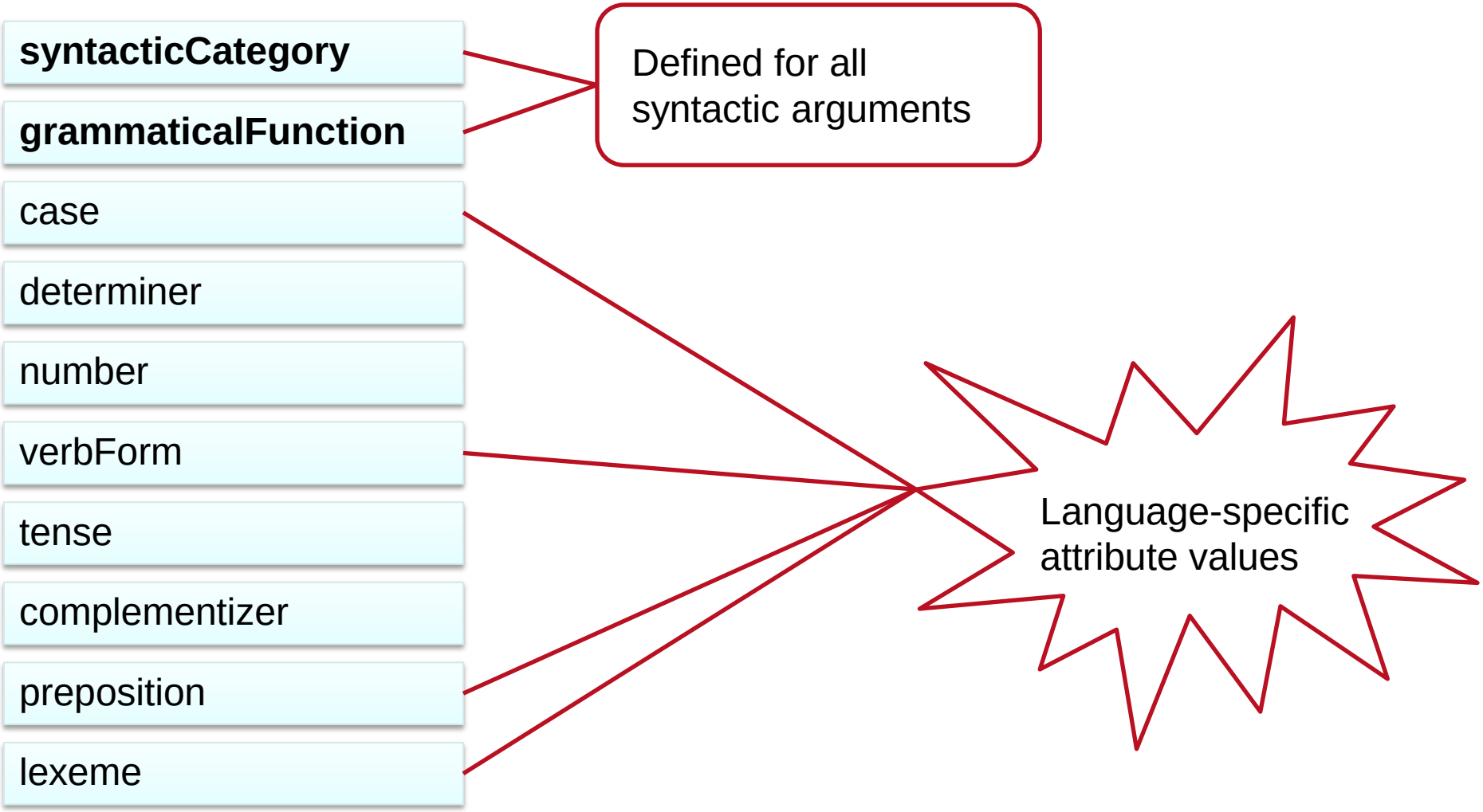
laugh

NP V PP {about}

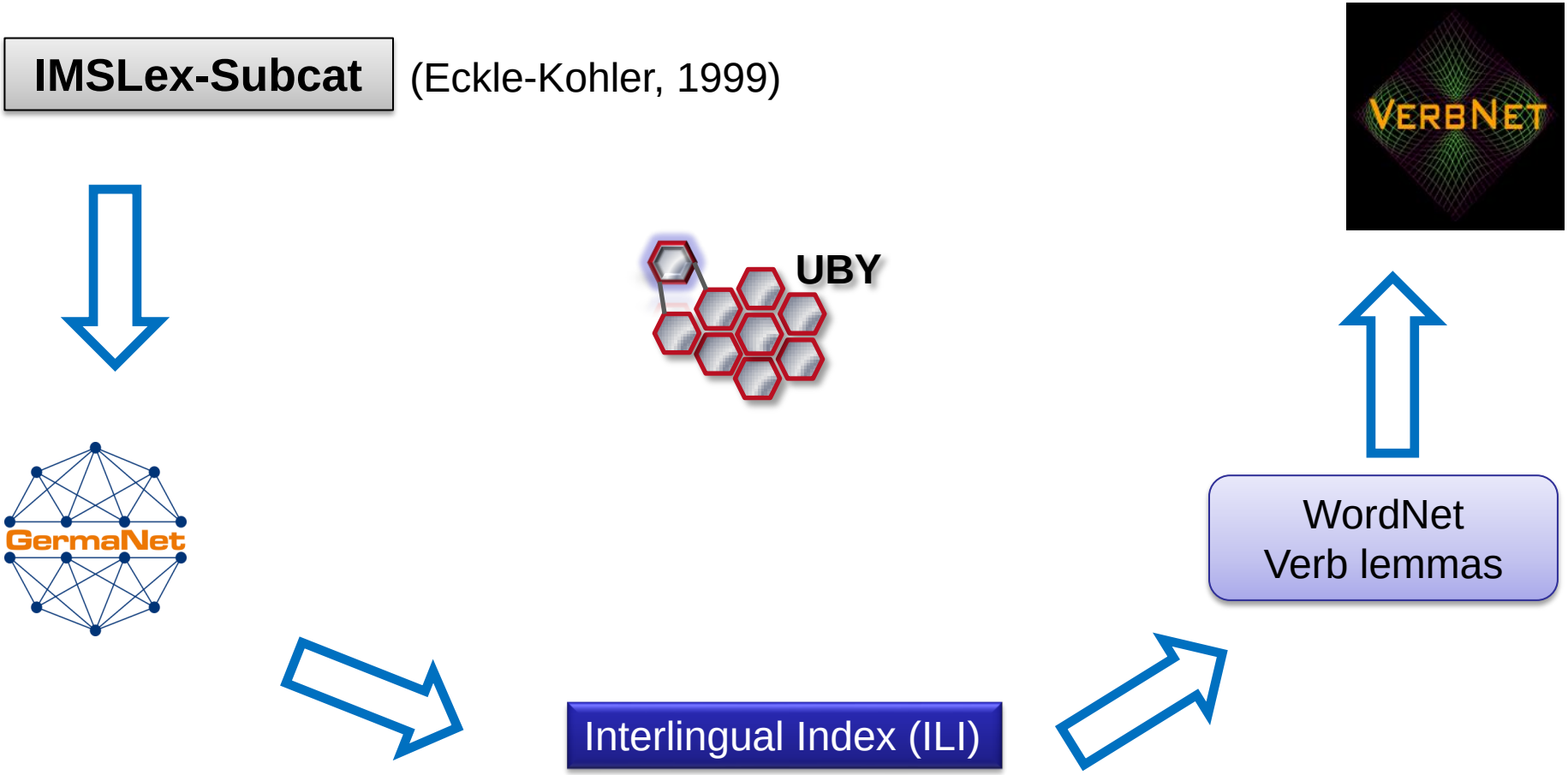
UBY

GermaNet Syntactic Arguments	**VerbNet Syntactic Arguments**
syntacticCategory: nounPhrase grammaticalFunction: subject	syntacticCategory: nounPhrase grammaticalFunction: subject
syntacticCategory: prepositionalPhrase	syntacticCategory: prepositionalPhrase
grammaticalFunction: prepositionalComplement	grammaticalFunction: prepositionalComplement
Preposition: -	Preposition: about

Syntactic Arguments in UBY-LMF: Attributes



Cross-lingual Linking of Verb Senses



Cross-lingual Linking of Verb Senses – Case Study

TECHNISCHE UNIVERSITÄT DARMSTADT

IMSLex-Subcat

UBY

VERBNET

Verbs from IMSLex-Subcat which can be used with zu-infinitive or complement clause: **784 verbs**

GermaNet

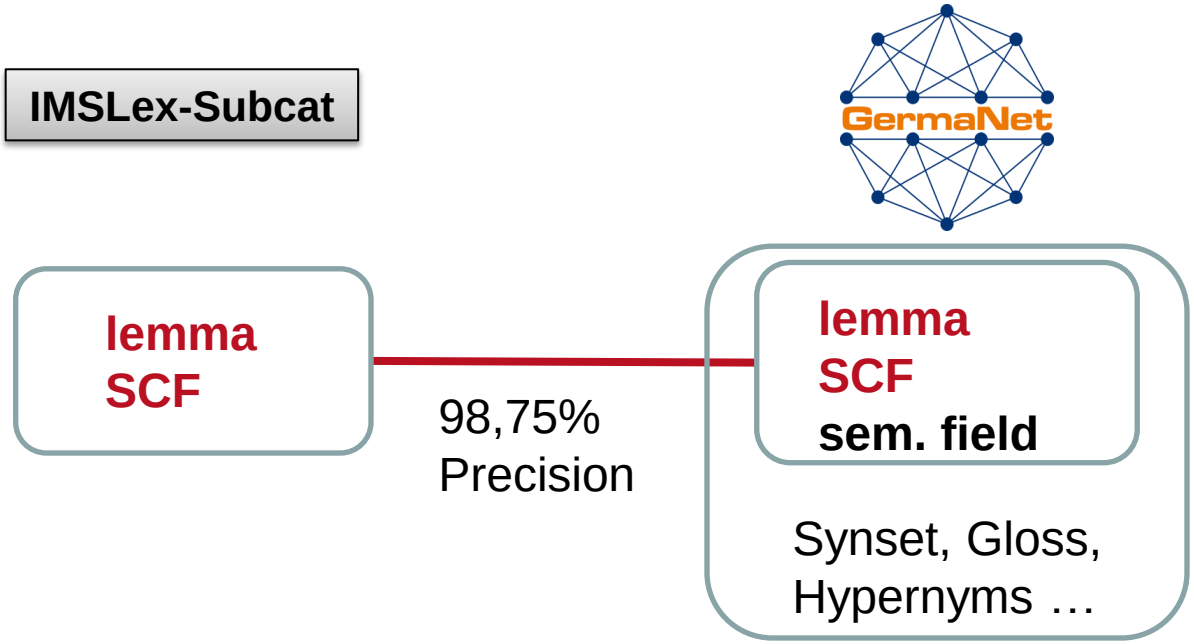
WordNet
Verb lemmas

Interlingual Index (ILI)

How to Exploit UBY-LMF Compliant SCFs: Linking of IMSLex-Subcat and GermaNet

TECHNISCHE UNIVERSITÄT DARMSTADT

- **syntacticCategory**
- **gramm.Function**
- case
- determiner
- number
- verbForm
- tense
- complementizer
- preposition
- lexeme



- Same number of syntactic arguments
- syntacticCategory , case und complementizer values are compatible

→ **lexicon entries are linked**

44 (out of 784) IMSLex verbs do not occur in GermaNet

How to Exploit UBY-LMF Compliant SCFs: Linking of IMSLex-Subcat und VerbNet

TECHNISCHE UNIVERSITÄT DARMSTADT

syntacticCategory

gramm.Function

case

determiner

number

verbForm

tense

complementizer

preposition

lexeme

IMSLex-Subcat

VERBNET

lemma
SCF
lemma EN

72,68% Precision

lemma
SCF
VerbNet class

Semantic roles,
Selectional preferences …

- Same number of syntactic arguments
- syntacticCategory values are compatible

→ **lexicon entries are linked**

- Based on translation via ILI

(140 WordNet verbs do not occur in VerbNet)

Future Work: Cross-lingual Verb Sense Linking on the Semantic Web

lemonUby is the Semantic Web version of UBY

http://lemon-model.net/lexica/uby/

- *lemon*: lexicon model for the Semantic Web
- mapping from UBY-LMF to *lemon*

lemonUby

- comprises a subset of UBY resources
- *lemonUby* is linked with other language resources in the Linguistic Linked Open Data (LLOD) cloud:
  - lexical resources: WordNet 3.0, WordNet 2.0, Wiktionary
  - linguistic terms in *lemonUby* are linked to the Ontologies of Linguistic Annotations (OLiA) (Chiarcos, 2012)

Take Home Messages

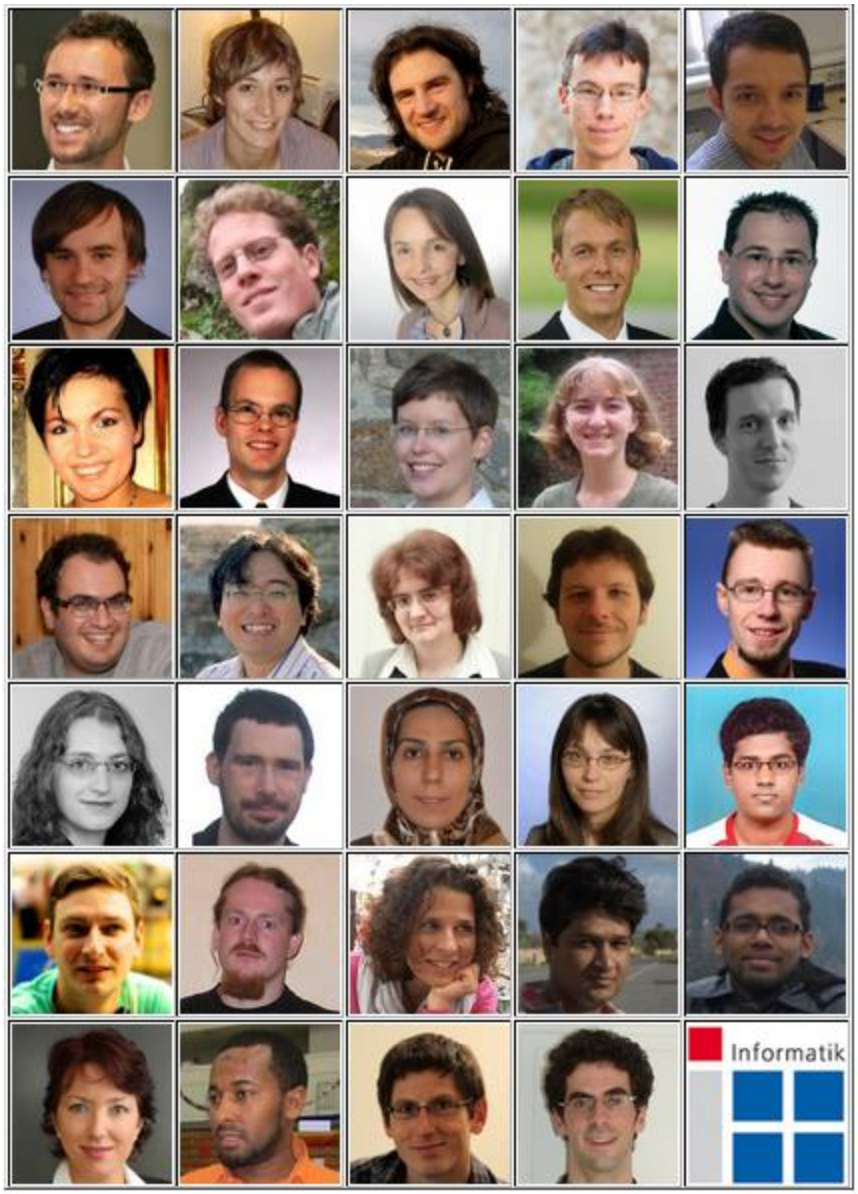
Automatically linking lexical resources at the sense level

- is a difficult NLP task that involves WSD
- benefits from lexical resource standardization

Many linked lexical resources exist and their potential for NLP is waiting to be tapped.

Unique features of UBY are

- the wide variety of lexical resources it integrates
- its ISO-compliant standardized format



Thank You!

TECHNISCHE UNIVERSITÄT DARMSTADT

Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt

KTS
Klaus Tschira Stiftung gemeinnützige GmbH

VolkswagenStiftung

Deutsche Forschungsgemeinschaft DFG

LOEWE

Bundesministerium für Wirtschaft und Technologie

Emmy Noether-Programm
Deutsche Forschungsgemeinschaft DFG

UKP

Selected References – Classic Lexical Resources

WordNet, GermaNet

- Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, USA.
- Claudia Kunze and Lothar Lemnitzer. 2002. GermaNet – representation, visualization, application. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), pages 1485–1491, Las Palmas, Canary Islands, Spain.

FrameNet

- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98, pages 86–90, Montreal, Canada.
- Charles J. Fillmore. 1982. Frame Semantics. In The Linguistic Society of Korea, editor, Linguistics in the Morning Calm, pages 111–137. Hanshin Publishing Company, Seoul, Korea.

VerbNet

- Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A Large-scale Classification of English Verbs. Language Resources and Evaluation, 42:21–40.

IMSLex-Subcat

- Judith Eckle-Kohler. 1999. Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora, Logos, Berlin.

Selected References – Collaborative Resources

Wikipedia, Wiktionary, OmegaWiki

- Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, (7):154–165.
- Christian M. Meyer and Iryna Gurevych. To Exhibit is not to Loiter: A Multilingual, Sense-Disambiguated Wiktionary for Measuring Verb Similarity, in: Proceedings of the 24th International Conference on Computational Linguistics (COLING), Vol. 4, p. 1763–1780, December 2012. Mumbai, India.
- Michael Matuschek , Christian M. Meyer and Iryna Gurevych. Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Translation Applications. In: Translation: Corpora, Computation, Cognition (TC3), vol. 3, no. 1, p. 87-118, July 2013. ISSN 2193-6986.
- Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), pages 1646–1652, Marrakech, Morocco.

Selected References – Linked Lexical Resources I

- Jordi Atserias, Luıs Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. 2004. The Meaning Multilingual Central Repository. In Proceedings of the second international WordNet Conference (GWC 2004), pages 23–30, Brno, Czech Republic.
- Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), pages 513–522, New York, NY, USA.
- Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer and Christian Wirth. UBY - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), p. 580--590, April 2012.
- Egoitz Laparra and German Rigau. 2010. eXtended WordFrameNet. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pages 1214–1419, Valletta, Malta.
- Clifton J. McFate and Kenneth D. Forbus. 2011. NULEX: an open-license broad coverage lexicon. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, HLT '11, pages 363–367, Portland, OR, USA.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 216–225, Uppsala, Sweden.

Selected References – Linked Lexical Resources II

- Roberto Navigli and Simone Paolo Ponzetto. 2010. Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1522–1531, Uppsala, Sweden.
- Martha Palmer. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In Proceedings of the Generative Lexicon Conference (GenLex-09), pages 9–15, Pisa, Italy.
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In Proceedings of the 16th International Conference on World Wide Web, pages 697–706, Banff, Canada.

Selected References – Linking Lexical Resources I

- 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), p. 883–892, November 2011. Chiang Mai, Thailand.
- Oscar Ferrandez, Michael Ellsworth, Rafael Munoz, and Collin F. Baker. 2010. Aligning FrameNet and WordNet based on Semantic Neighborhoods. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pages 310–314, Valletta, Malta.
- Judith Eckle-Kohler, John P. McCrae and Christian Chiarcos. lemonUBY- a large, interlinked, syntactically-rich resource for ontologies, in: Semantic Web Journal (submitted).
- Silvana Hartmann and Iryna Gurevych. FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), vol. 1, p. 1363-1373, Association for Computational Linguistics, August 2013.
- Egoitz Laparra and German Rigau. 2009. Integrating WordNet and FrameNet using a Knowledge-based
- Word Sense Disambiguation Algorithm. In Proceedings of the International Conference RANLP-2009, pages 208–213, Borovets, Bulgaria.
- Michael Matuschek and Iryna Gurevych. Dijkstra-WSA: A Graph-Based Approach to Word Sense Alignment. In: Transactions of the Association for Computational Linguistics (TACL), vol. 1, p. 151-164, May 2013.

Selected References – Linking Lexical Resources II

- Elisabeth (geb. Wolf) Niemann and Iryna Gurevych. 2011. The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. In Proceedings of the International Conference on Computational Semantics (IWCS), pages 205–214, Singapore.
- Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In Computational Linguistics and Intelligent Text Processing, pages 100–111. Springer, Berlin Heidelberg.
- Sara Tonelli and Daniele Pighin. 2009. New Features for FrameNet - WordNet Mapping. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), pages 219 - 227, Boulder, CO, USA.

Selected References – Standardizing Lexical Resources I

- Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A Data Category Registry- and Component-based Metadata Framework. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), pages 43–47, Valletta, Malta.
- Christian Chiarcos. 2012. Ontologies of Linguistic Annotation: Survey and Perspectives, in: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), pp 303-310, May 2012. Istanbul, Turkey.
- Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek and Christian M. Meyer. UBY-LMF - Exploring the Boundaries of Language-Independent Lexicon Models. In: Gil Francopoulo: LMF Lexical Markup Framework, chap. 10, p. 145-156, ISTE - HERMES - Wiley, 2013. ISBN 978 184 821 4309.
- Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek and Christian M. Meyer. UBY-LMF - A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In: Nicoletta Calzolari and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), p. 275--282, May 2012.
- Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pages 233–236, Genoa, Italy.
- Gil Francopoulo (editor) LMF Lexical Markup Framework, ISTE / Wiley 2013 (ISBN 978-1-84821-430-9).

Selected References – Standardizing Lexical Resources II

- John McCrae, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, Tobias Wunner. (2012) Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation 46:701–719.

Selected References – Semantic Tagging

- Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In Proceedings of the 2003 conference on Empirical methods in natural language processing (EMNLP '03).
- Massimiliano Ciaramita and Yasemin Altun (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In Proceedings of the 2006 Conference on Empirical Methodsin Natural Language Processing, pages 594{602, Sydney, Australia.
- Likun Qiu, Yunfang Wu, and Yanqiu Shao. 2011. Combining contextual and structural information for supersense tagging of chinese unknown words. In Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I (CICLing'11), Alexander Gelbukh (Ed.), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, 15-28.