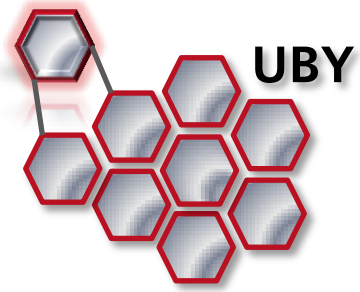
Lexical Resources for Natural Language Processing

Christian M. Meyer and **Hatem Mousselly Sergieh**

TECHNISCHE UNIVERSITÄT DARMSTADT



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

OntoWiktionary

SALSA II

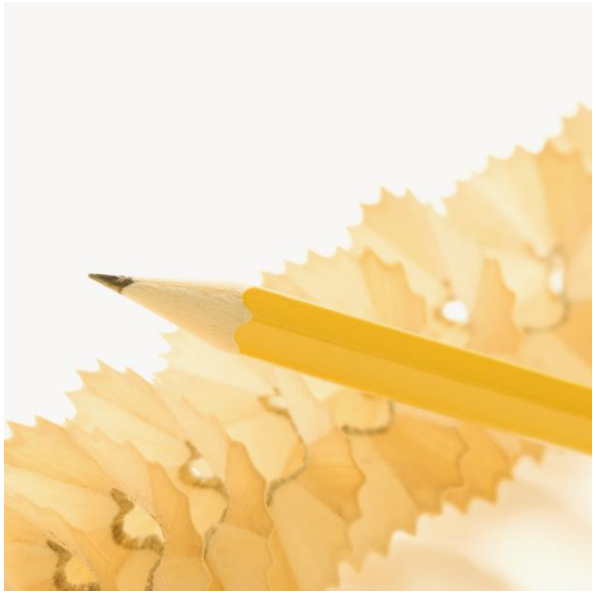
IMSLex-Subcat

Try it yourself! – Preparation

TECHNISCHE UNIVERSITÄT DARMSTADT

You need a JDK >= 6 and a Maven-ready IDE

- Download the UBY 0.7.0 h2 database and the corresponding code snippets from: http://uby.ukp.informatik.tu-darmstadt.de/uby/gscl2015/
- Unzip everything
- Import the demo source files in your workspace
- Put the h2 database in the embeddedUby folder of your project folder
- Optional: Download the tutorial slides as well



Alternative: https://uby.ukp.informatik.tu-darmstadt.de/uby-browser/

Lexical Resources for NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

Introduction

Dictionaries — Try it!

Wordnets and Thesauri

Multilingual and Aligned Resources — Try it!

– Break –

Deep Semantic Resources

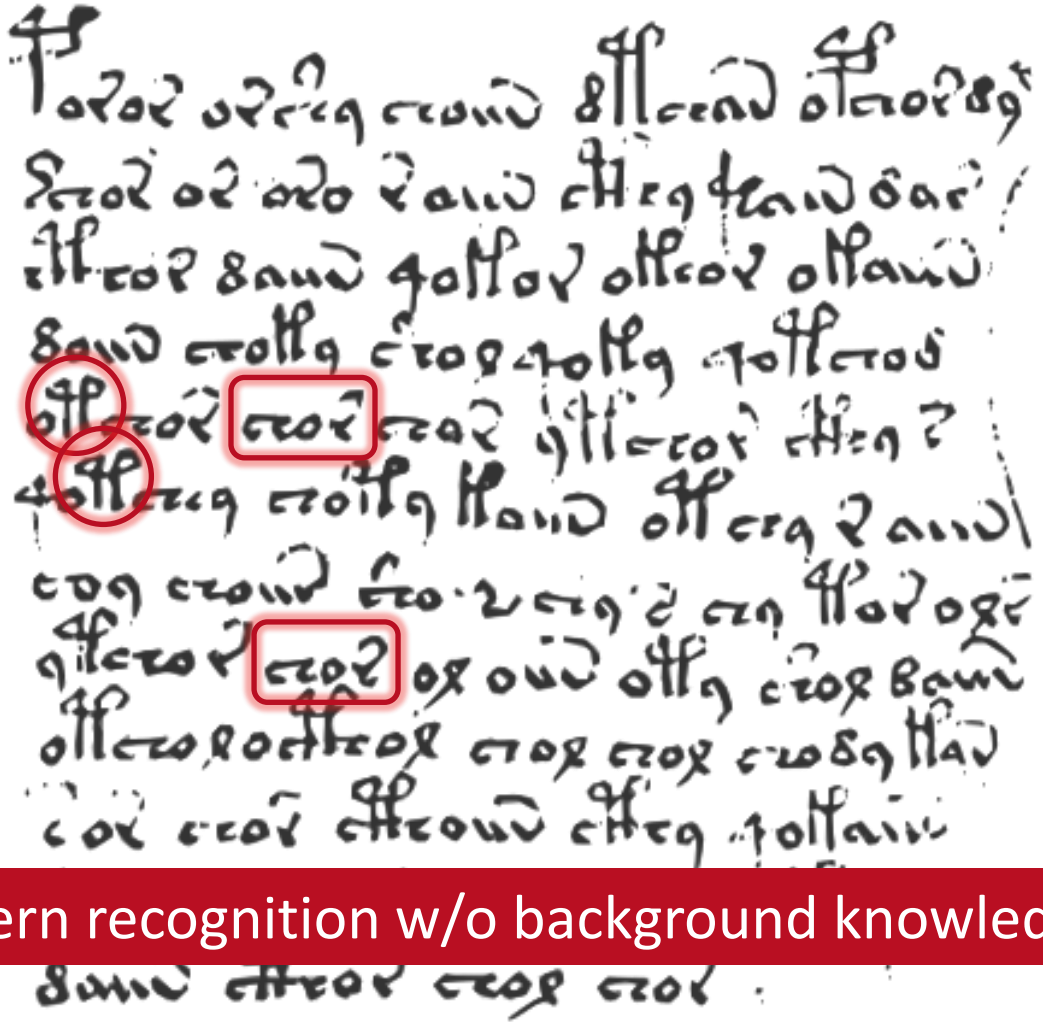
Syntactic Resources — Try it!

Lexical Resources in Action

Wrap-up — Try it!

http://pixabay.com/en/coffee-cup-drink-cafe-brown-mug-312521/ (CC0)

Knowledge-Poor Approaches



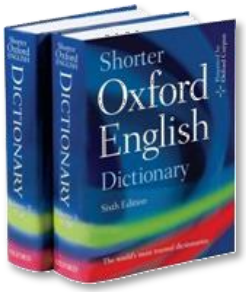
→ Pattern recognition w/o background knowledge

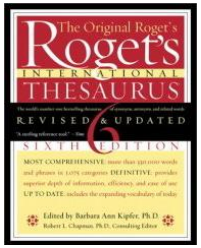
http://de.wikipedia.org/wiki/Voynich-Manuskript#mediaviewer/File:Voynich_manuscript_excerpt.svg (05.12.2009)

Knowledge-Rich Approaches

Background Knowledge

TECHNISCHE UNIVERSITÄT DARMSTADT



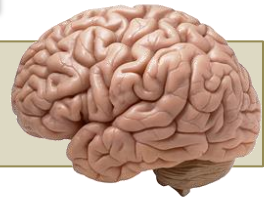


Dictionaries **Encyclopedias** **Thesauri** **Wordnets** **Many others…**

Lexical resources

Corpora

Intuition



https://www.flickr.com/photos/flamephoenix1991/8376271918 (CC-BY-SA 2.0; Flickr user _DJ_)

UKP

"Who cares about lexical resources; we have corpora!"

Lexical Resources and Corpora

TECHNISCHE UNIVERSITÄT DARMSTADT

Corpus

- Collected from real-world text and speech
- Contains multiple occurrences of a lemma
- Frequent phenomena occur more often
- Shows how language is used
- Provides typical contexts and frequencies

Lexical Resource

- Derived from corpora (aggregated view)
- A lemma usually occurs only once
- Rare & frequent phenomena are treated equally
- Describes how language is used
- Provides meta information (e.g., sense definition)

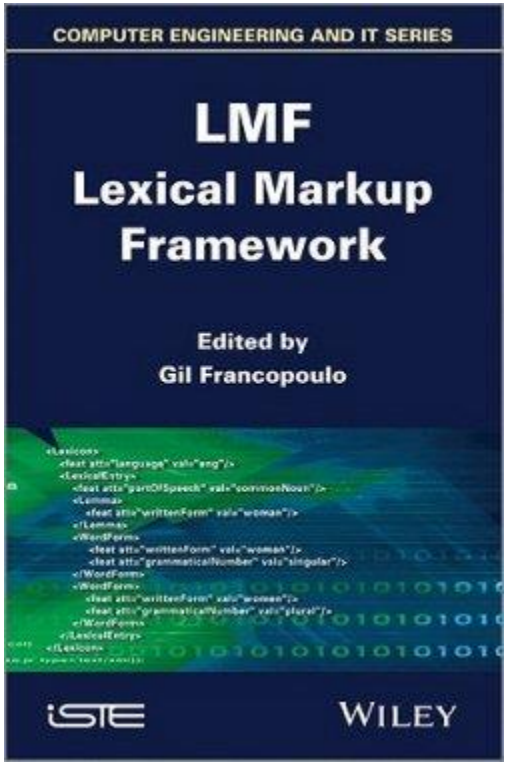
Typical Questions to Lexical Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

1. What is the meaning of *(to) sing*?
   - → text understanding
   - → word sense disambiguation
2. What are typical syntactic usages of the verb *(to) sing*?
   - → natural language generation
   - → grammar exercises
3. Does *bunny* have a special meaning when used in a sports report?
   - → domain adaptation
   - → genre classification
4. What is another word for *promising*?
   - → writing aid
   - → text simplification
5. What is a French equivalent of the English noun *plant*?
   - → foreign language learning
   - → automatic translation

Terminology

- **Lexical Markup Framework (LMF),** ISO 24613:2008
- LMF is an abstract model for representing electronic lexical resources
- **UBY-LMF** is first large-scale implementation involving different types of resources (e.g., collaboratively built ones)

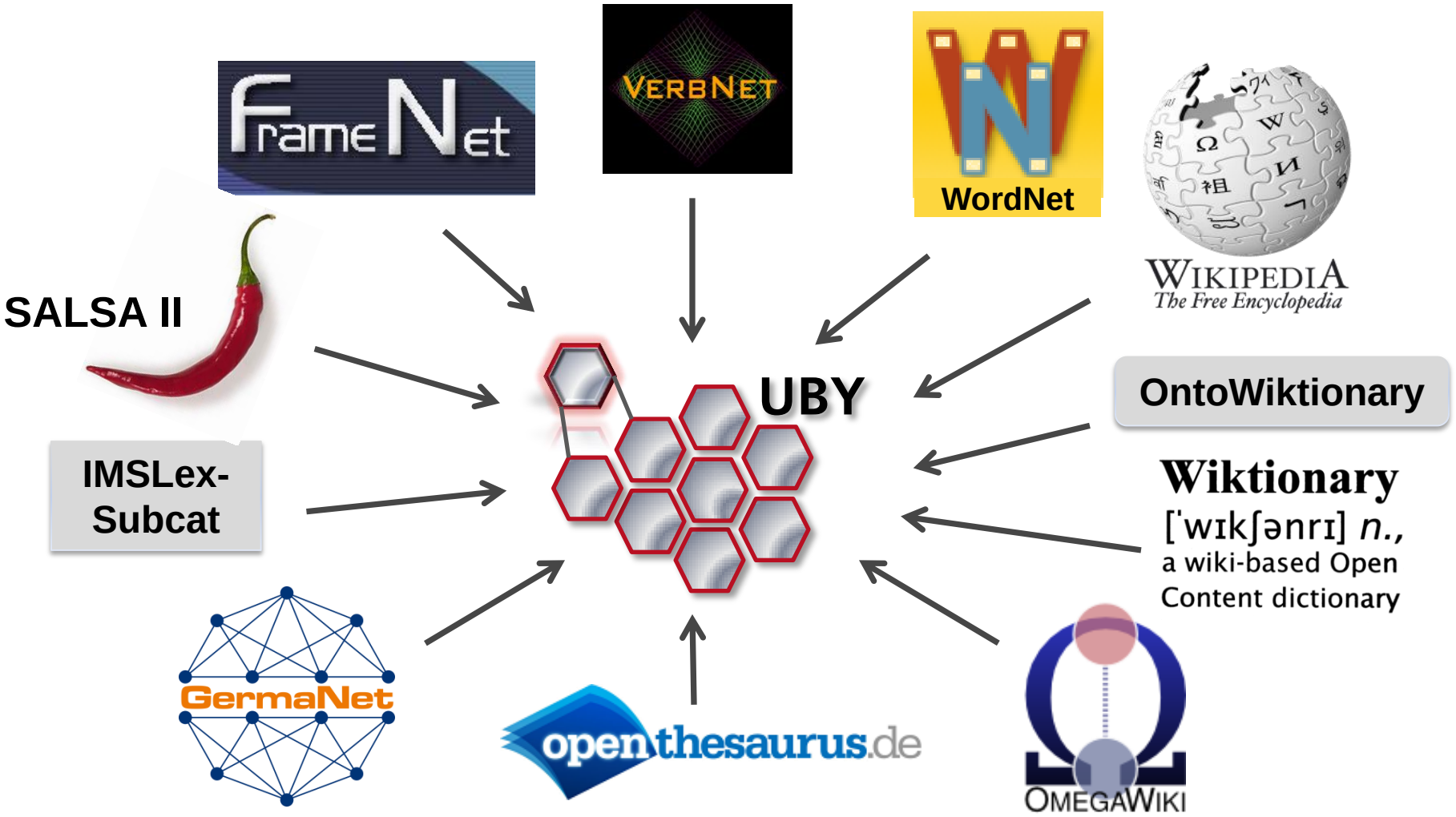


Gil Francopoulo (Ed.): **LMF: Lexical Markup Framework**, London: Wiley-ISTE, 2013.
ISBN: 978-1-84821-430-9.

http://www.lexicalmarkupframework.org/

UBY – Linked Lexical Resource

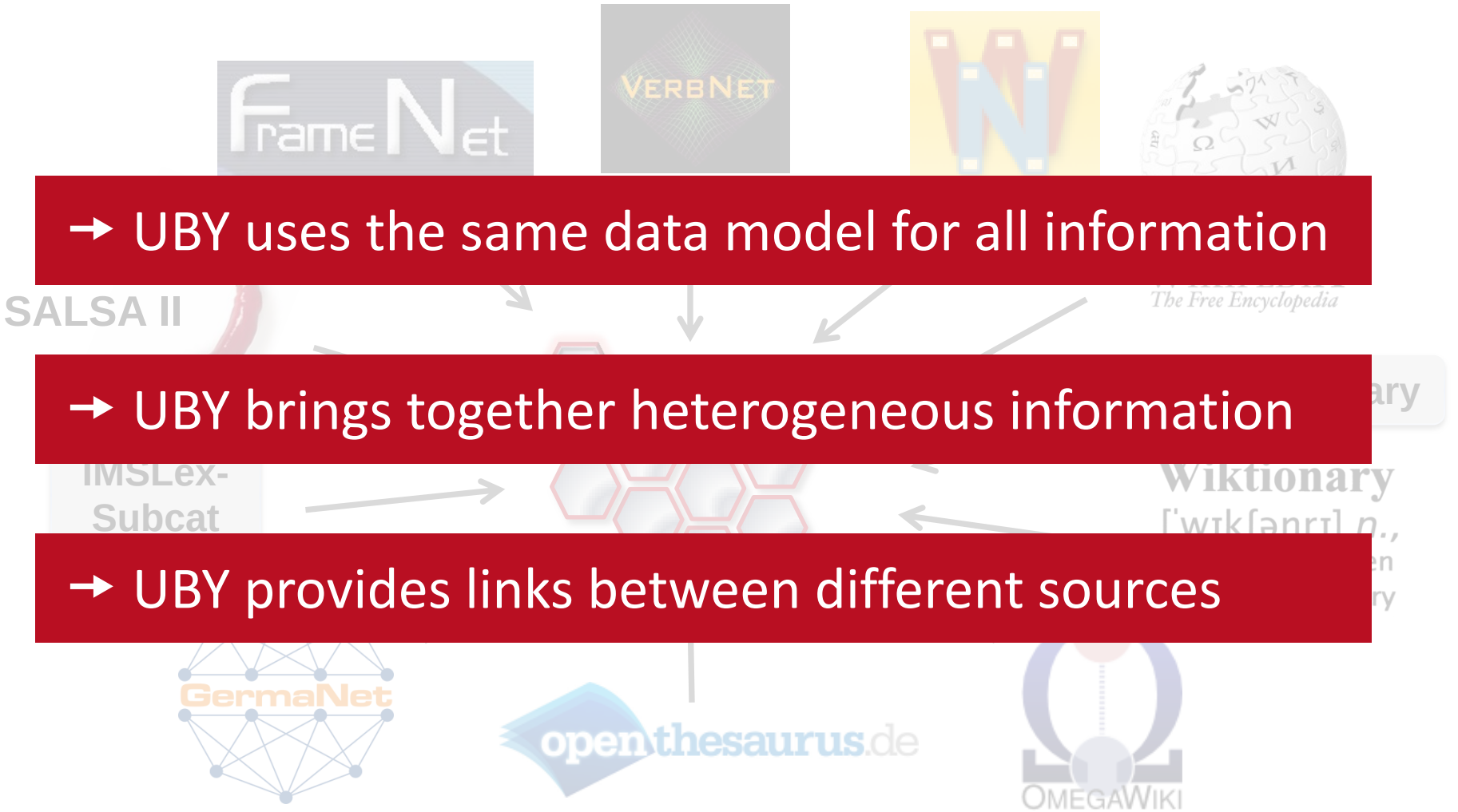
TECHNISCHE UNIVERSITÄT DARMSTADT



UKP

Why UBY?

TECHNISCHE UNIVERSITÄT DARMSTADT



→ UBY uses the same data model for all information

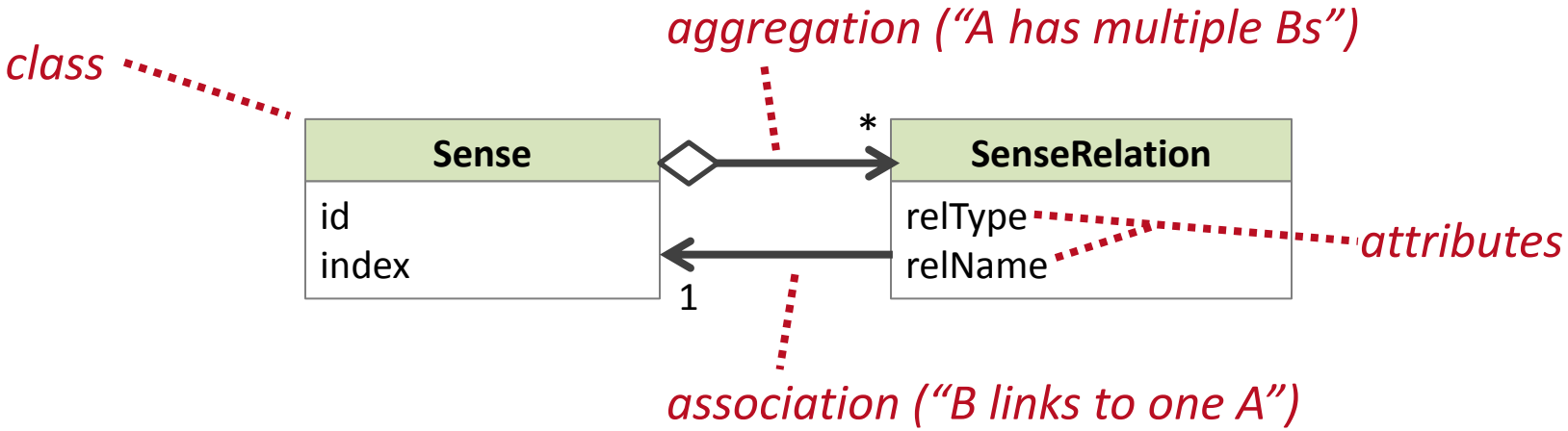
→ UBY brings together heterogeneous information

→ UBY provides links between different sources

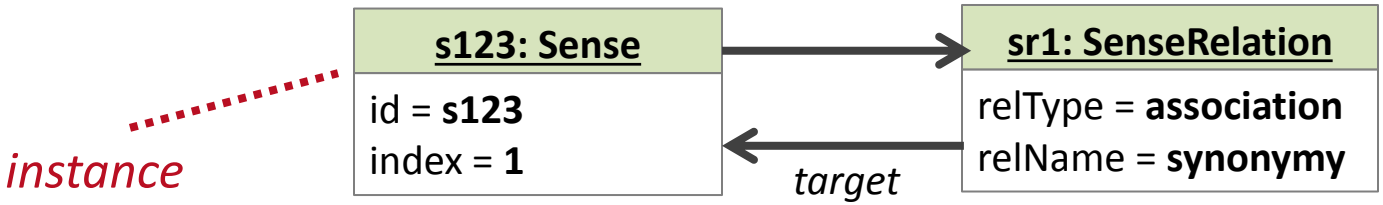
UKP

Notation

UML class diagrams:



UML object diagrams:



Reading Suggestions

TECHNISCHE UNIVERSITÄT DARMSTADT

- **[LMF-Paper]** G. Francopoulo/M. George/N. Calzolari/M. Monachini/N. Bel/M. Pet/C. Soria: Lexical Markup Framework (LMF), in: *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC), pp. 233–236, 2006. Genoa, Italy.
- **[LMF-Book]** G. Francopoulo (Ed.): LMF: *Lexical Markup Framework*, London: Wiley-ISTE, 2013.
- **[LMF-Standard]** *Language resource management – Lexical markup framework (LMF)*, ISO 24613:2008, International Organization for Standardization, Geneva, Switzerland, 2008.
- **[TEI-Standard]** *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Version 2.3.0, TEI Consortium, Charlottesville, VA, 2013.
- **[RDF-Standard]** *Resource Description Framework (RDF)*: Concepts and Abstract Syntax, W3C Recommendation 10 February 2004, World Wide Web Consortium, Cambridge, MA, 2004.
- **[UBY]** I. Gurevych/J. Eckle-Kohler/S. Hartmann/M. Matuschek/Ch.M. Meyer/Ch. 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), pp. 580–590, 2012. Avignon, France.
- **[UBY-LMF]** J. Eckle-Kohler/I. Gurevych/S. Hartmann/M. Matuschek/Ch.M. Meyer: UBY-LMF – A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF, in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp. 275–282, 2012. Istanbul, Turkey.

Lexical Resources for NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

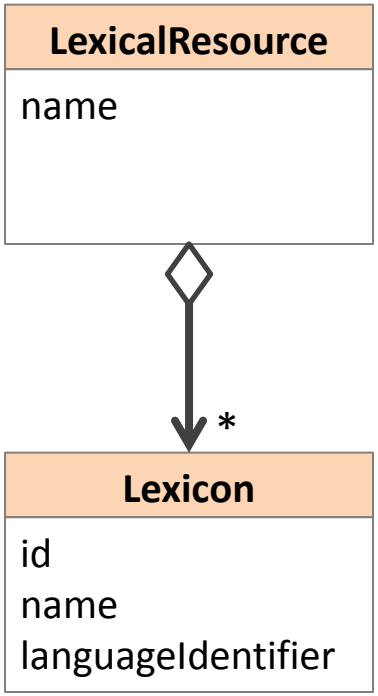
- Introduction
- **Dictionaries** — Try it!
- Wordnets and Thesauri
- Multilingual and Aligned Resources — Try it!

– Break –

- Deep Semantic Resources
- Syntactic Resources — Try it!
- Lexical Resources in Action
- Wrap-up — Try it!

http://pixabay.com/en/coffee-cup-drink-cafe-brown-mug-312521/ (CC0)

Lexical Resource & Lexicon



Lexical resource

- a.k.a. lexical database, lexical knowledge base
- Consists of one or several lexicons
- Parent for all further lexical information

Lexicon

- Language-specific
- Contains multiple lexical entries, syntactic representations, semantic representations, etc.

Example: a bilingual dictionary is one lexical resource that consists of two lexicons (e.g., Italian→German / German→Italian)

http://upload.wikimedia.org/wikipedia/commons/f/fe/Langenscheidt_3W%C3%B6rterb%C3%BCcher.JPG (Public Domain)

Electronic Dictionaries (Examples)

TECHNISCHE UNIVERSITÄT DARMSTADT

Wiktionary

- Free, collaboratively created online dictionary
- http://www.wiktionary.org

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

Digitales Wörterbuch der Deutschen Sprache

- Large-scale German dictionary project at BBAW
- http://www.dwds.de

DWDS

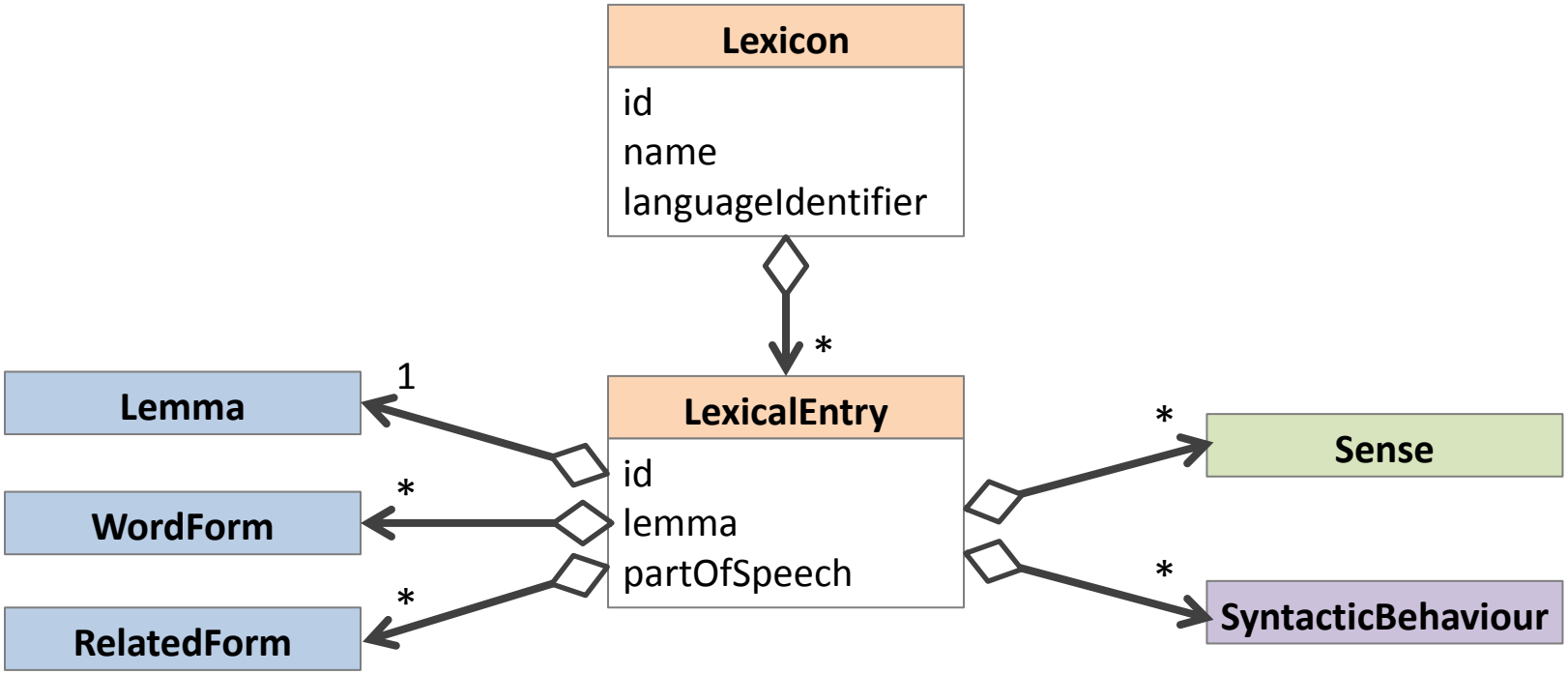
Online-Wortschatz-Informationssystem Deutsch

- Dictionary portal at IDS, Mannheim
- http://www.owid.de

OWID

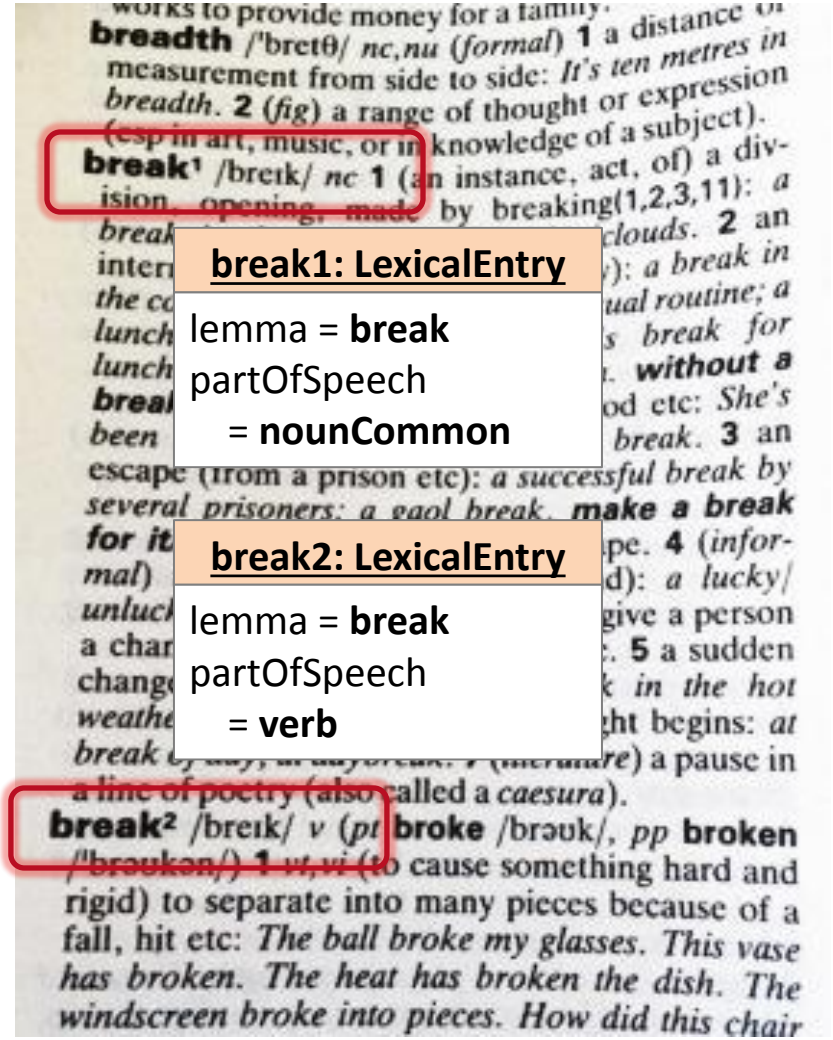
Many other electronic dictionaries…

Lexical Entry

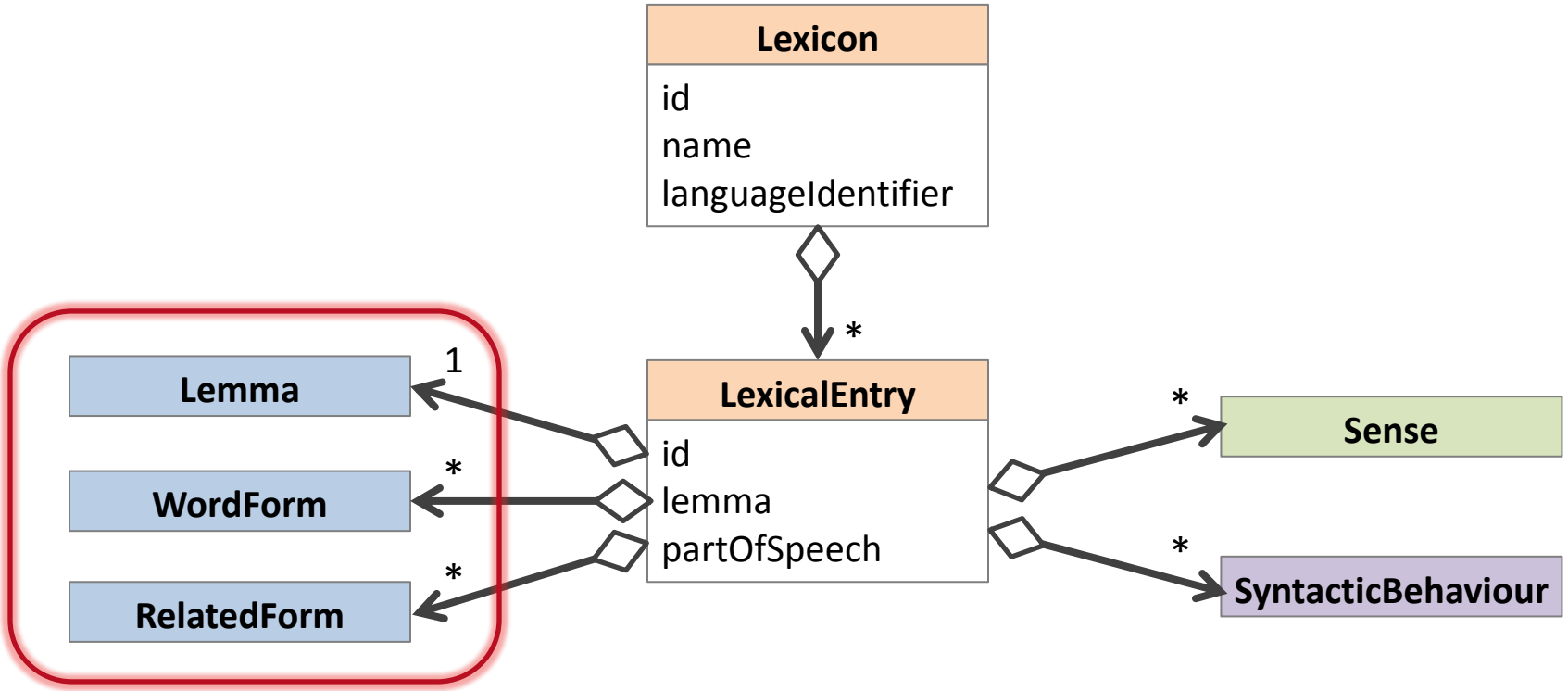


Lexical entry (a.k.a. lexeme, headword): container for managing multiple word forms, meanings, syntactic behaviours; equivalent to a dictionary article; defined by a lemma and part of speech tag.

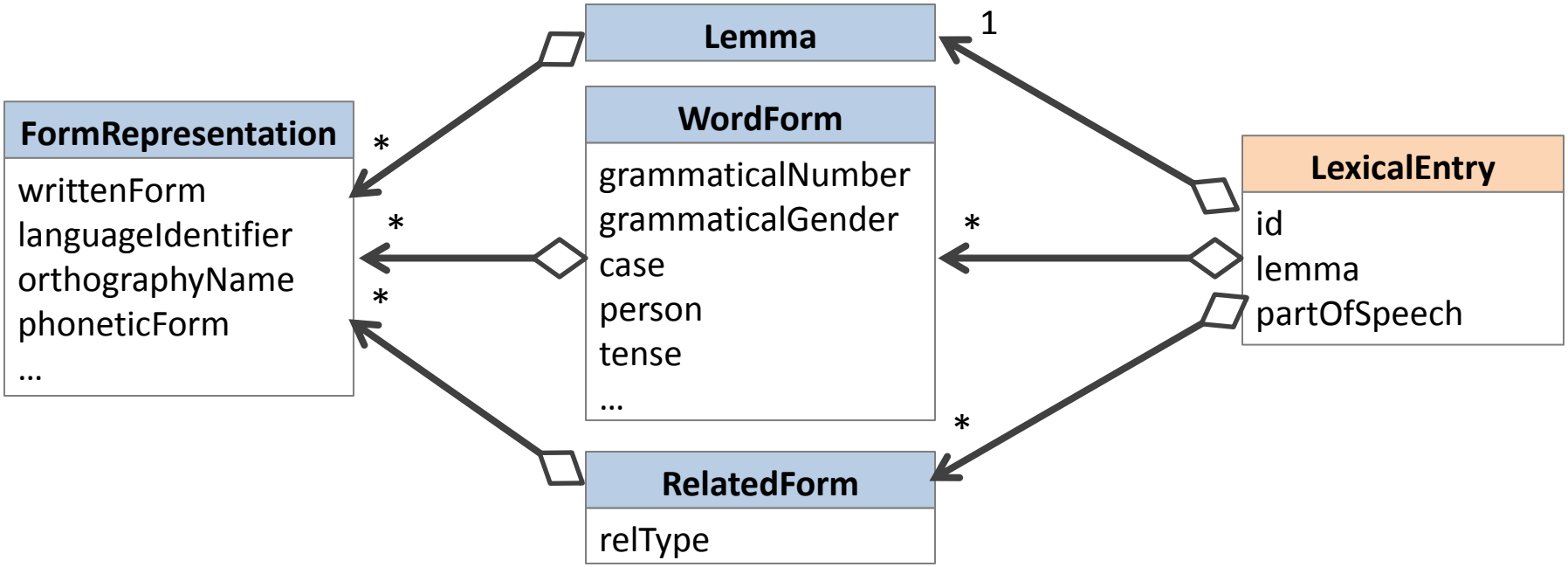
Lexical Entry: Examples



Word Forms

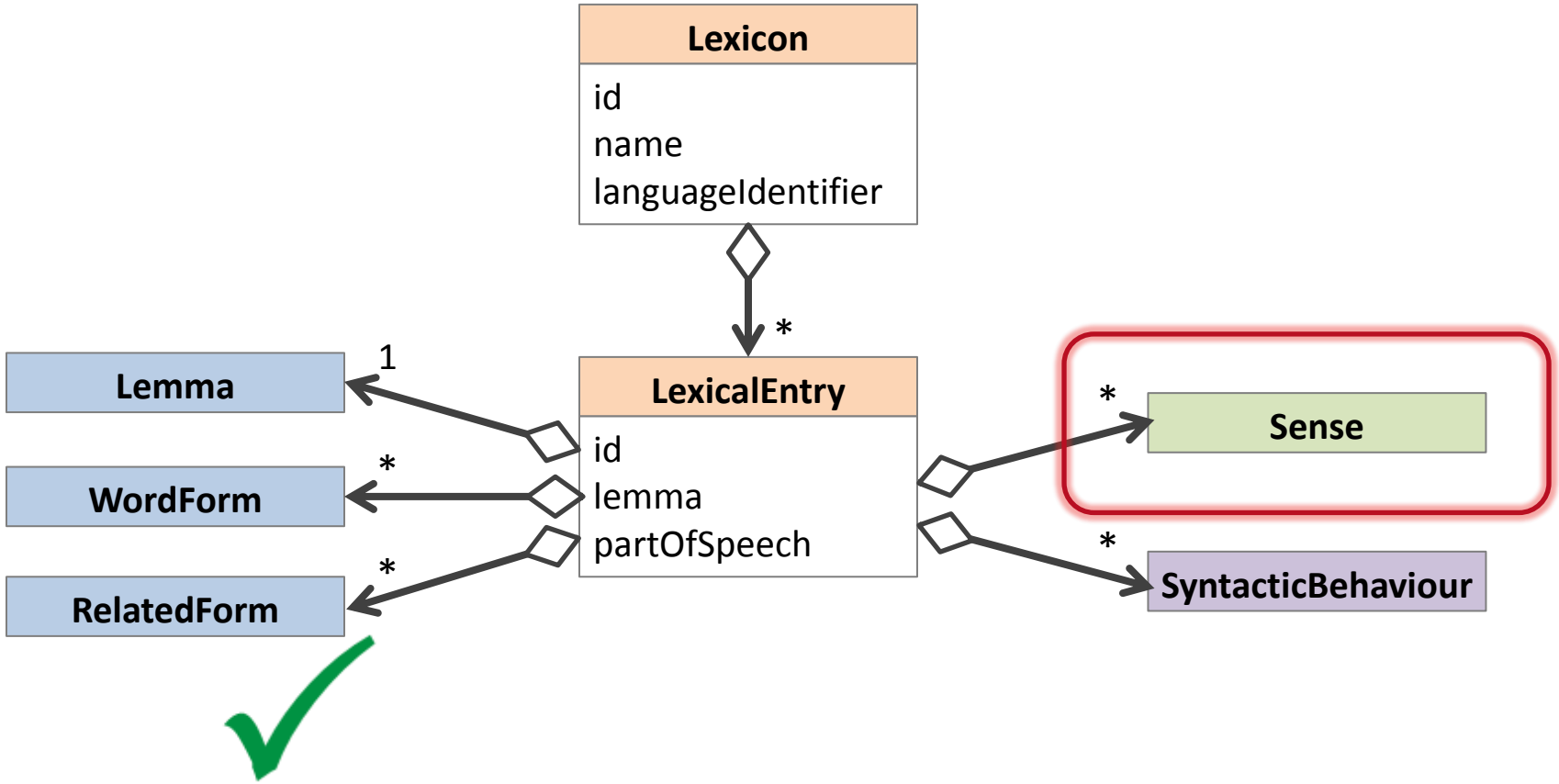

Lexicon
id
name
languageIdentifier
*
LexicalEntry
id
lemma
partOfSpeech
1
Lemma
*
WordForm
*
RelatedForm
*
Sense
*
SyntacticBehaviour

Word Forms: Data Model



- **Lemma**: conventional form to represent a lexical entry
- **Word form**: any form that a lexical entry can take when used in a sentence or phrase; inflected by number, gender, person,…
- **Related form**: a similar form; related by derivation, compounding,…

Meaning


Lexicon
id
name
languageIdentifier
*
1
Lemma
LexicalEntry
id
lemma
partOfSpeech
*
WordForm
*
RelatedForm
*
Sense
*
SyntacticBehaviour

Meaning: Example

sing (*third-person singular simple present* **sings**, *present participle* **singing**, *simple past* **sang**, *past participle* **sung** *or* (*archaic*) **sungen**)

1. (*intransitive*) To produce musical or harmonious sounds with one's voice.
   *"I really want to **sing** in the school choir." said Vera.*
2. (*transitive*) To express audibly by means of a harmonious vocalization. [quotations ▼]
3. (*transitive*) To soothe with singing.
   *to **sing** somebody to sleep*
4. (*intransitive*, *slang*) To confess under interrogation.
5. To make a small, shrill sound. [quotations ▼]
   *The air **sings** in passing through a crevice.*
6. To relate in verse; to celebrate in poetry. [quotations ▼]

- In Wiktionary, the verb (to) sing has 6 **senses**
- Each of them is described by a **definition** (a.k.a. paraphrase, gloss)
- 3 have a usage example (a.k.a. **sense example**)

Meaning: Example

s1: Sense

index = **1**

sing: LexicalEntry

lemma = **sing**
partOfSpeech
= **verb**

polysemy

s4: Sense

index = **4**

Meaning: Example

sing: LexicalEntry

lemma = **sing**
partOfSpeech
= **verb**

s1: Sense

index = **1**

s4: Sense

index = **4**

d1: Definition

e1: SenseExample

exampleType
= **senseInstance**

d4: Definition

t1: TextRepr

writtenText = **To produce musical or harmonious sounds with one's voice.**

t2: TextRepr

writtenText = **"I really want to sing in the school choir." said Vera.**

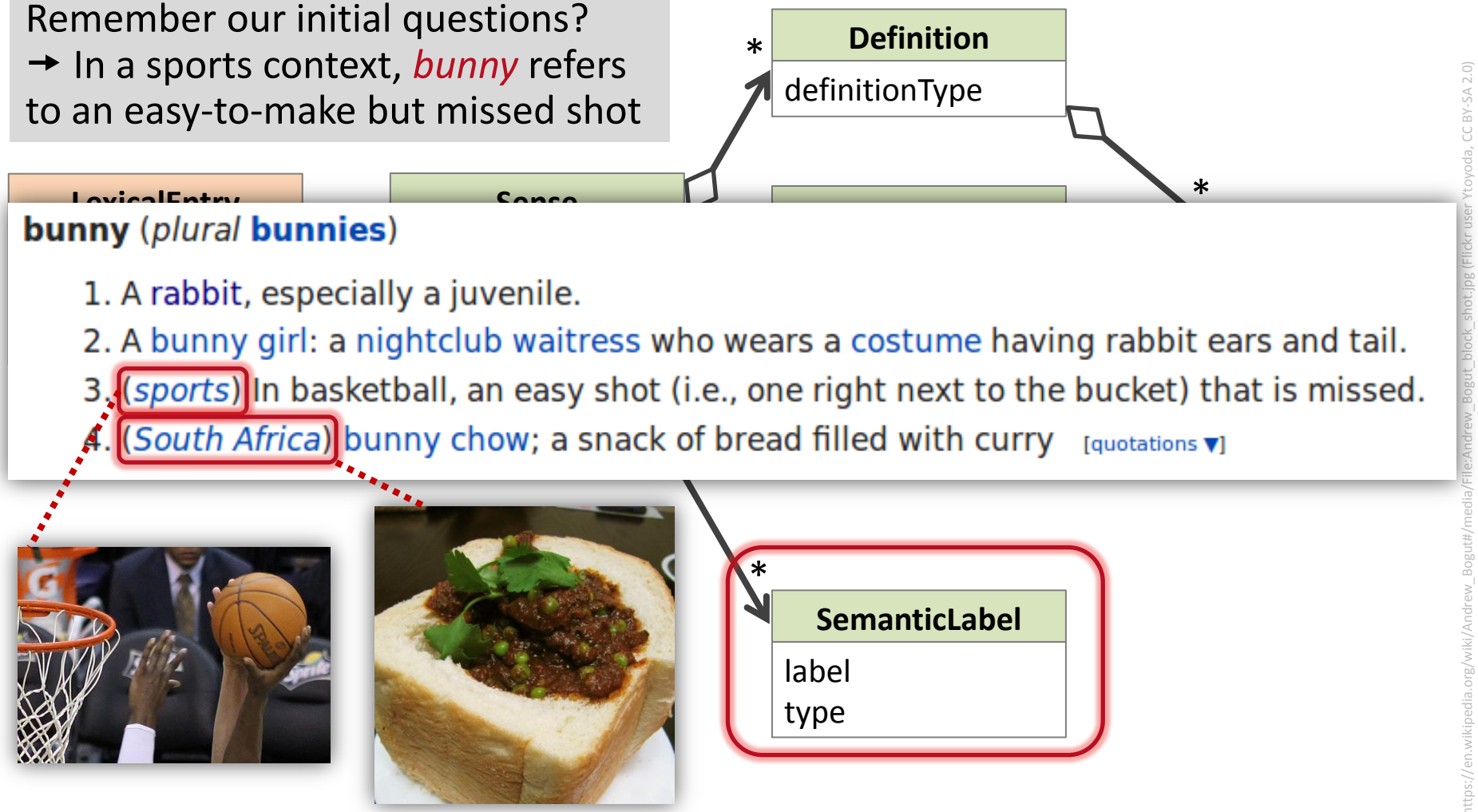
t3: TextRepr

writtenText = **To confess under interrogation.**

Remember our initial questions?
→ There are different meanings for *(to) sing* depending on the context!

Meaning: Semantic Labels

Remember our initial questions?
→ In a sports context, *bunny* refers to an easy-to-make but missed shot



https://en.wikipedia.org/wiki/Andrew_Bogut#/media/File:Andrew_Bogut_block_shot.jpg (Flickr user Ytoyoda, CC BY-SA 2.0)
https://www.flickr.com/photos/ramenfuel/6173617758 (Flickr user Jo del Corro, CC-BY 2.0)

Semantic Label Types

TECHNISCHE UNIVERSITÄT DARMSTADT

- **domain** (e.g., sports, chemistry)
- **regionOfUsage** (e.g., South Africa, Bavaria, Scottish)
- **timePeriodOfUsage** (e.g., 1800s, old fashioned)
- **register** (e.g., formal, slang)
- **sentiment** (e.g., negative judgment)
- **semanticNounClass** (e.g., onlyPlural, toponym [place name])
- **semanticField** (e.g., person, substance)

and many other types…

Meaning: Equivalents

TECHNISCHE UNIVERSITÄT DARMSTADT

Equivalents/Translations for *bunny*:

- Catalan: catxap (ca) *m*, conillet *m*
- Chinese:
  Mandarin: 小兔 (xiǎotù), 小兔子 (xiǎotùzi)
- Dutch: konijn (nl)
- Finnish: pupu (fi), kani (fi)
- French: lapereau (fr) *m*
- German: Kaninchen (de) *n*, Hase (de) *m*, Häschen *n* (*diminutive*), Häslein *n* (*diminutive*), Hoppelhäslein *n* (*diminutive*), Karnickel (de) *n* (*colloquial*), Schlappohr *n* (*humorous*)
- Macedonian: зајаче *n* (zajače)
- Norwegian: hare (no) *m*, harepus *m*
- Polish: króliczek (pl) *m*
- Portuguese: coelhinho (pt) *m*
- Romanian: iepuraș (ro)
- Russian: (*rabbit*) кролик (ru) *m* (králik), (*young rabit*) крольчонок (ru) *m* (krol'čónok), (*young hare*) зайчик (ru) *m* (zájčik)
- Sicilian: cunighieddu (scn) *m*
- Spanish: conejito *m*, gazapo (es) *m*

UKP

Try it yourself! – Assignment 1

TECHNISCHE UNIVERSITÄT DARMSTADT

- Unzip and import Java source files
- Unzip the downloaded UBY database and move it to the "embeddedUby" folder in your workspace
- Open `org.dkpro.uby.examples.Assignment1`

1) Explore which lexicons are in your database
2) Print the sense definitions of the noun *book* in FrameNet
3) List the word forms for the English *peculiarity* and the German *gut* in OntoWiktionary
4) Identify the semantic labels and their types of *bridge* in the English Wiktionary

15 minutes

- Alternative: https://uby.ukp.informatik.tu-darmstadt.de/uby-browser/

http://pixabay.com/en/stopwatch-racing-timer-timing-153398/ (CC0)

Reading Suggestions

TECHNISCHE UNIVERSITÄT DARMSTADT

- **[Lexicography]** B.T.S. Atkins/M. Rundell: *The Oxford Guide to Practical Lexicography*, Oxford: Oxford University Press, 2008.
- **[Lexicography]** R.H. Gouws/U. Heid/W. Schweickard/H.E. Wiegand (Eds.): *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (= Handbooks of Linguistics and Communication Science Series 5.4), Berlin/New York: de Gruyter, 2013.
- **[Lexicography]** Academic network on internet lexicography: http://www.internetlexikografie.de
- **[English dictionaries]** R. Lew: Online Dictionaries of English, in P.A. Fuertes-Olivera/H. Bergenholtz (Eds.): *E-Lexicography: The Internet, Digital Initiatives and Lexicography*, pp. 230–250, London/New York: Continuum, 2011.
- **[German dictionaries]** M. Mann (Ed.): *Digitale Lexikographie. Ein- und mehrsprachige elektronische Wörterbücher mit Deutsch: aktuelle Entwicklungen und Analysen* (= Germanistische Linguistik 223–224). Hildesheim/Zürich/New York: Olms, 2014.
- **[Wiktionary]** Ch.M. Meyer/I. Gurevych: Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography, chapter 13 in S. Granger/M. Paquot (Eds.): *Electronic Lexicography*, pp. 259-291, Oxford: Oxford University Press, 2012.
- **[Wiktionary]** Ch.M. Meyer: *Wiktionary: The Metalexicographic and the Natural Language Processing Perspective*, Dissertation, Technische Universität Darmstadt, tuprints 3654, 2013.

Lexical Resources for NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

- Introduction
- Dictionaries
- **Wordnets and Thesauri**
- Multilingual and Aligned Resources
- *– Break –*
- Deep Semantic Resources
- Syntactic Resources
- Lexical Resources in Action
- Wrap-up

Try it!

Try it!

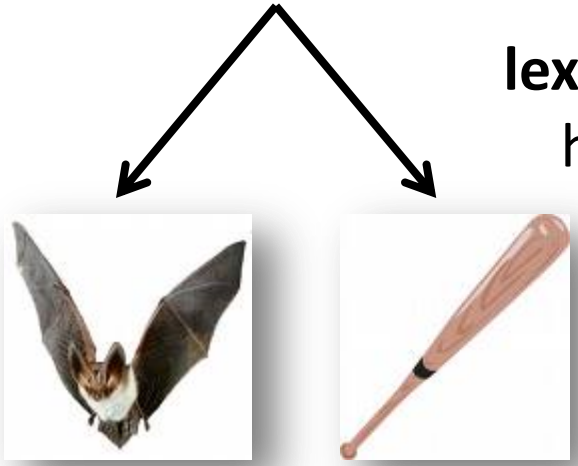
Try it!

Try it!

http://pixabay.com/en/coffee-cup-drink-cafe-brown-mug-312521/ (CC0)

Lexical Ambiguity vs. Synonymy

He hit the ball with the **bat**.



lexical ambiguity: words can have several meanings

bat **baseball racket**

synonymy: the same meaning can be expressed by different words

Synonymy

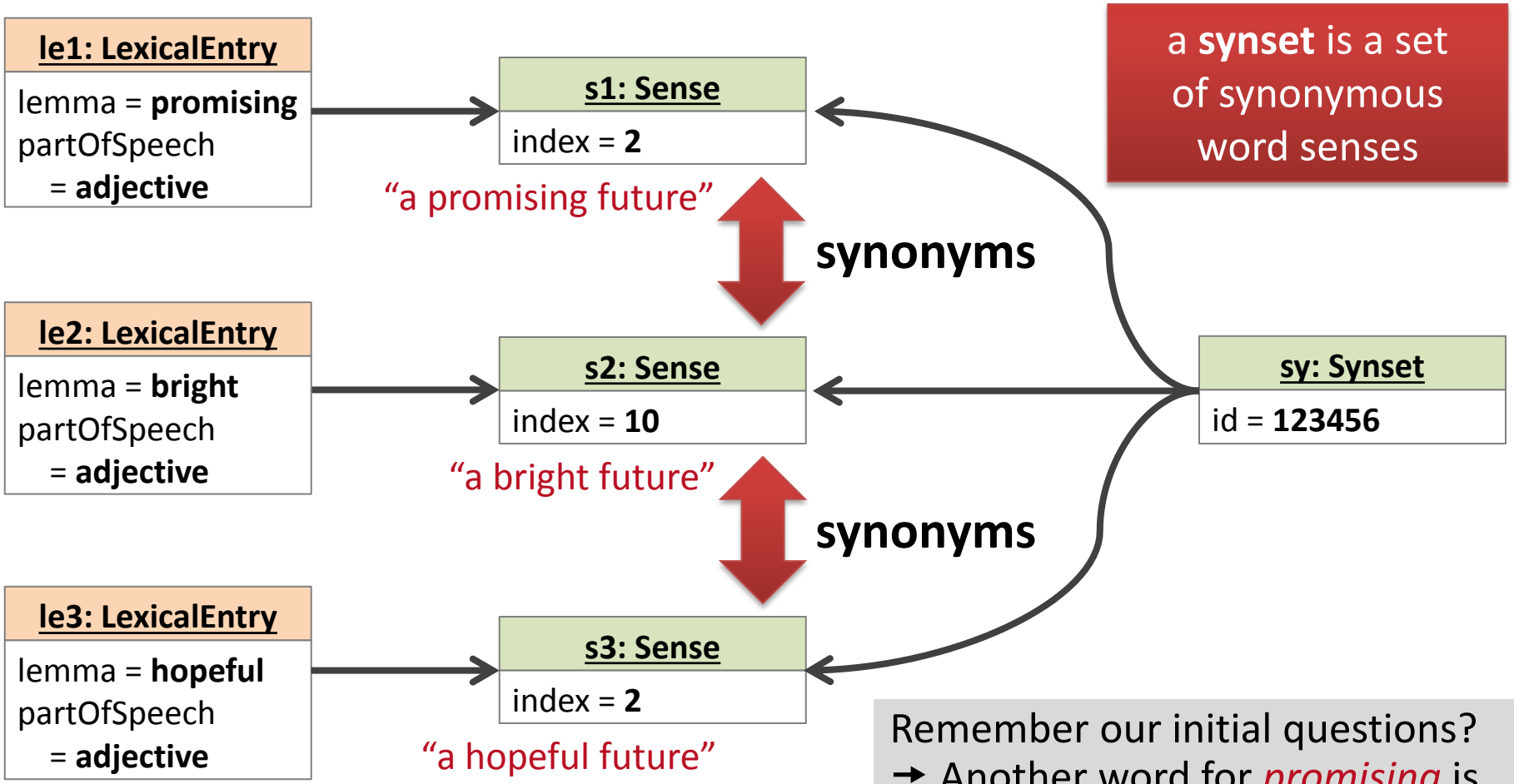
TECHNISCHE UNIVERSITÄT DARMSTADT

- Are `big` and `large` synonyms?
  - How `big/large` is that plane?
  - Would I be flying on a `big/large` plane?

- How about here:
  - Miss Nelson became a kind of `big` sister to Benjamin.
  - ? Miss Nelson became a kind of `large` sister to Benjamin.

- Synonymy is a relation between **senses** rather than **word forms**.
  - `big` has a sense of being older/grown-up (“große Schwester”)
  - `large` lacks this sense

Synonymy: Example



Wordnets and Thesauri (Examples)

TECHNISCHE UNIVERSITÄT DARMSTADT

Princeton WordNet

- Started in the mid 1980s by George Miller and team at Princeton University
- http://wordnet.princeton.edu/wordnet/

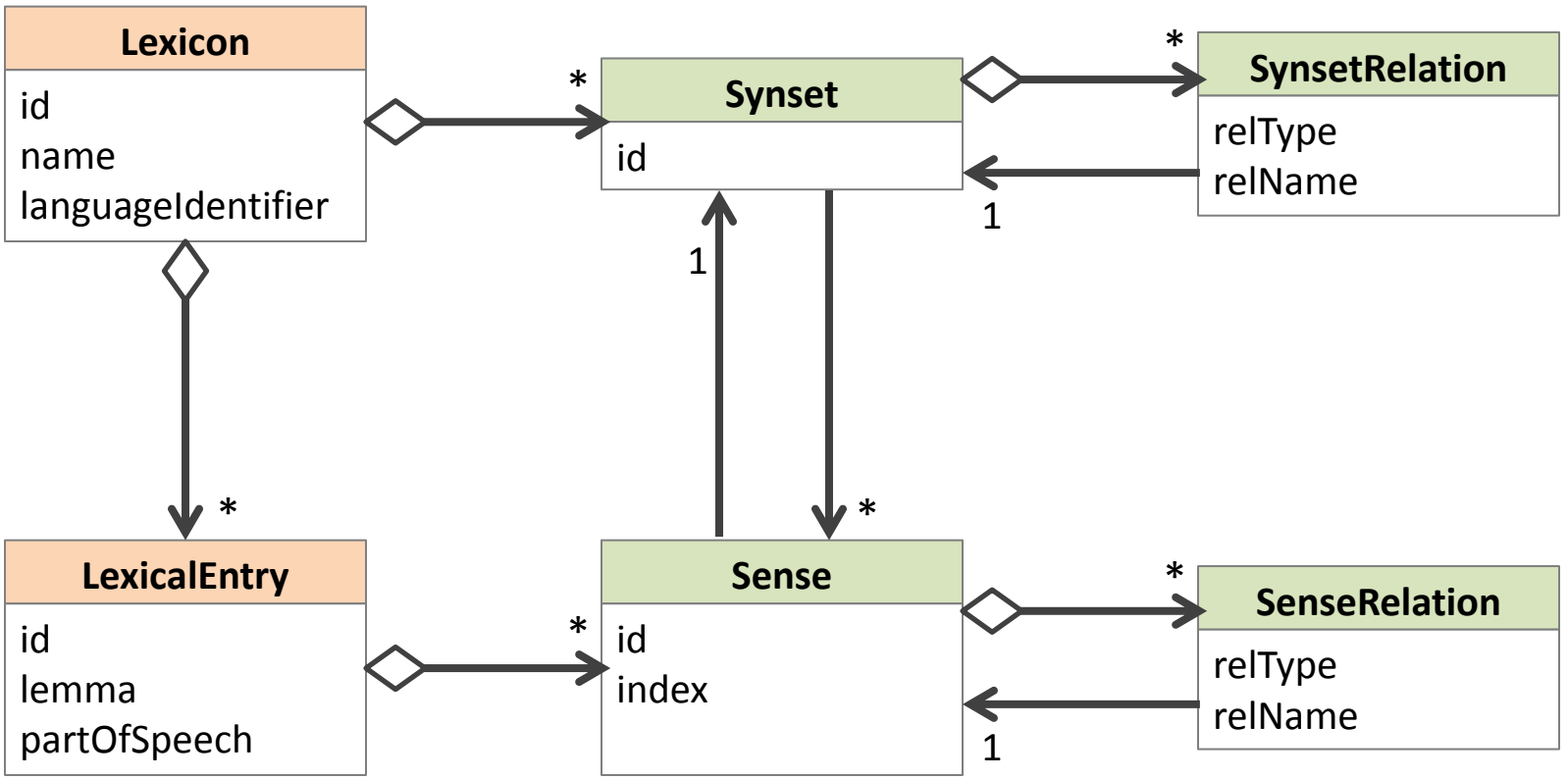
GermaNet

- German wordnet started in the 1990s at Tübingen University
- http://www.sfs.uni-tuebingen.de/GermaNet/

OpenThesaurus

- Freely available synonymy lexicon
- https://www.openthesaurus.de/

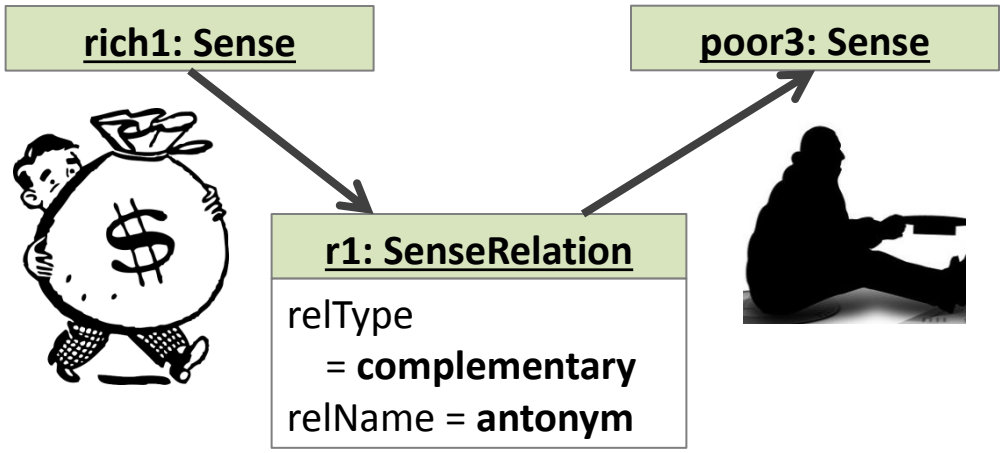
Synsets: Data Model



Antonymy

Relation between two senses having opposite meanings

- rich / poor
- rise / fall
- dark / light
- short / long
- hot / cold
- up / down
- leader / follower
- increase / decrease
- stable / unstable



http://pixabay.com/en/moneybags-wealth-rich-wealthy-30556/ (CC0)

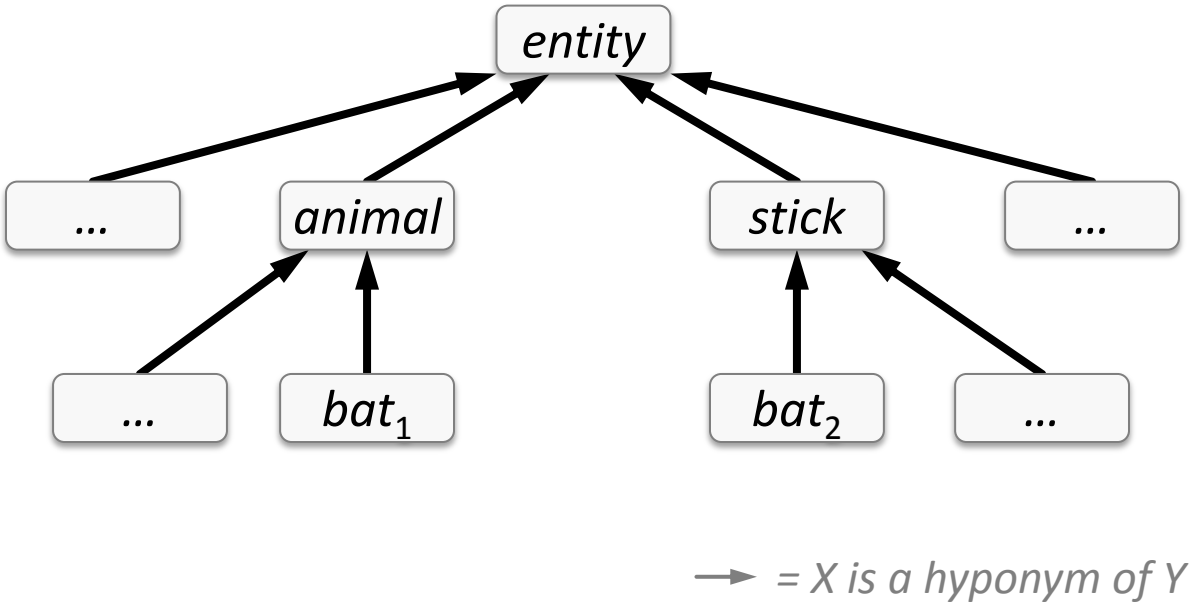
Hypernymy and Hyponymy

Taxonomic/hierarchical relation between two senses

- car is a hyponym of vehicle
- car is a hypernym of taxi

Examples:

- car / vehicle
- dog / animal
- mango / fruit
- oak / tree



Overview of Relation Types

TECHNISCHE UNIVERSITÄT DARMSTADT

relType	relName	Example	Description
complementary	antonym	rich → poor	opposite meaning
taxonomic	hypernym	car → vehicle	broader meaning
taxonomic	hyponym	car → taxi	narrower meaning
taxonomic	cohyponym	cat → dog	same hypernym (here: *pet*)
taxonomic	troponym	sleep → nap	"hyponymy for verbs"
partWhole	holonym	door → car	X is the whole of Y
partWhole	meronym	car → door	X is a part of Y
association	synonym	stack → pile	same meaning
association	seeAlso	bread → baker	related meaning
…	…	…	…

Sense vs. Synset

TECHNISCHE UNIVERSITÄT DARMSTADT

Sense:

- pair of form and meaning
- associated information is limited to a particular sense and its usage

- **die** to stop being alive
- **kick the bucket** *(phrase, humorous)* to die.
- **perish** *(mainly literary)* to die, usually because of an illness or something that happens suddenly

(Taken from the online Macmillan Dictionary)

Synset:

- contains multiple senses
- associated information applies to all senses of the synset

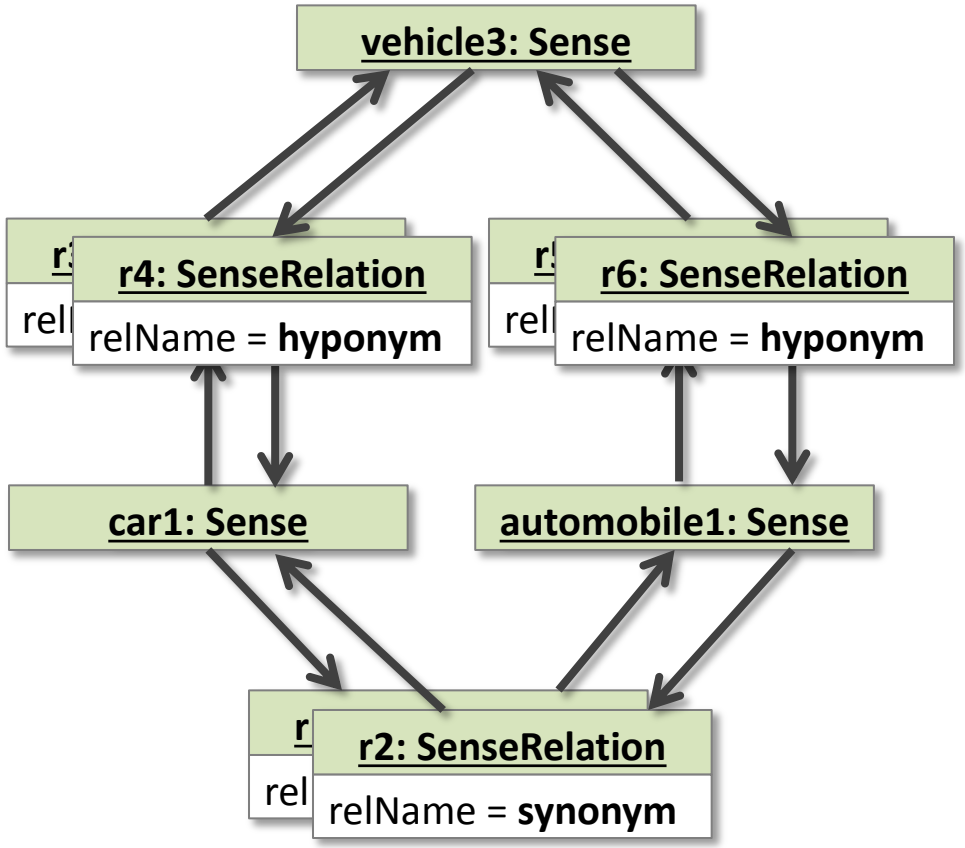
{die, kick the bucket, perish,…}

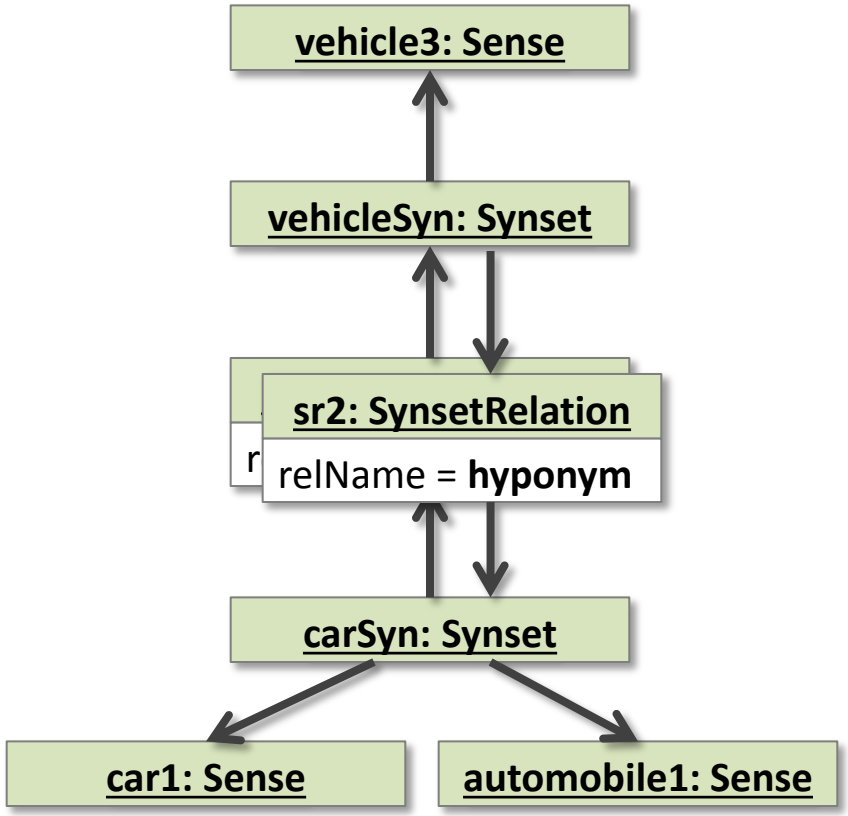
- pass from physical life […]

(Taken from WordNet)

Sense Relations vs. Synset Relations

Modeling alternatives – Compare:





Why not restrict to one alternative?

TECHNISCHE UNIVERSITÄT DARMSTADT

Avoid redundancy:

- {car, automobile, motorcar} is hyponym of {motor vehicle, automotive vehicle} – 1 synset relation vs. 6 sense relations

Deal with lexicon gaps:

- Synonyms: island → oasis, oasis → island, oasis → refuge
- But not: refuge → oasis, island → refuge, refuge → island

Allow relations between specific forms:

- {unvoiced, voiceless, surd, hard}

- {voiced, sonant, soft}

UKP

Reading Suggestions

- **[Princeton WordNet]** Ch. Fellbaum (Ed.): *WordNet: An Electronic Lexical Database* (= Language, Speech, and Communication), Cambridge, MA: MIT Press, 1998.
- **[GermaNet]** B. Hamp/H. Feldweg: GermaNet – a Lexical-Semantic Net for German, in: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pp. 9–15, 1997. Madrid, Spain.
- **[GermaNet]** V. Henrich/E. Hinrichs: GernEdiT – The GermaNet Editing Tool, in: *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (LREC), pp. 2228–2235, 2010. Valletta, Malta.
- **[OpenThesaurus]** D. Naber: OpenThesaurus: ein offenes deutsches Wortnetz, in B. Fisseni/H.-C. Schmitz/B. Schröder/P. Wagner (Eds.) : *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung*, pp. 422–433, Frankfurt: Peter Lang, 2005.
- **[Wordnet-LMF]** C. Soria/M. Monachini/P. Vossen: Wordnet-LMF: Fleshing out a Standardized Format for Wordnet Interoperability, in: *Proceedings of the 2009 International Workshop on Intercultural Collaboration*, pp. 139–146, 2009. Palo Alto, CA, USA.
- **[Synsets]** M. Matuschek/I. Gurevych: Beyond the Synset: Synonyms in Collaboratively Constructed Semantic Resources, in : *Re-thinking synonymy: semantic sameness and similarity in languages and their description: Book of Abstracts*, pp. 58–59, 2010. Helsinki, Finland.
- **[Wordnets]** Global WordNet association: http://www.globalwordnet.org

Lexical Resources for NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

Introduction

Dictionaries

Wordnets and Thesauri

Multilingual and Aligned Resources

– Break –

Deep Semantic Resources

Syntactic Resources

Lexical Resources in Action

Wrap-up

Try it!

Try it!

Try it!

Try it!

http://pixabay.com/en/coffee-cup-drink-cafe-brown-mug-312521/ (CC0)

Multilingual Resources (Examples)

TECHNISCHE UNIVERSITÄT DARMSTADT

Wikipedia

- Huge open-licensed encyclopedia in over 200 languages
- http://www.wikipedia.org

WIKIPEDIA
The Free Encyclopedia

OmegaWiki

- Free dictionary based on multilingual synsets
- http://www.omegawiki.org

OMEGAWIKI

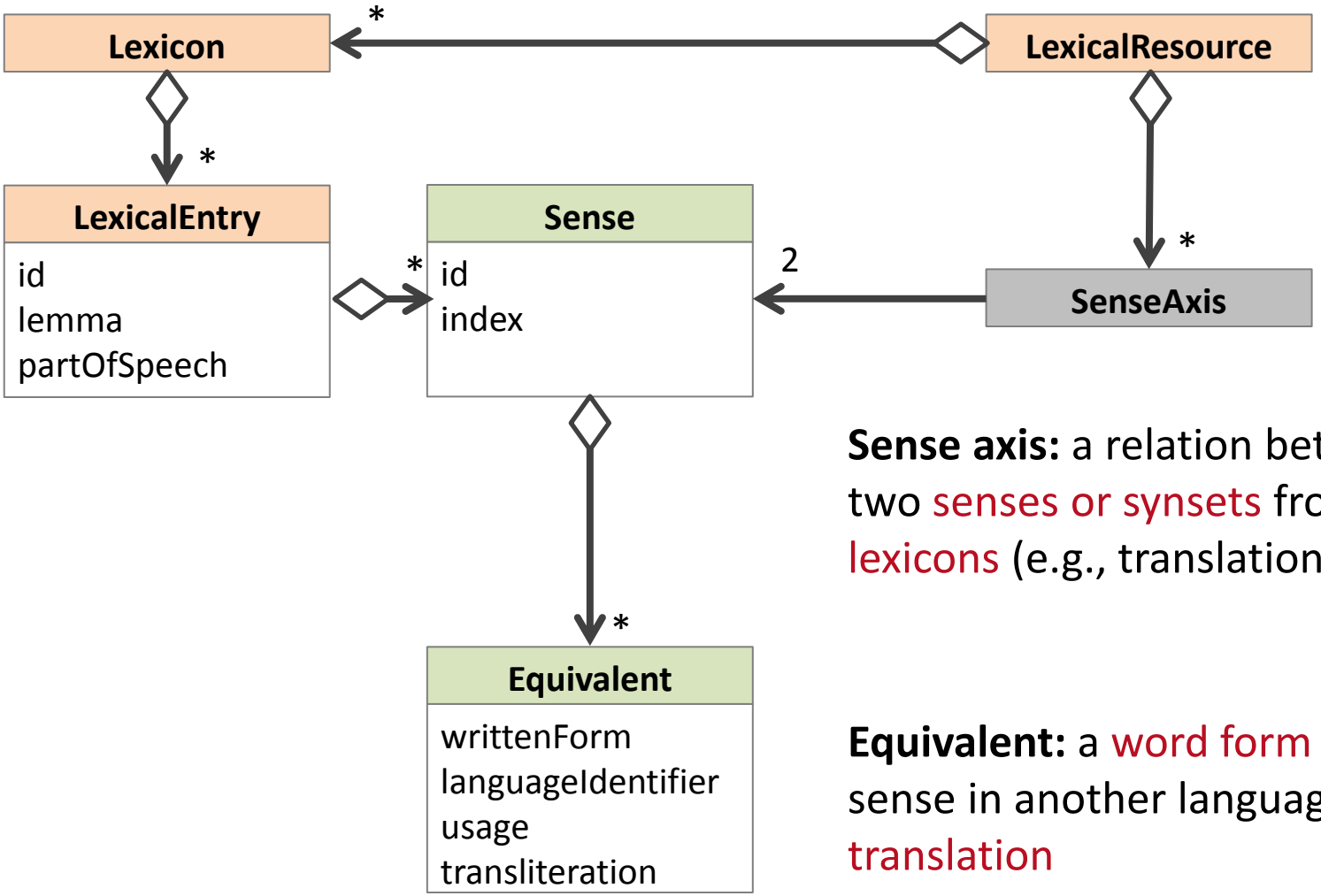
EuroWordNet

- Multilingual wordnet for several European languages
- http://www.illc.uva.nl/EuroWordNet/

EuroWordNet

Many other examples (e.g., bilingual dictionaries)

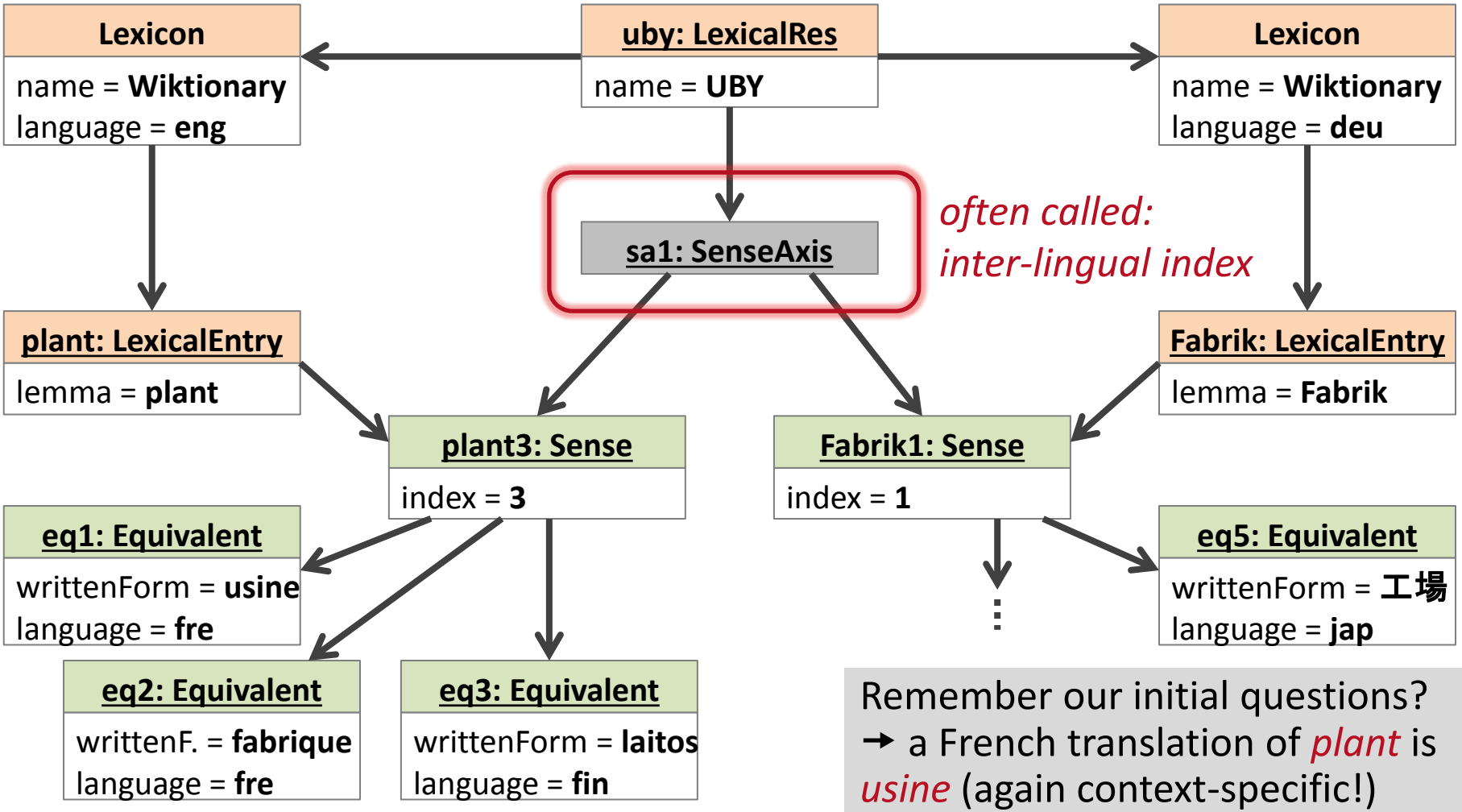
Equivalent vs. SenseAxis



Sense axis: a relation between two senses or synsets from different lexicons (e.g., translations)

Equivalent: a word form expressing a sense in another language; a translation

Equivalent vs. SenseAxis

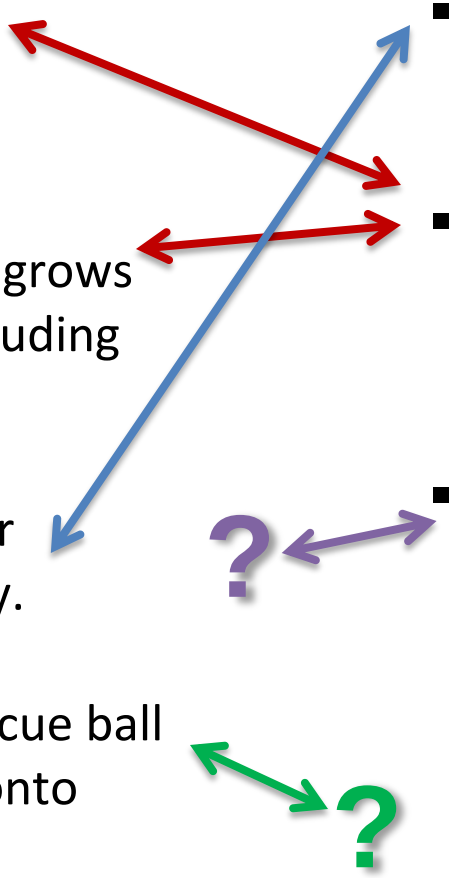


Word Sense Alignment

TECHNISCHE UNIVERSITÄT DARMSTADT

plant in Wiktionary

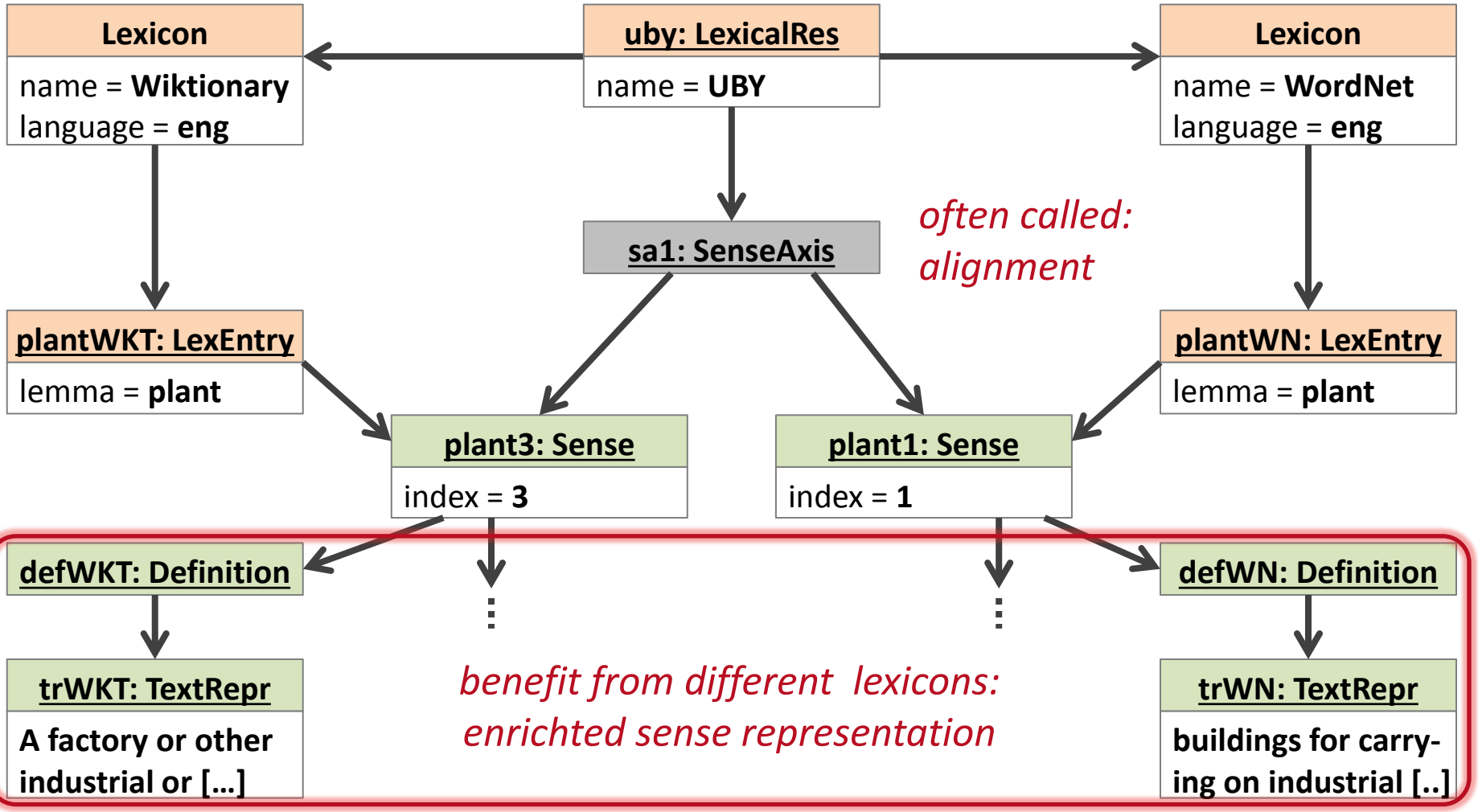
- (**botany**) An organism of the kingdom Plantae […]
- (proscribed as biologically inaccurate) Any **creature** that grows on soil or similar surfaces, including plants and fungi.
- A **factory** or other industrial or institutional building or facility.
- (**snooker**) A play in which the cue ball knocks one (usually red) ball onto another […]



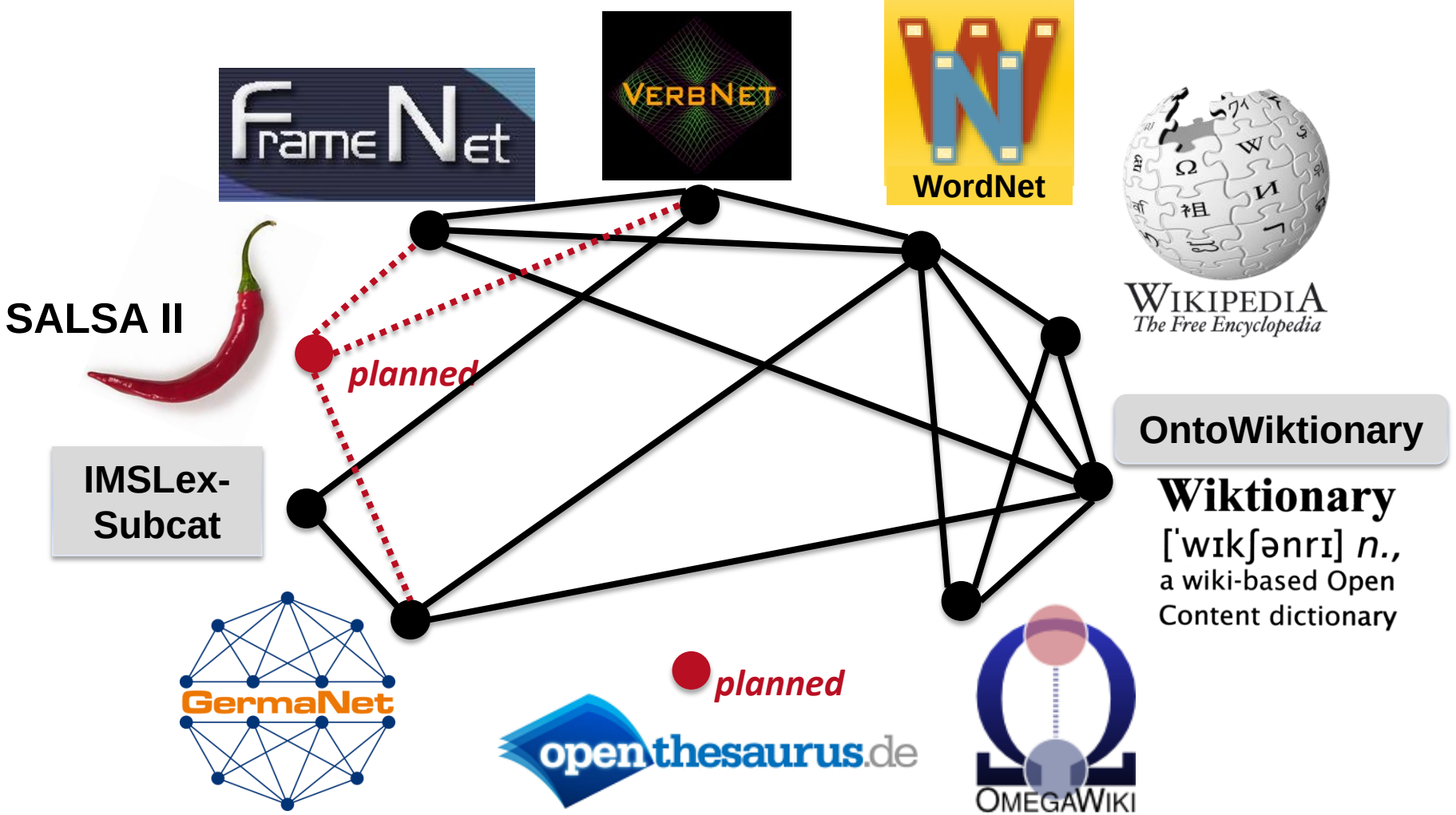
plant in WordNet

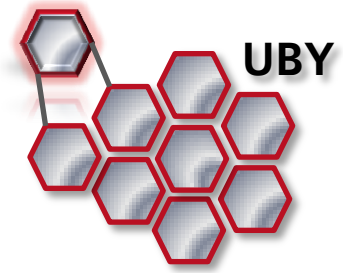
- buildings for carrying on **industrial labor**
- (**botany**) a living organism lacking the power of locomotion
- an **actor** situated in the audience whose acting is rehearsed but seems spontaneous to the audience

Word Sense Alignment



Alignments in UBY





OntoWiktionary

OntoWiktionary

A prototype of a structurally enriched resource

Wiktionary: Collaborative Dictionary

TECHNISCHE UNIVERSITÄT DARMSTADT

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

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
Special pages
Printable version
Permanent link
Page information
Cite this page
Add definition

In other projects
Wikipedia

Visibility

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

boat

English [edit]

Etymology [edit]

From Middle English *boot*, *bot*, *boet*, *boyt* ("boat"), from Old English *bāt* ("boat"), from Proto-Germanic **baitaz*, **baitan* ("boat, small ship"), from Proto-Indo-European **bheid-* ("to break, split"). Cognate with Old Norse *beit* ("boat").

Old Norse *bātr* (whence Icelandic *bátur*, Norwegian *båt*), Dutch *boot*, German *Boot*, and French *bateau* are all ultimately borrowings from the Old English word.

Wikipedia has an article on:
Boat

A **boat** kept on land

Pronunciation [edit]

- (*RP*) enPR: bōt, IPA: /bəʊt/, X-SAMPA: /b@Ut/
- Rhymes: -əʊt
- (*GenAm*) enPR: bōt, IPA: /boʊt/, X-SAMPA: /boUt/
- Rhymes: -oʊt
- Audio (US) 0:00 MENU

Noun [edit]

boat (*plural* **boats**)

1. A craft used for transportation of goods, fishing, racing, recreational cruising, or military use on or in the water, propelled by oars or outboard motor or inboard motor or by wind.
2. (*poker slang*) A full house.
3. (*chemistry*) One of two possible conformations of cyclohexane rings (the other being chair), shaped roughly like a boat.

Usage notes [edit]

There's no explicit limit, but the word ***boat*** usually refers to a relatively small watercraft that is generally smaller than a "ship" and

UKP

Java-based Wiktionary Library (JWKTL)

TECHNISCHE UNIVERSITÄT DARMSTADT

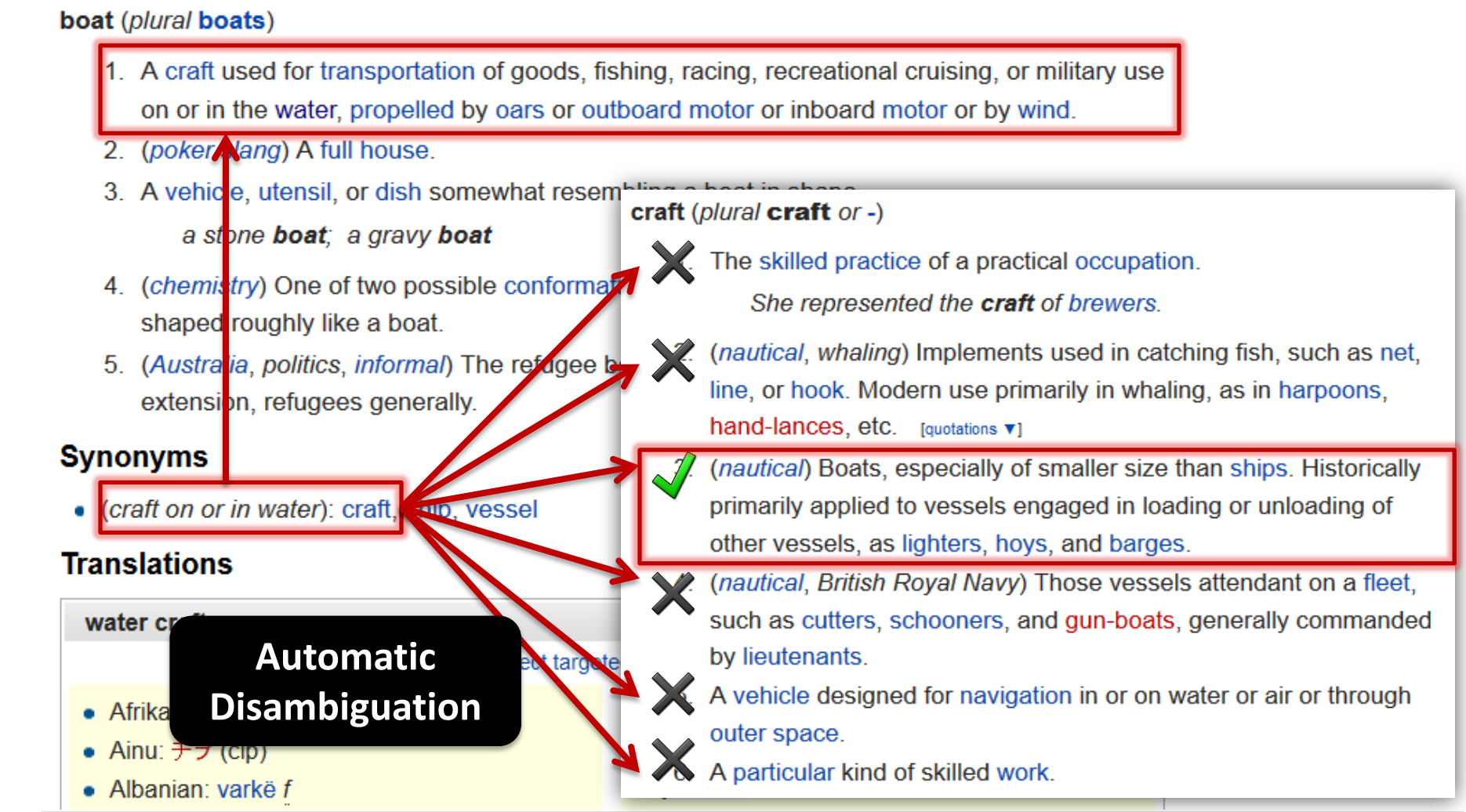
- Wiktionary articles are encoded in a wiki markup language
- Extraction software required to access the data!

```
====Verb====
{{en-verb|pays|paying|paid|past2=payed|past2_qual=archaic}}

# {{context|transitive|lang=en}} To [[give]] [[money]] or other compensation to in
exchange for goods or services.
#: {{usex|lang=en|he '''paid''' him to clean the place up;  he '''paid''' her off
the books and in kind where possible}}
#* {{quote-book|year=1918|author={{w|W. B. Maxwell}}|chapter=17
|title=[http://openlibrary.org/works/OL1097634W The Mirror and the Lamp]
|passage=This time was most dreadful for Lilian. Thrown on her own resources and almost
penniless, she maintained herself and '''paid''' the rent of a wretched room near the
hospital by working as a charwoman, sempstress, anything.}}
#* {{quote-magazine|date=2013-06-21|author={{w|Oliver Burkeman}}
|volume=189|issue=2|page=48|magazine={{w|The Guardian Weekly}} […]
```

UKP community project at GitHub:
http://dkpro.org/dkpro-jwktl/

Disambiguation of Relations/Translations



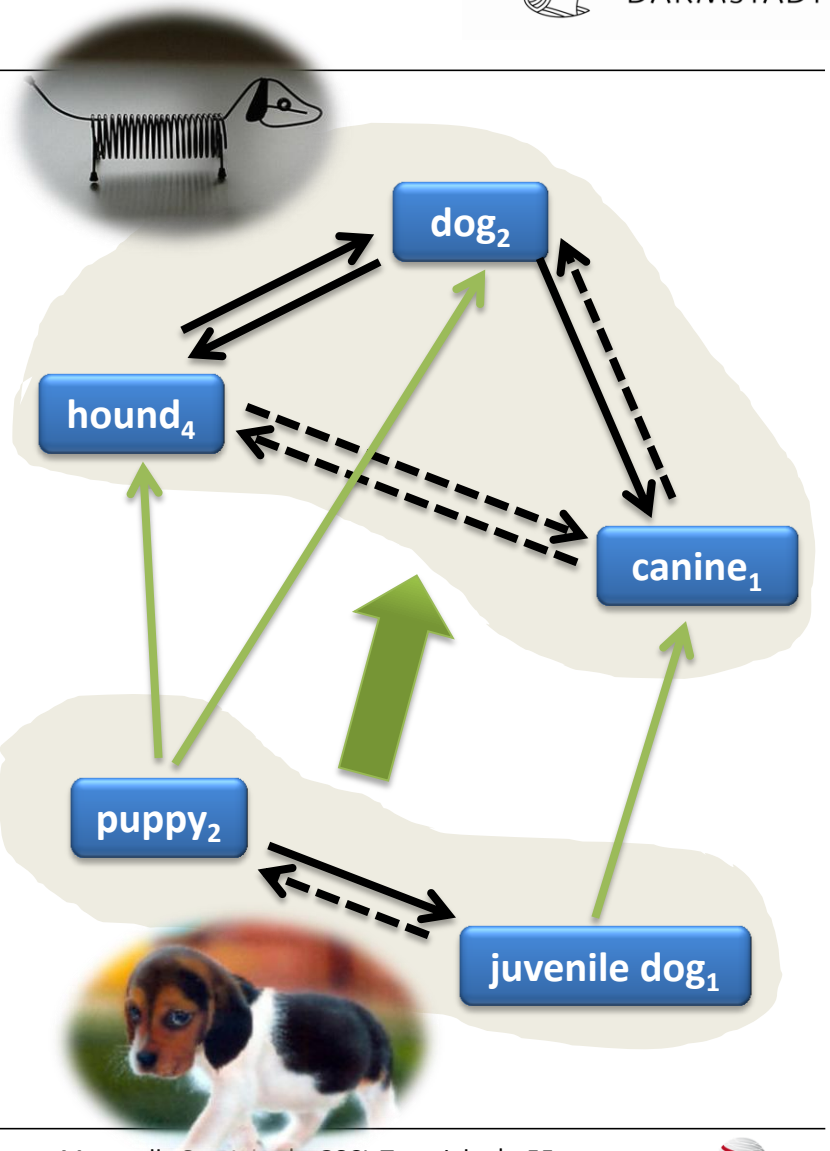
Creation of Synsets

1. Synset formation

- Start with Wiktionary senses
- Create synonymy graph
- Calculate the transitive hull

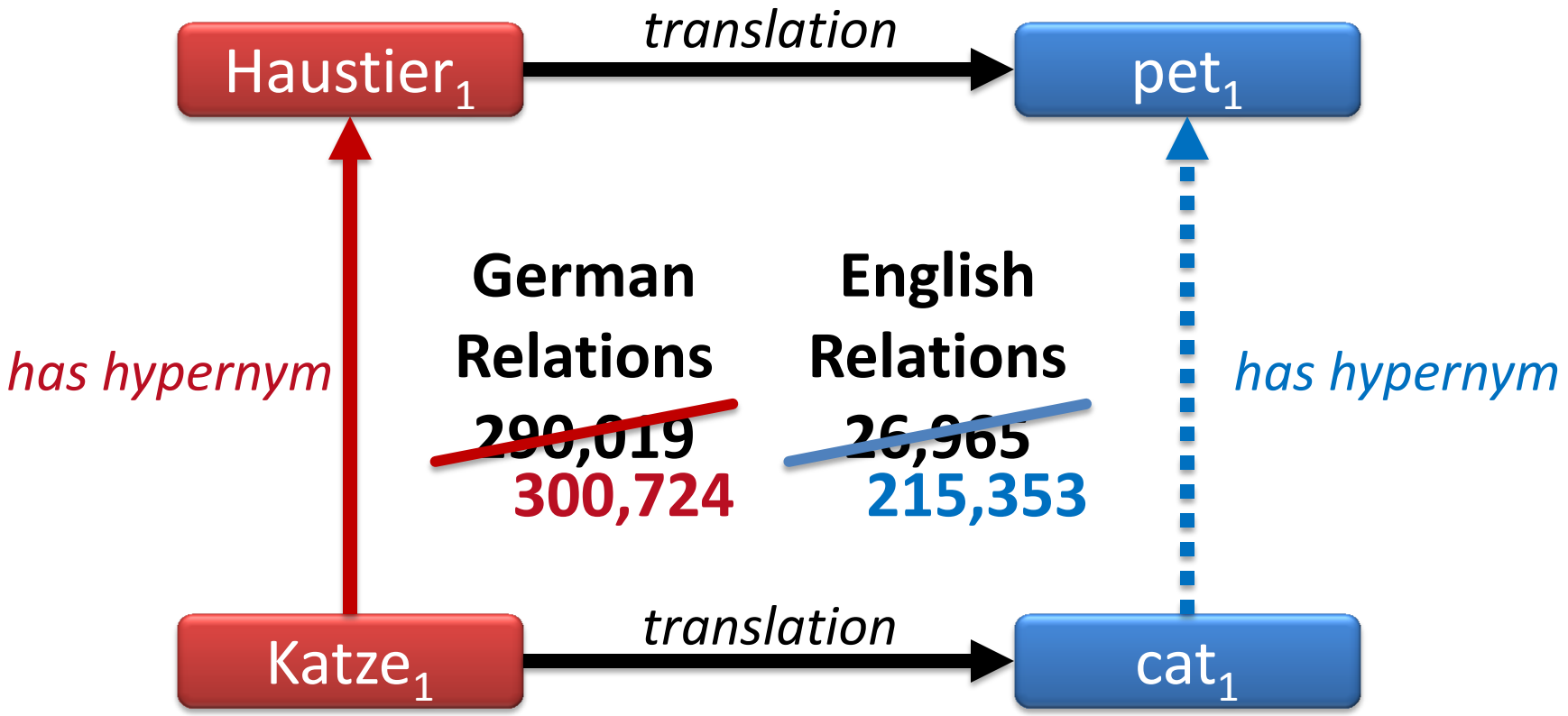
2. Synset relations

- From sense relations to synset relations



http://www.flickr.com/photos/sillydog/6546575/ (CC BY-SA 2.0 by Flickr user sillydog)
http://www.flickr.com/photos/7946788@N08/5126883524/ (CC BY-SA 2.0 by Flickr user filmismylove)

Inference of Semantic Relations



OntoWiktionary

TECHNISCHE UNIVERSITÄT DARMSTADT

Language edition: English OntoWiktionary (excerpt) Load OntoWiktionary

Page: 1312 Concepts per page: 3 Go

< Previous Next >

Lemma Exact Search Clear

Search Results:

Concept ID	Lexicalization ID	Lemma	Gloss
3935	322367:0:1	egocentric	selfish, self-centered
	549591:0:1	idiocentric	characteriz… oneself or … the ways of…
Subsumes (0)			
SubsumedBy (0)			
Related (2)	742106:0:2	individualistic	Interested in… …than others; egocentric
3936	206262:0:1	micronutrient	A mineral, vitamin or other substance that is essential, even in very small quantities, for growth or metabolism
	1582847:0:1	micromineral	A mineral of which only trace amounts are needed in the diet.
Subsumes (1)			
SubsumedBy (1)	550155:0:1	trace element	A chemical element present in a sample in very small quantities.
Related (3)			

UBY

Now part of UBY!

https://www.ukp.tu-darmstadt.de/data/ontowiktionary/

Try it yourself! – Assignment 2

TECHNISCHE UNIVERSITÄT DARMSTADT

Open `org.dkpro.uby.examples.Assignment2`

1) Explore the hypothesis that synonymy relations hold between senses rather than forms
2) Find the WordNet definitions
3) Access a lexicon using a synset iterator
4) Explore the noun *submarine* in the English OntoWiktionary
5) Create an enriched sense representation based on sense alignments

We start again with the second part at 12:00

http://pixabay.com/en/stopwatch-racing-timer-timing-153398/ (CC0)

Reading Suggestions

TECHNISCHE UNIVERSITÄT DARMSTADT

- **[EuroWordNet]** P. Vossen (Ed.): *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer, 1998.
- **[OmegaWiki]** M. Matuschek/Ch.M. Meyer/I. Gurevych: Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Translation Applications, *Translation: Computation, Corpora, Cognition: Special Issue "Language Technology for a Multilingual Europe"* 3 (1): 87–118, 2013.
- **[OntoWiktionary]** Ch.M. Meyer/I. Gurevych: OntoWiktionary -- Constructing an Ontology from the Collaborative Online Dictionary Wiktionary, chapter 6 in M.T. Pazienza/A. Stellato (Eds.) : *Semi-Automatic Ontology Development: Processes and Resources*, pp. 131–161, Hershey, PA: IGI Global,, 2012.
- **[OntoWiktionary]** Ch.M. Meyer: *Wiktionary: The Metalexicographic and the Natural Language Processing Perspective*, Dissertation, Technische Universität Darmstadt, tuprints 3654, 2013. http://tuprints.ulb.tu-darmstadt.de/3654/
- **[Sense Alignment]** Ch.M. Meyer/I. 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, 2011. Chiang Mai, Thailand.
- **[Sense Alignment]** M. Matuschek: *Word Sense Alignment of Lexical Resources*. Dissertation, Technische Universität, Darmstadt, tuprints 4355, 2015. http://tuprints.ulb.tu-darmstadt.de/4355/

Lexical Resources for NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

- Introduction
- Dictionaries
- Wordnets and Thesauri
- Multilingual and Aligned Resources

– Break –

- **Deep Semantic Resources**
- Syntactic Resources
- Lexical Resources in Action
- Wrap-up

Try it!

Try it!

Try it!

Try it!

http://pixabay.com/en/coffee-cup-drink-cafe-brown-mug-312521/ (CC0)

“Deep” Semantic Resources (Examples)

TECHNISCHE UNIVERSITÄT DARMSTADT

FrameNet

- English resource based on frame semantics
- http://framenet.icsi.berkeley.edu/

SALSA II – the SAarbrücken Lexical Semantics Acquisition project

- German resource based on frame semantics
- http://www.coli.uni-saarland.de/projects/salsa/

VerbNet

- English verb lexicon based on PropBank semantics
- http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

Multiple multilingual FrameNet versions…

Frame Semantics

TECHNISCHE UNIVERSITÄT DARMSTADT

- Semantic theory initiated by Charles J. Fillmore in the 1970ies
- Model "prototypical situations", their **participants or props** as well as the **role** each participant or prop plays

Frame: COMMERCE_BUY	
Seller	**BMW** **bought** **Rover** from **British Aerospace**.
Buyer	**Rover** was **bought** by **BMW**, which financed […] the new Range Rover.
Goods	**BMW**, which **acquired** **Rover** in 1994, is now dismantling the company.
Money	**BMW**'s **purchase** of **Rover** for **$1.2 billion** was a good move.

Frame Semantics: Terminology

TECHNISCHE UNIVERSITÄT DARMSTADT

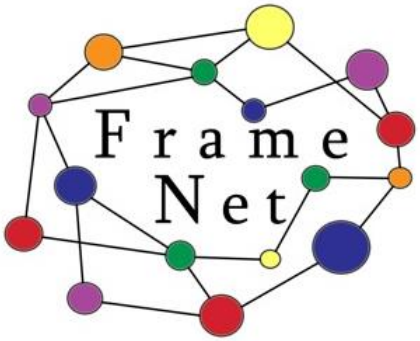
- **Frame**: a script-like description of a type of event, relation, state, or object
  - e.g. **COMMUNICATION_MANNER**
- **Frame Elements (FEs)**: participants in the frame and their **role**
  - **Speaker**: the person producing a message
  - **Addressee**: the person to whom the speaker is communicating
  - **Message**: the content which is communicated by the speaker
- **Lexical units (LUs)**: word senses which evoke a certain frame
  - also called: **frame-evoking elements**
  - e.g. `babble`, `lisp`, `mumble`, `shout`, `sing`, `stutter`, `whisper`,…
- **Frame-to-frame relations**: relationships between frames
  - e.g. **COMMUNICATION_MANNER** inherits from **COMMUNICATION**

FrameNet: Example

TECHNISCHE UNIVERSITÄT DARMSTADT

Communication_manner

Lexical Unit Index



Definition:

The words in this frame describe Manners of verbal communication. All of them can occur with quoted expressions.
He SLURRED his confession.

FEs:

Core:

Addressee [Add]
Semantic Type: Sentient
Addressee is the person to whom the Speaker is communicating. When expressed, the Addressee occurs as a PP Complement:
The taxi driver CHATTERED away to me about gardening.

Message [Msg]
Semantic Type: Message
Message is the content which is communicated by the Speaker. The Message may be a direct quote, a finite complement clause or an NP Object:
"I- It was an accident," Jo STAMMERED.

Jo STAMMERED that it was an accident.

Jo STAMMERED an apology.

Why Frame Semantics?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Normalization of **syntactic alternations**

  [Fred]$_{\text{Agent}}$ **hit**$_{\text{Cause\_Impact}}$ [the ball]$_{\text{Impactee}}$

  [The ball]$_{\text{Impactee}}$ was **hit**$_{\text{Cause\_Impact}}$

  [John]$_{\text{Donor}}$ **gave**$_{\text{Giving}}$ [Mary]$_{\text{Recipient}}$ [a book]$_{\text{Theme}}$

  [John]$_{\text{Donor}}$ **gave**$_{\text{Giving}}$ [a book]$_{\text{Theme}}$ [to Mary]$_{\text{Recipient}}$

- Normalization of **lexical alternations**
  (within and across parts of speech)

  [Marylin]$_{\text{Speaker}}$ **spoke**$_{\text{Statement}}$ about [her past]$_{\text{Topic}}$

  [Marylin]$_{\text{Speaker}}$'s **statement**$_{\text{Statement}}$ about [her past]$_{\text{Topic}}$

  [Marylin]$_{\text{Speaker}}$ **talked**$_{\text{Statement}}$ about [her past]$_{\text{Topic}}$

Frame-Evoking Word Classes

TECHNISCHE UNIVERSITÄT DARMSTADT

Verbs:

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

Predicate-like nouns:

The **development**$_{\text{Product\_development}}$ of [a new mobile phone]$_{\text{Product}}$ …

Predicate-like adjectives:

[That ladder]$_{\text{Entity}}$ is [really]$_{\text{Degree}}$ **tall**$_{\text{Measureable\_attributes}}$

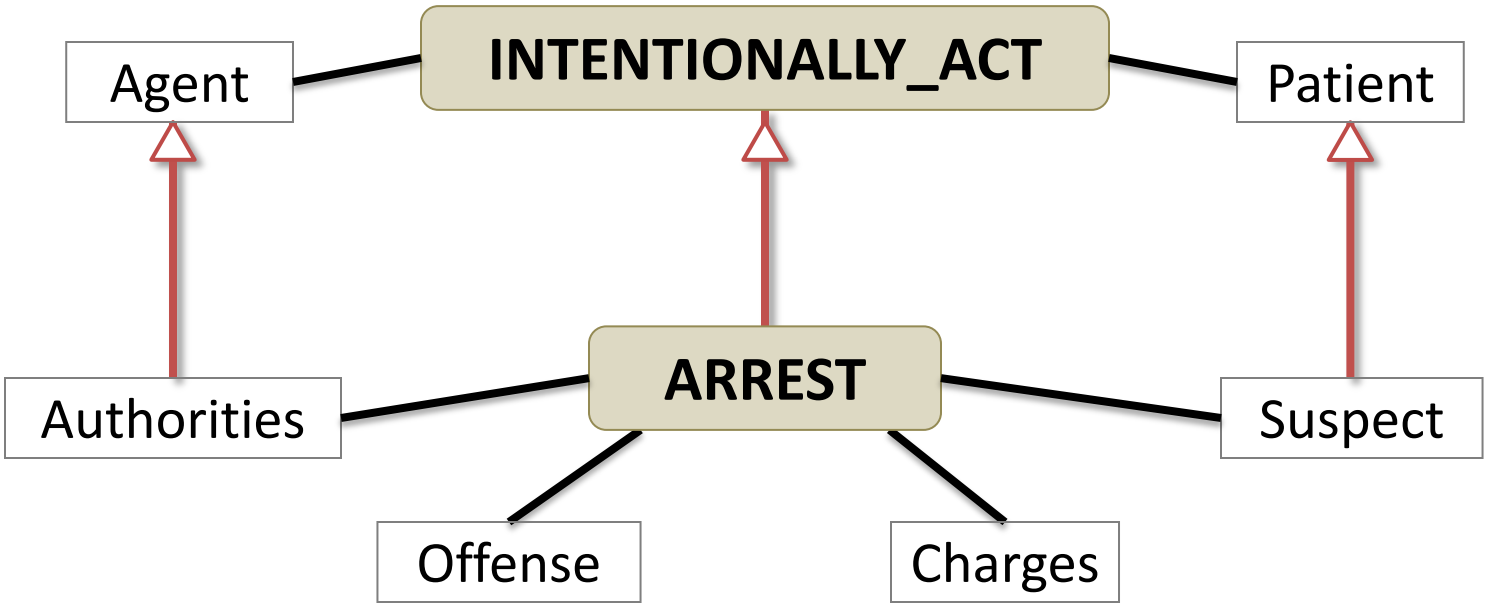
UKP

Frame-to-Frame Relations

TECHNISCHE UNIVERSITÄT DARMSTADT

Inheritance relation

- a frame inherits all frame elements of one or more "super" frame(s)



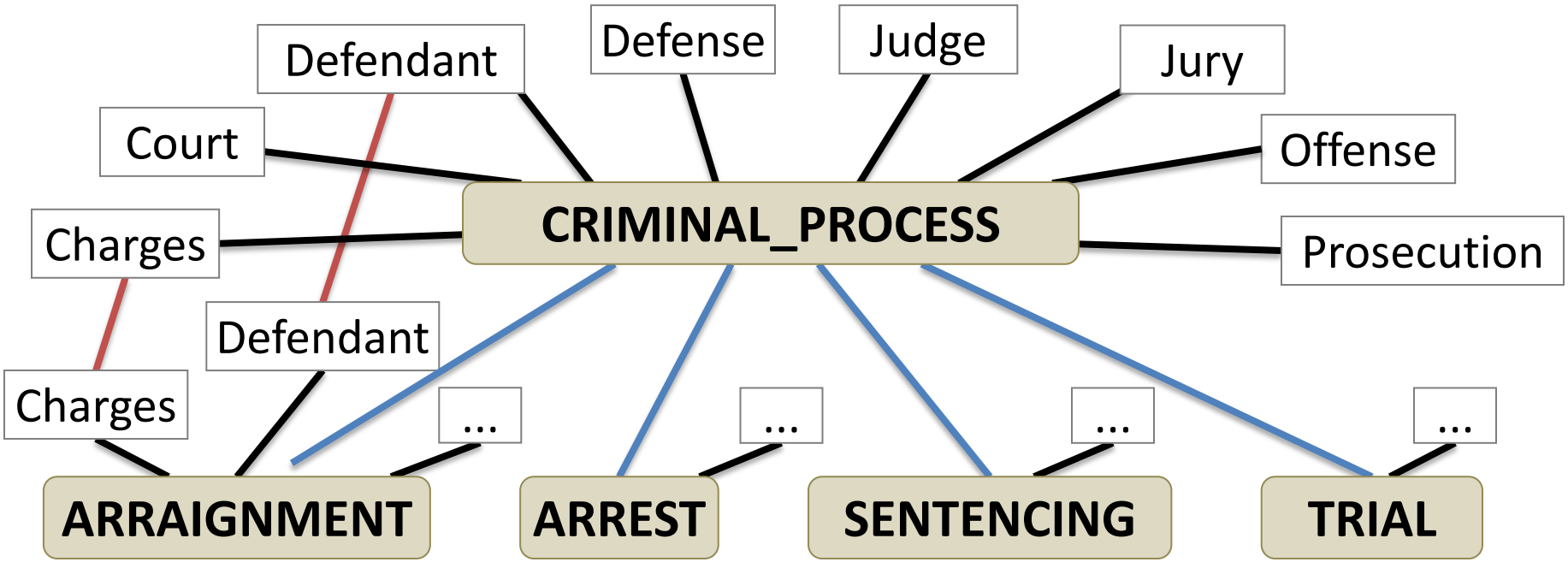
UKP

Frame-to-Frame Relations

TECHNISCHE UNIVERSITÄT DARMSTADT

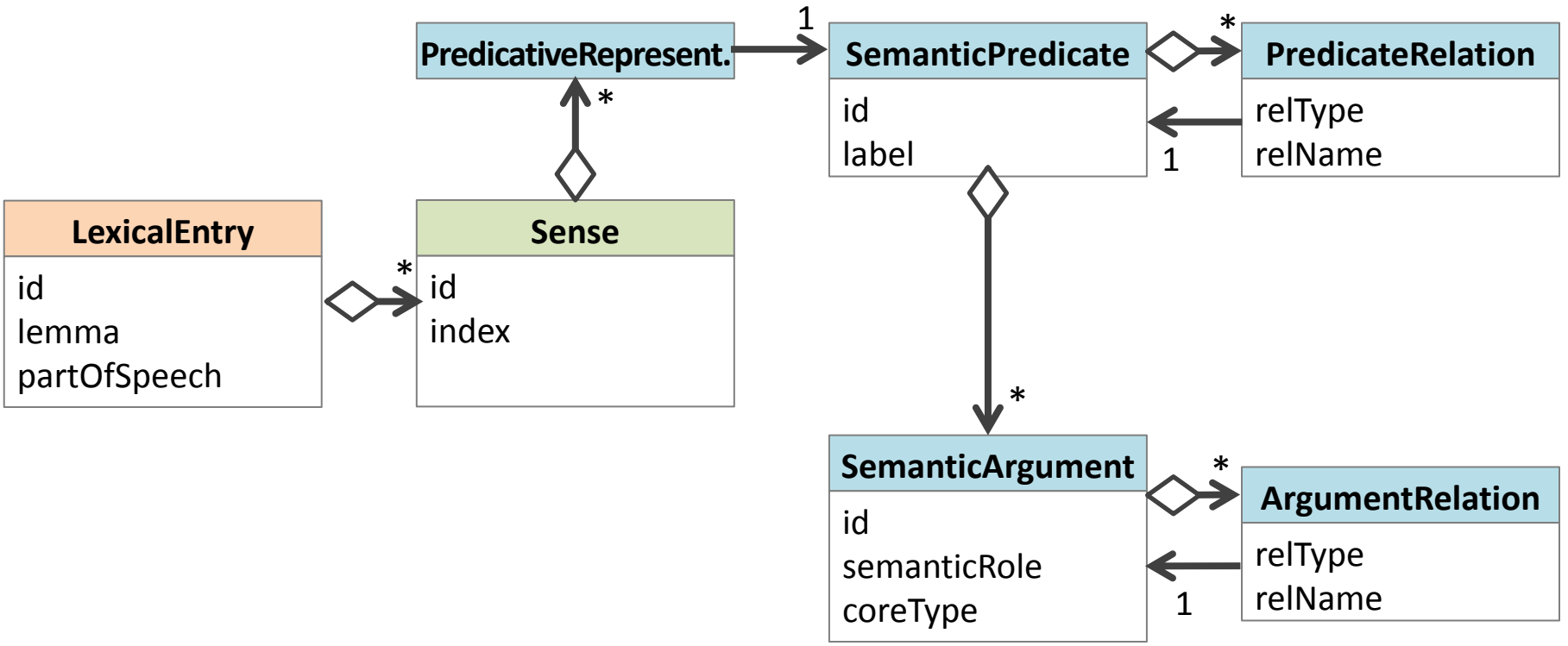
Subframe relation

- Super frame represents complex event
- Subframes usually inherit *some* roles of the super frame

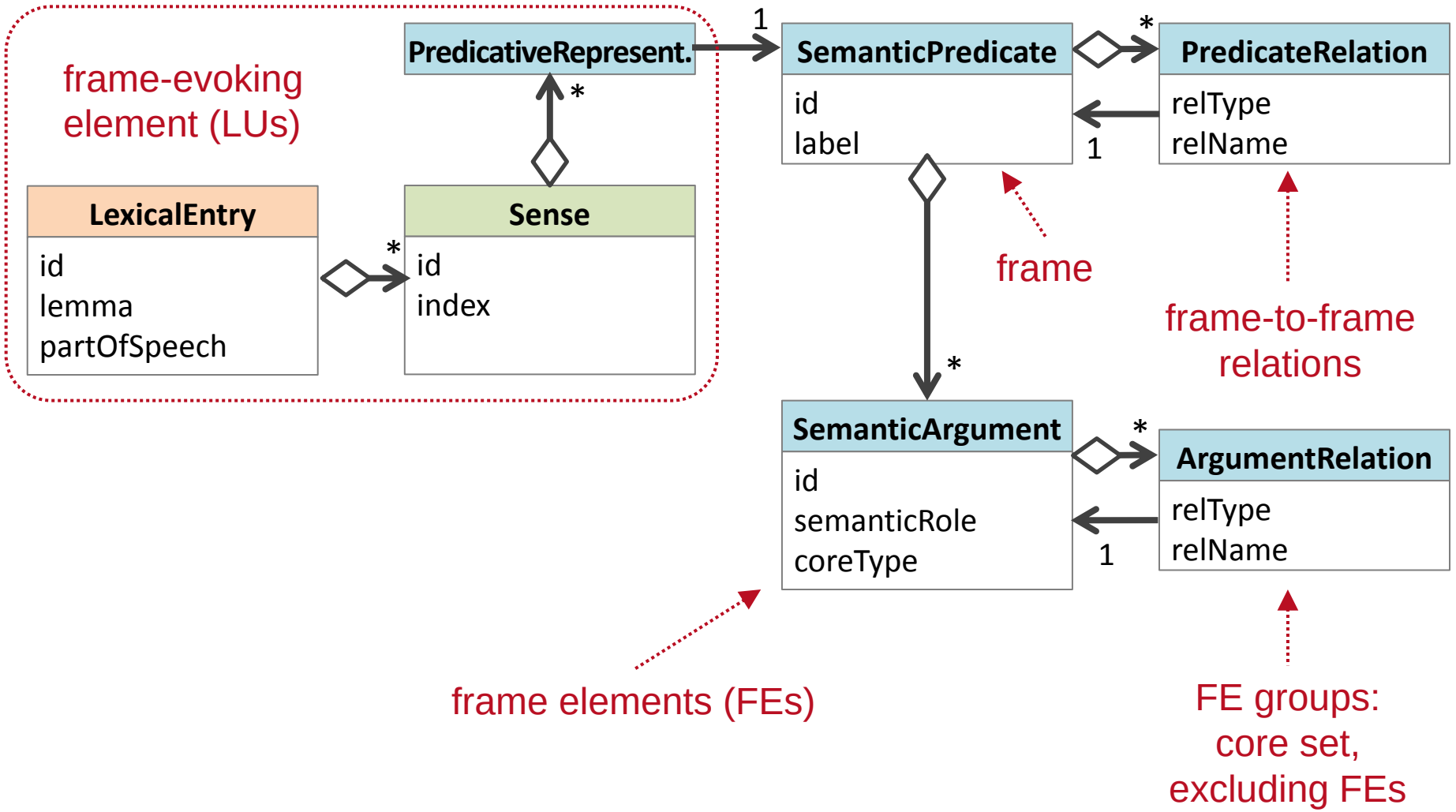


UKP

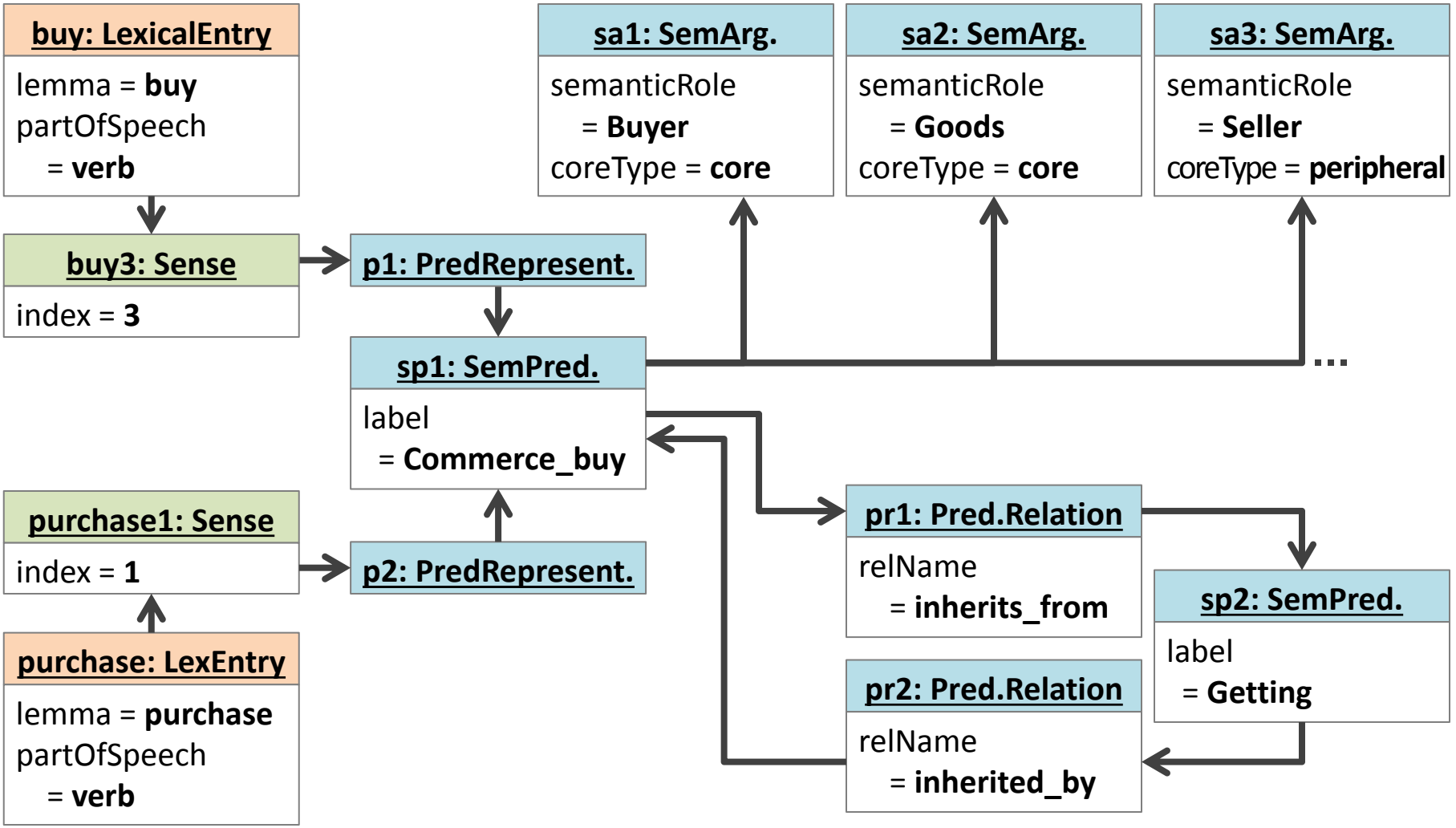
Semantic Predicates: Data Model



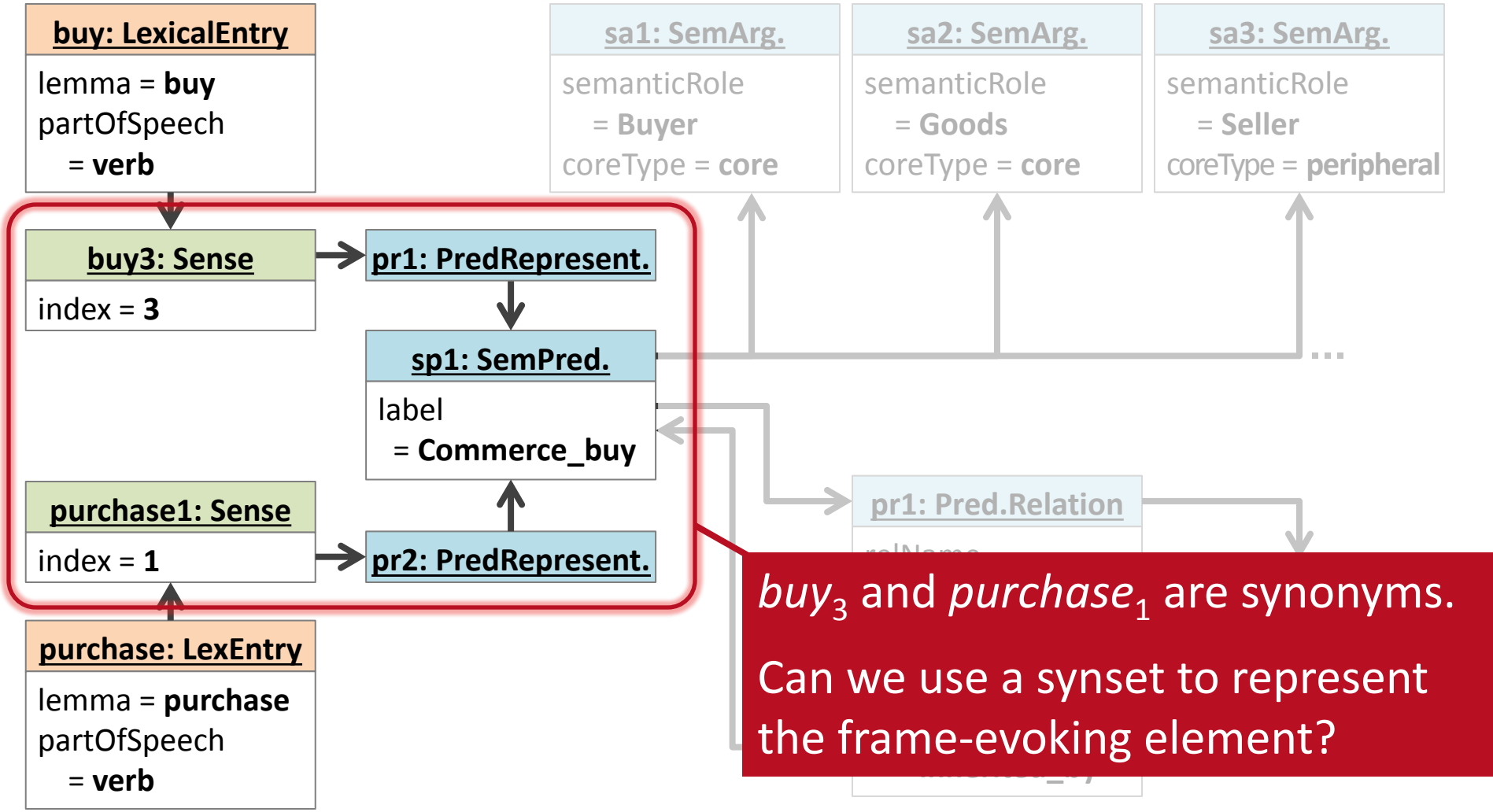
Mapping FrameNet to UBY-LMF



Semantic Predicates: Example



Semantic Predicates: Example



Frame-evoking Elements vs. Synonymy

TECHNISCHE UNIVERSITÄT DARMSTADT

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

- Senses sharing a frame (i.e., the frame-evoking elements) are **semantically related, but <u>not</u> necessarily synonymous!**
  - `love` and `hate` both evoke the **EXPERIENCER_FOCUS** frame, but they are antonyms

Therefore: Synsets are not appropriate to group frame-evoking elements!

Reading Suggestions

- **[Frame semantics]** Ch.J. Fillmore: Frame Semantics and the Nature of Language, in: *Annals of the New York Academy of Sciences 280: Conference on the Origin and Development of Language and Speech*, pp. 20–32, 1976.
- **[FrameNet]** C.F. Baker/Ch.J. Fillmore/J.B. Lowe: The Berkeley FrameNet project, in: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics* (ACL/COLING), pp. 86–90 , 1998. Montreal, Canada.
- **[Multilingual FrameNet]** H.C. Boas: *Multilingual FrameNets in Computational Lexicography: Methods and Applications* (= Trends in Linguistics. Studies and Monographs 2), Berlin: Mouton de Gruyter, 2009.
- **[Multilingual FrameNet]** S. Hartmann/I. 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), pp. 1363–1373, 2013.
- **[SALSA]** A. Burchardt/K. Erk/A. Frank/A. Kowalski/S. Padó/M. Pinkal: The SALSA Corpus: a German Corpus Resource for Lexical Semantics, in: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 969–974, Genoa, Italy.
- **[PropBank and VerbNet]** E. Loper/S. Yi/M. Palmer: *Combining Lexical Resources: Mapping Between PropBank and VerbNet*, in: Proceedings of the 7th International Workshop on Computational Linguistics, 2007. Tilburg, the Netherlands.

Lexical Resources for NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

- Introduction
- Dictionaries
- Wordnets and Thesauri
- Multilingual and Aligned Resources
- *– Break –*
- Deep Semantic Resources
- **Syntactic Resources**
- Lexical Resources in Action
- Wrap-up

Try it!

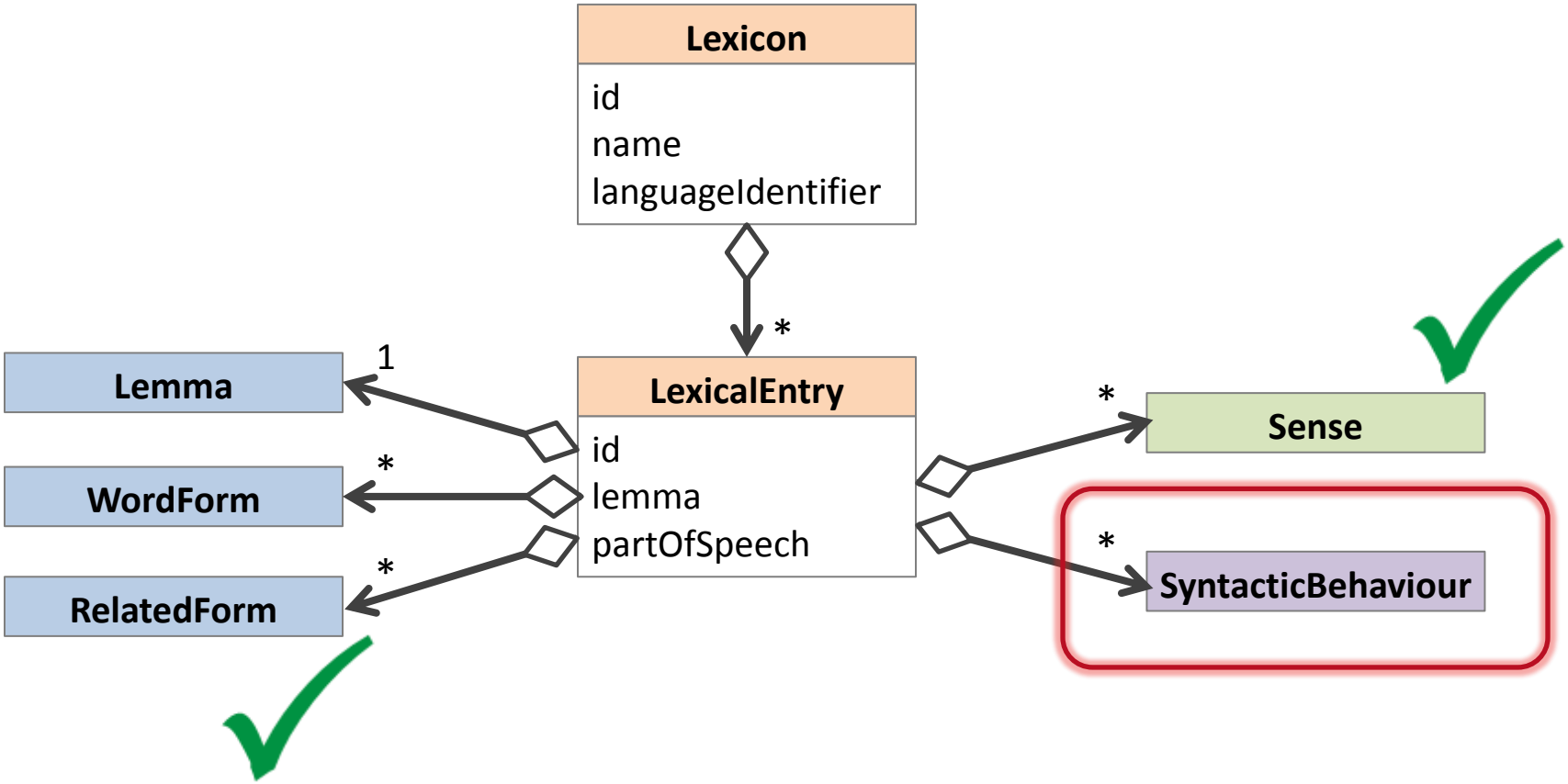
Try it!

Try it!

Try it!

http://pixabay.com/en/coffee-cup-drink-cafe-brown-mug-312521/ (CC0)

Syntax


Lexicon
id
name
languageIdentifier
*
1
Lemma
*
WordForm
*
RelatedForm
LexicalEntry
id
lemma
partOfSpeech
*
Sense
*
SyntacticBehaviour

Syntactic Resources (Examples)

TECHNISCHE UNIVERSITÄT DARMSTADT

VerbNet

- English verb lexicon based on Levin classes
- http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

IMSLex-Subcat

IMSLex-Subcat

- German verb lexicon based on Levin classes
- http://www.logos-verlag.de/cgi-bin/engbuchmid?isbn=301&lng=deu&id=
- http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/IMSLex.html

Few other electronic valency dictionaries…

Subcategorization and Valency

TECHNISCHE UNIVERSITÄT DARMSTADT

The syntactic behavior of lexical entries (mostly: verbs) is described by **subcategorization** or **valency**

- **Subcategorization frames (SCF)**: typical sentence „patterns“
- **Syntactic arguments**: the components of this pattern
- **Syntactic categories**: NP-nominative, NP-accusative, PP-as,…
- **Grammatical functions**: subject, object,…

Intransitive usage:

She [subject, nominative]	`is singing.`

Transitive usage:

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

Subcategorization Frame and Sense

TECHNISCHE UNIVERSITÄT DARMSTADT

Example: (to) sing

Sense 1: "produce tones with the voice"

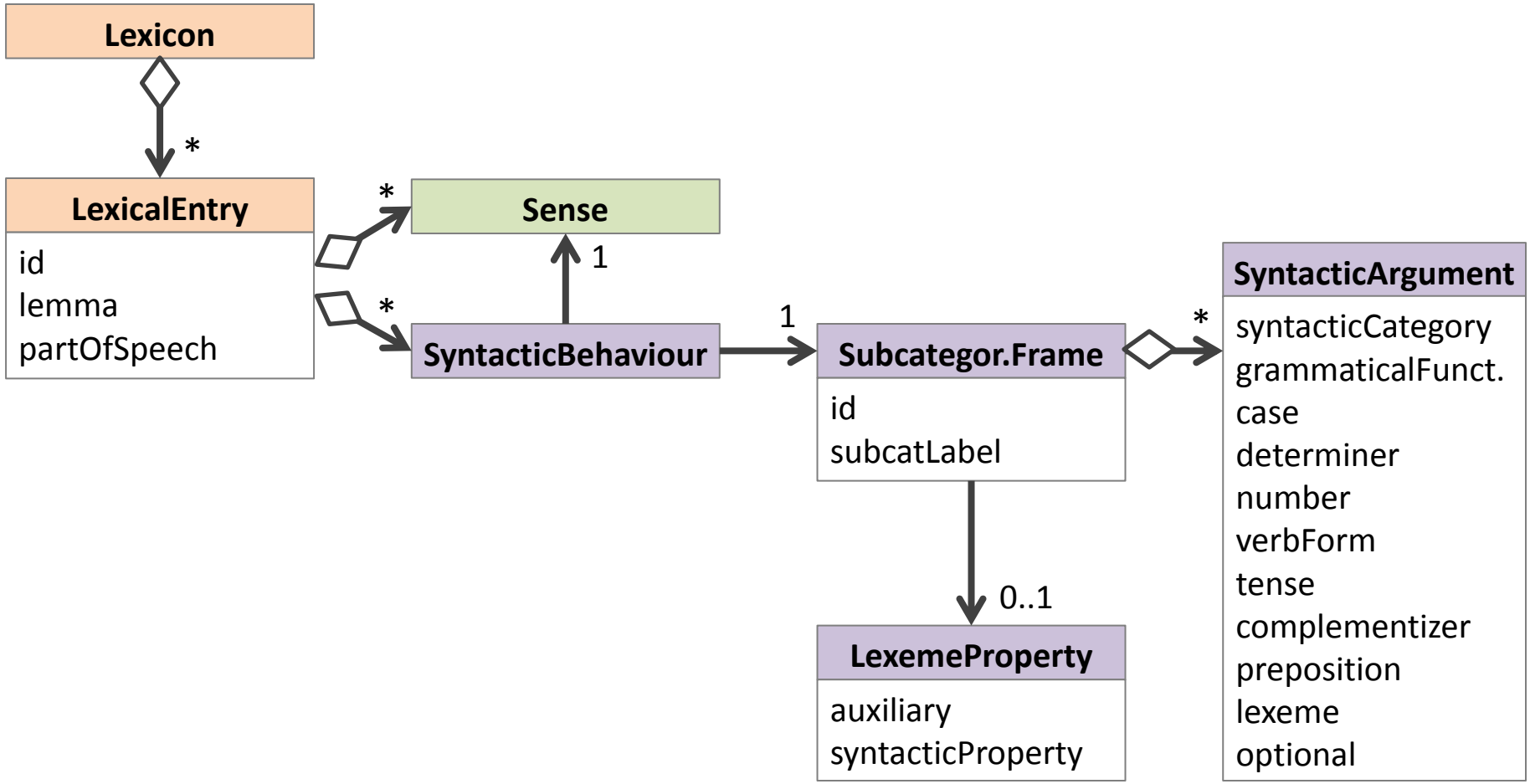
- can be used with and without accusative object:
  - They **sing**.
  - They **sing** Christmas carols.

Sense 2: "divulge confidential information or secrets"

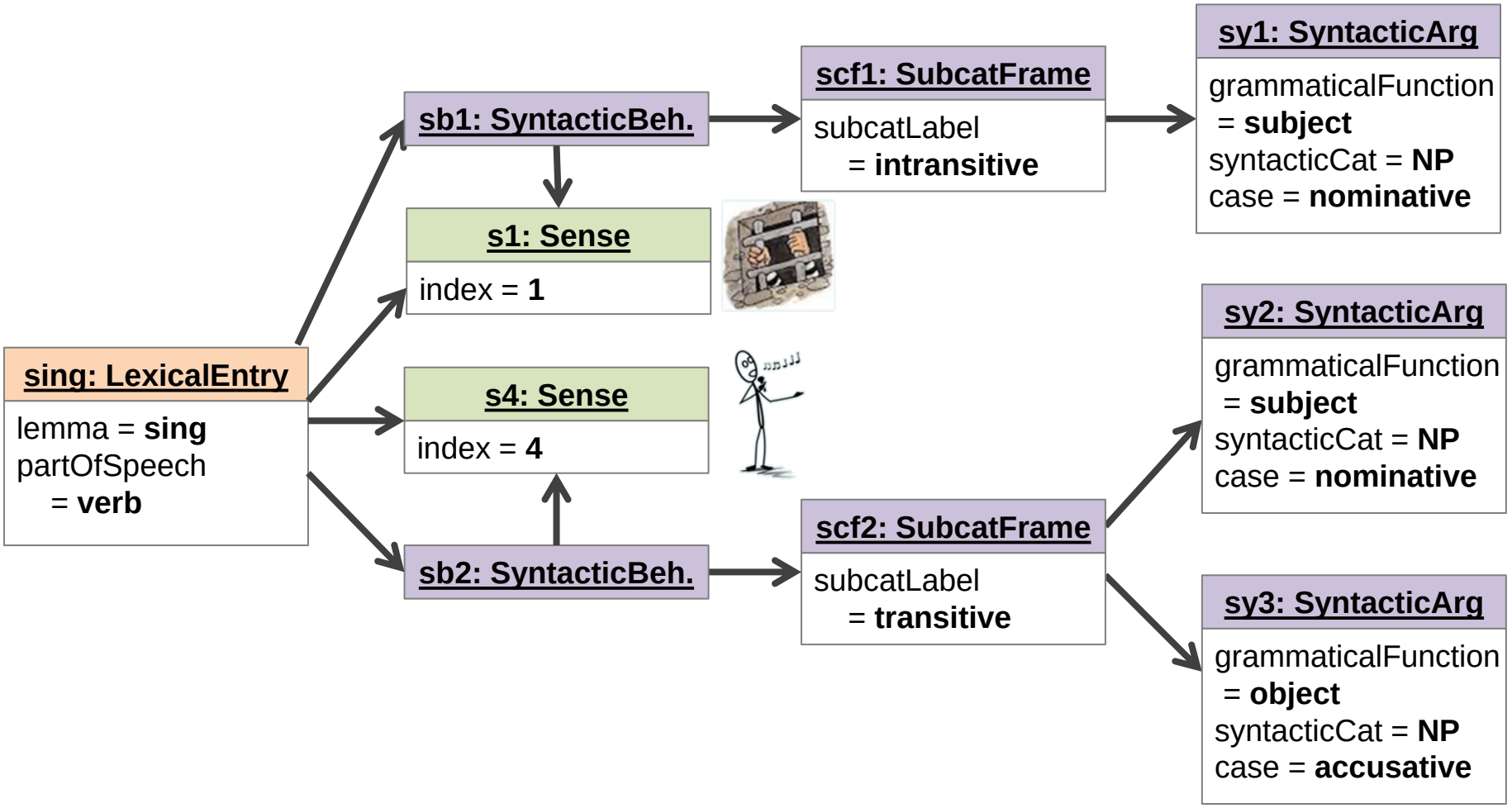
- usually not used with accusative object:
  - The informant will **sing** very soon.
  - ? The informant will **sing** the secrets very soon.

Remember our initial questions? → *(to) sing* is typically used with and without accusative object, depending on the context (there are many other usages…)

Syntax: Data Model



Syntactic Behavior: Example



Levin-Style Verb Classes

TECHNISCHE UNIVERSITÄT DARMSTADT

Syntactic alternation (e.g., dative alternation)

- [Martha]$_{NOM}$ **gave** [an apple]$_{ACC}$ [to Myrna]$_{PP}$.
- [Martha]$_{NOM}$ **gave** [Myrna]$_{DAT}$ [an apple]$_{ACC}$.
- Verbs taking part in this alternation share particular meaning components, e.g., change of possession verbs like *give* and *sell*

Levin's hypothesis

- Verbs with **similar syntactic alternation behavior** share common semantic properties ("**meaning components**")
  → **Verb meaning and verb syntax correspond**

UKP

Levin-Style Verb Classes

TECHNISCHE UNIVERSITÄT DARMSTADT

Levin's verb classes group verbs that share the **same predicate-argument structure**, i.e.,

- **subcategorization frames** and syntactic argument alternations
- **semantic roles** and **selectional preferences**
- **semantic predicate** based on the event decomposition (Moens and Steedman, 1988)

VerbNet

- electronic lexicon grouping verb by their verb class
- roughly 4,000 English verbs

VerbNet: Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

Levin class ID

Subclass relations

GO TO COMMENTS

get-13.5.1

Members: 25, Frames: 7

POST COMMENT

CLASS HIERARCHY

GET-13.5.1*

GET-13.5.1-1

Verbs sharing the same predicate-argument structure

MEMBERS

KEY

ATTAIN (G 1)	CONSERVE (WN 4; G 1)	PICK (FN 1; WN 2; G 2)	SHOOT (FN 1, 2, 3; WN 2; G 1)
BOOK (WN 2; G 1)	FIND (FN 1, 2; WN 3, 7; G 1)	PLUCK (FN 1, 2; WN 1, 6; G 1)	SLAUGHTER (FN 1; WN 1; G 1)
BUY (FN 1; WN 1, 3; G 1)	GATHER (FN 1, 2; WN 1, 6; G 1, 2)	PROCURE (WN 1, 2)	VOTE (WN 5; G 5)
CALL (FN 1, 2, 3, 4; WN 5, 23; G 1, 7)	HIRE (FN 1; WN 1, 2, 3; G 1)	PULL (FN 1, 2, 3; WN 2, 6, 17; G 1, 3)	WIN (WN 2; G 2)
CATCH (WN 4, 5, 8; G 1, 2)	LEASE (FN 1, 2; WN 2, 4; G 1)	REACH (WN 8; G 3)	
CHARTER (FN 1; WN 3)	ORDER (FN 1; WN 2; G 2)	RENT (FN 1; WN 3, 4; G 2)	
CHOOSE (FN 1; WN 1, 2; G 1)	PHONE (WN 1; G 1)	RESERVE (WN 3, 4; G 2)	

Alignments

ROLES

REF

- AGENT [+ANIMATE | +ORGANIZATION]
- THEME
- SOURCE [+CONCRETE]
- BENEFICIARY [+ANIMATE | +ORGANIZATION]
- ASSET [-LOCATION & -REGION]

Thematic roles

Selectional restrictions

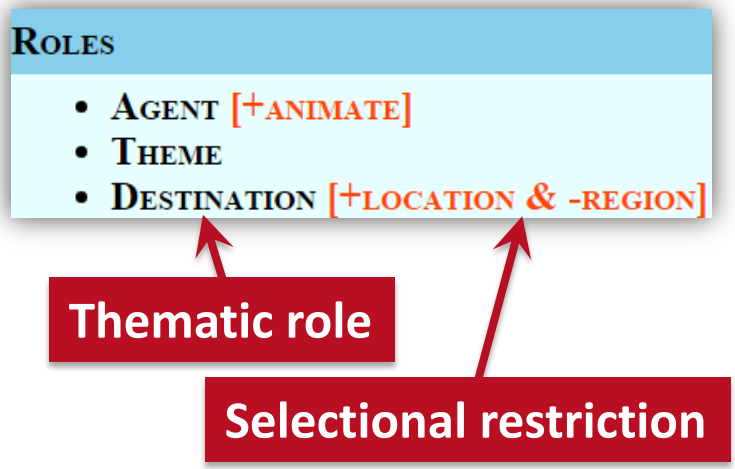
UKP

VerbNet: Roles

TECHNISCHE UNIVERSITÄT DARMSTADT

Thematic roles

- Small set of roles used across all classes (≠ fine-grained FrameNet roles)
- [Sandy]$_{\text{AGENT}}$ **shattered** [the glass]$_{\text{PATIENT}}$.
- [The glass]$_{\text{PATIENT}}$ **shattered**.



Selectional restrictions

- Constraint to limit the type of the "filler" of a role
- Existence (+animate) vs. absence (–animate)
- Basic logical operators OR (*a* | *b*) and AND (*a* & *b*)

UKP

VerbNet: Syntax and Semantics

TECHNISCHE UNIVERSITÄT DARMSTADT

Subcategorization frame

Example sentence

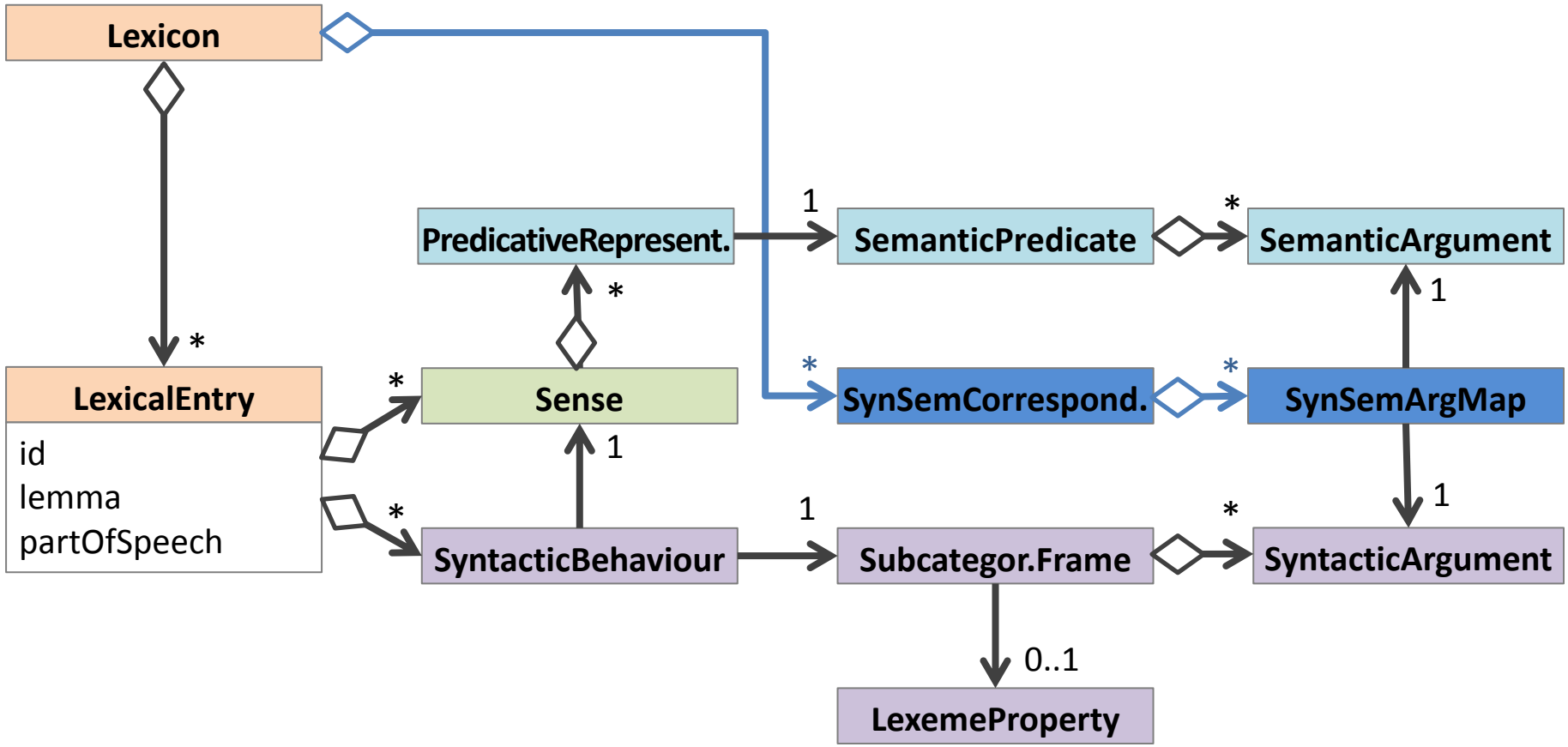
Syntax/Semantic interface

Semantic representation

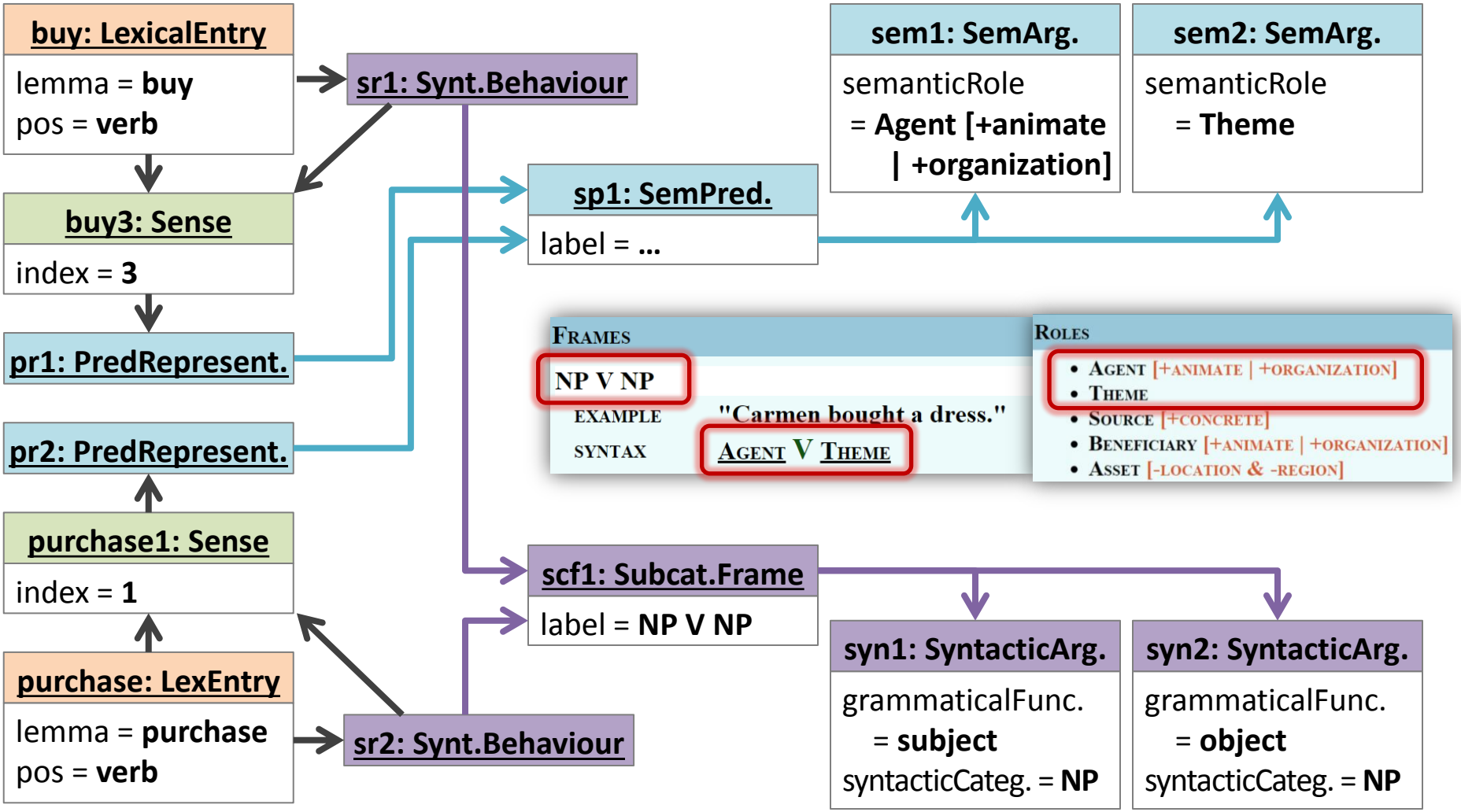
FRAMES	REF KEY
NP V NP	
EXAMPLE	"Carmen bought a dress."
SYNTAX	AGENT V THEME
SEMANTICS	HAS_POSSESSION(START(E), ?SOURCE, THEME) TRANSFER(DURING(E), THEME) HAS_POSSESSION(END(E), AGENT, THEME) CAUSE(AGENT, E)
NP V NP PP.SOURCE	
EXAMPLE	"Carmen bought a dress from Diana."
SYNTAX	AGENT V THEME {FROM} SOURCE
SEMANTICS	HAS_POSSESSION(START(E), SOURCE, THEME) TRANSFER(DURING(E), THEME) HAS_POSSESSION(END(E), AGENT, THEME) CAUSE(AGENT, E)
NP V NP PP.BENEFICIARY	
EXAMPLE	"Carmen bought a dress for Mary."
SYNTAX	AGENT V THEME {FOR} BENEFICIARY
SEMANTICS	HAS_POSSESSION(START(E), ?SOURCE, THEME) TRANSFER(DURING(E), THEME) HAS_POSSESSION(END(E), AGENT, THEME) CAUSE(AGENT, E) BENEFIT(E, BENEFICIARY)
NP V NP.BENEFICIARY NP	
EXAMPLE	"Carmen bought Mary a dress."
SYNTAX	AGENT V BENEFICIARY THEME
SEMANTICS	HAS_POSSESSION(START(E), ?SOURCE, THEME) TRANSFER(DURING(E), THEME) HAS_POSSESSION(END(E), AGENT, THEME) CAUSE(AGENT, E) BENEFIT(E, BENEFICIARY)

UKP

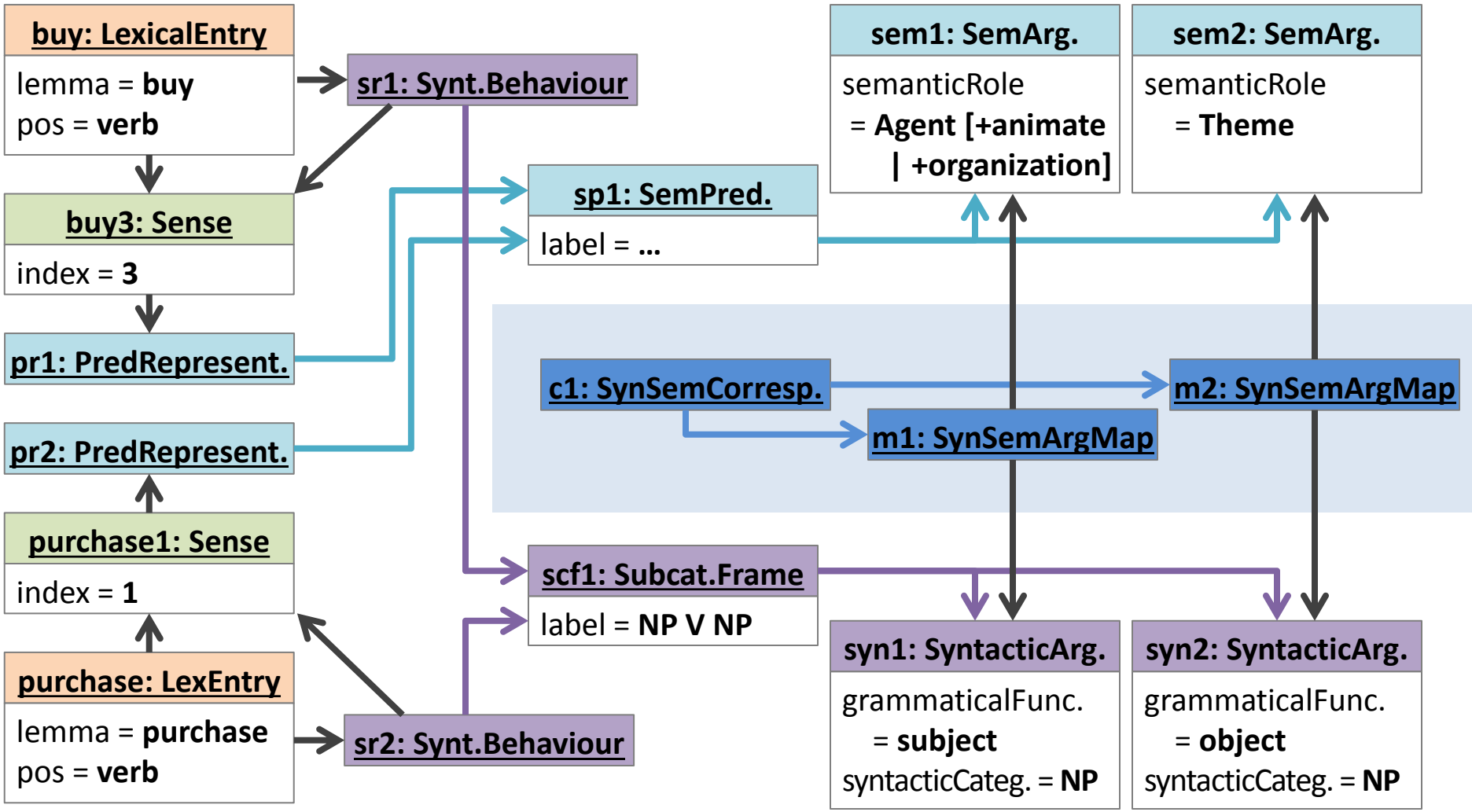
Syntax/Semantics Interface: Model



Syntax/Semantics Interface: Example



Syntax/Semantics Interface: Example



Different Sense Groupings

TECHNISCHE UNIVERSITÄT DARMSTADT

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

VerbNet verb class
manner_speaking-37.3
103 members, 14 frames

sing

WordNet synset
01734912-v
1 member

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

FrameNet frame
Communication_manner
24 members

UKP

Try it yourself! – Assignment 3

TECHNISCHE UNIVERSITÄT DARMSTADT

Open `org.dkpro.uby.examples.Assignment3`

1) Find all semantic arguments of *(to) read* in FrameNet
2) Explore the subcategorization frames of *(to) sing* in VerbNet
3) Compare the syntactic and semantic arguments of *(to) sing* in VerbNet and WordNet

10 minutes

http://pixabay.com/en/stopwatch-racing-timer-timing-153398/ (CC0)

Reading Suggestions

- **[Verb classes]** B. Levin: *English Verb Classes and Alternations: A Preliminary Investigation*, Chicago/London: The University of Chicago Press, 1993.
- **[Multilingual Verb classes]** B. Levin: *Verb classes within and across languages*, chapter 39 in A. Malchukov/B. Comrie (Eds.): Valency Classes in the World's Languages, Volume 2: Case Studies from Austronesia, the Pacific, the Americas, and Theoretical Outlook, pp. 1627–1670, Berlin/New York: De Gruyter, 2015.
- **[VerbNet]** K. Kipper Schuler: VerbNet: a broad-coverage, comprehensive verb lexicon, Dissertation, University of Pennsylvania, 2005. http://verbs.colorado.edu/~kipper/
- **[VerbNet]** K. Kipper/A. Korhonen/N. Ryant/M. Palmer: A Large-scale Classification of English Verbs, *Language Resources and Evaluation Journal* 42(1): 21–40, 2008.
- **[IMSLex-Subcat]** J. Eckle-Kohler: *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora*, Berlin: Logos-Verlag, 1999.
- **[Subcat-LMF]** J. Eckle-Kohler/I. Gurevych: Subcat-LMF: Fleshing out a standardized format for subcategorization frame interoperability, in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), pp. 550–560, 2012.
- **[Syntax/Semantics]** J. Eckle-Kohler/I. Gurevych/S. Hartmann/M. Matuschek/Ch.M. Meyer: UBY-LMF – Exploring the Boundaries of Language-Independent Lexicon Models, chapter 10 in G. Francopoulo (Ed.): *LMF: Lexical Markup Framework*, pp. 145–156, London: Wiley-ISTE, 2013.

Lexical Resources for NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

- Introduction
- Dictionaries
- Wordnets and Thesauri
- Multilingual and Aligned Resources
- *– Break –*
- Deep Semantic Resources
- Syntactic Resources
- **Lexical Resources in Action**
- Wrap-up

Try it!

Try it!

Try it!

Try it!

http://pixabay.com/en/coffee-cup-drink-cafe-brown-mug-312521/ (CC0)

Lexical Resources in Action

TECHNISCHE UNIVERSITÄT DARMSTADT

Speech recognition
- Phonetic representations

Machine translation
- Word and phrase translations

Lexical substitution
- Synonyms

Word relatedness
- Semantic relations

Word sense disambiguation
- Sense definition, semantic labels

Natural language generation
- Inflected word forms

Grammar correction
- Syntactic frames

Question answering
- Semantic predicates & arguments

Text classification/understanding
- Different types of information as ML features

Many other tasks possible…

General Approaches

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Use/Extract knowledge from a lexical resource

- Pairs of lemma and POS tag as lexicon for a tagger
- List of word/phrase translations as an SMT background lexicon
- Semantic network to compute the relatedness of word pairs

2. Enrich/Annotate a text with lexical information

- Query tokens or larger chunks of a text in a lexical resource
  - e.g., identify synonyms for each token
- Usually requires preprocessing

Preprocessing

TECHNISCHE UNIVERSITÄT DARMSTADT

- Usually essential when querying a lexical resource
- Typically: Lemmatization, POS tagging
- But also: Multi-word identification, parsing or chunking, named entity recognition

Input	Allows Query
Token	→ WordForm
Lemmatized token + POS	→ LexicalEntry
Sense-disambiguated token	→ LexicalEntry + Sense
Syntactic parse	→ SubcategorizationFrame
Semantic parse	→ SemanticPredicate

DKPro Core

TECHNISCHE UNIVERSITÄT DARMSTADT

- A collection of software components for NLP based on the Apache UIMA framework
- Multiple components for preprocessing and linguistic annotation
- Integrates multiple freely available tools, such as:
  - TreeTagger
  - OpenNLP
  - Stanford NLP
  - LanguageTool
  - Mate Tools
- Multiple data formats (text, PDF, XML,…)

http://www.dkpro.org

Part of Speech Tags

TECHNISCHE UNIVERSITÄT DARMSTADT

- abbreviation
- abbreviationAcronym
- abbreviationInitialism
- adjective
- adverb
- adverbPronominal
- adpositionPreposition
- adpositionPostposition
- adpositionCircumposition
- affix
- affixPrefix
- affixSuffix
- contraction
- determiner
- determinerDefinite
- determinerPossessive
- determinerIndefinite
- determinerDemonstrative
- determinerInterrogative
- numeral
- interjection
- phraseme
- conjunction
- conjunctionCoordinating
- conjunctionSubordinating
- noun
- nounCommon
- nounProper
- nounProperFirstName
- nounProperLastName
- pronoun
- pronounPersonal
- pronounPossessive
- pronounDemonstrative
- pronounRelative
- pronounIndefinite
- pronounPersonalReflexive
- pronounPersonalIrreflexive
- pronounInterrogative
- particle
- particleNegative
- particleInfinitive
- particleComparative
- particleAnswer
- symbol
- verb
- verbAuxiliary
- verbModal
- verbMain

Part of Speech Tags

TECHNISCHE UNIVERSITÄT DARMSTADT

- abbreviation
- determinerDemonstrative
- pronounRelative
- adpositionPreposition
- conjunctionCoordinating
- particleNegative
- affixPrefix
- affixSuffix
- contraction
- nounProper
- **nounProperFirstName**
- nounProperLastName
- symbol
- verb
- verbAuxiliary

Why *nounProperFirstName*?

- Some resources tag "Joe" as a *noun*
- Some as *proper noun*
- Some as *first name*

Prefix notation allows queries at different granularities:

`noun` vs. `noun*` and `nounProper` vs. `nounProper*` …

But: Mapping between analysis tool and resource required!

e.g., DKPro Core to UBY: `de.tudarmstadt.ukp.uby.resource.UbyResourceUtils.corePosToUbyPos(String)`

UKP

Word Sense Disambiguation

TECHNISCHE UNIVERSITÄT DARMSTADT

Ambiguity for a Computer

- *The thief of last Friday's robbery sat on the **bank** of the Ruhr and counted his money.*

?

bank1: Sense
index = **1**

bank2: Sense
index = **2**

https://www.flickr.com/photos/68751915@N05/6629034769/in/photostream/ (Flickr user 401(K) 2012, CC-BY-SA)

UKP

WSD Strategies

TECHNISCHE UNIVERSITÄT DARMSTADT

Greedy

- Retrieve all senses of these entries

Heuristics

- Most frequent sense (e.g., WordNet); first sense (e.g., Wiktionary)
- Domain-specific sense (e.g., prefer senses with label *chemistry*)
- Synonyms occur in the text
- Filter by syntactic structure of the context (e.g., verb occurs with to-infinitive)

More advanced approaches

- Lesk's algorithm (Lesk, 1986)
- Topic-sensitive PageRank (Agirre and Soroa, 2009)
- Graph connectivity (Navigli and Lapata, 2010)
- …

Simplified Lesk

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Retrieve all **sense definitions** of the word to be disambiguated
2. Count the **words overlapping** between the context and each sense definition
3. Choose the sense that yields the **highest overlap**

How about this example?

*The thief of last Friday's robbery sat on the **bank** of the Ruhr and counted his money.*

bank$_1$ sloping land (especially the slope beside a body of water)

bank$_2$ a financial institution that accepts deposits and channels the money into lending activities

WSD is difficult… ☹

Super Sense Tagging

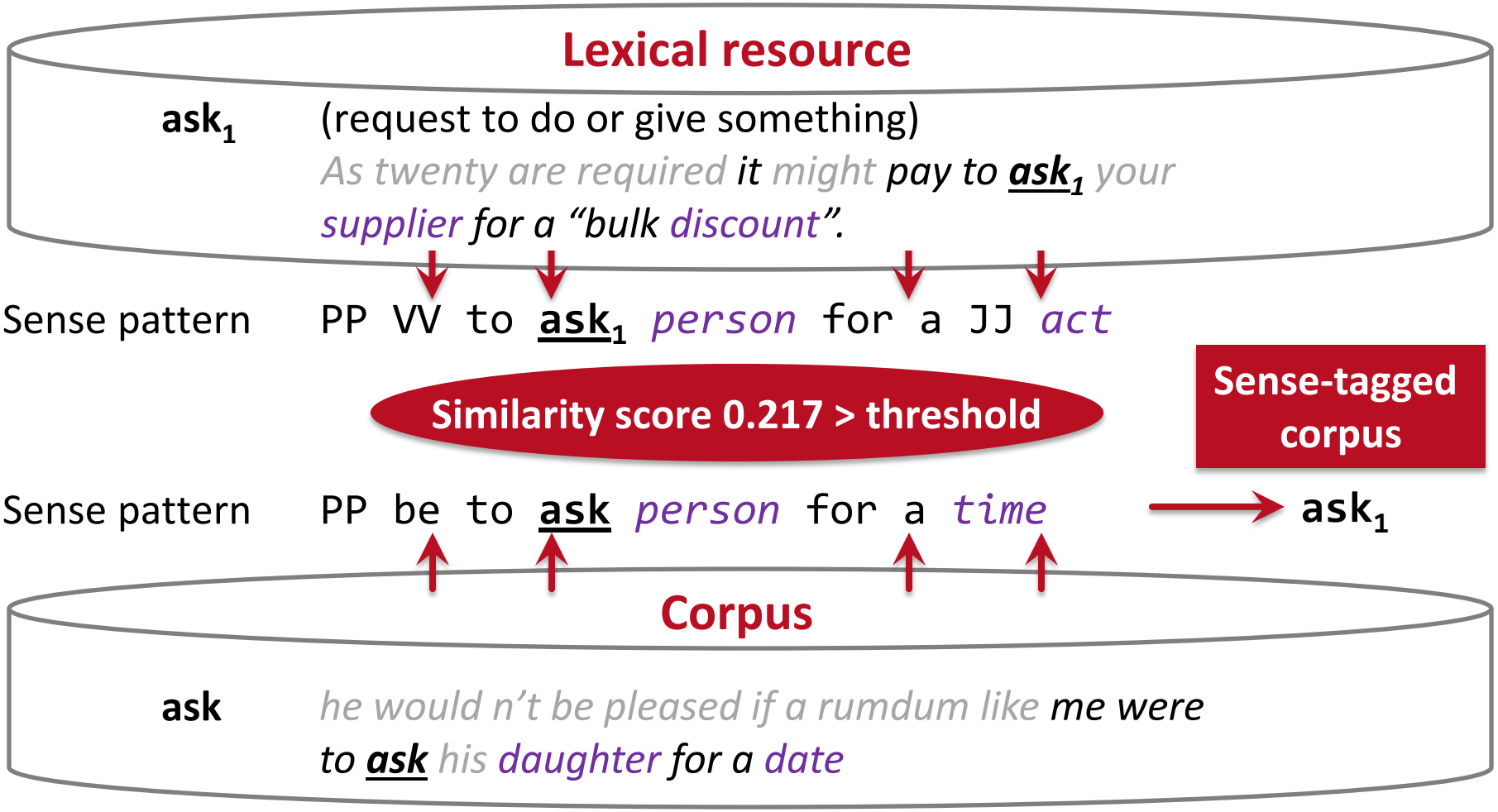
TECHNISCHE UNIVERSITÄT DARMSTADT

- Instead of a fine-grained inventory of senses, limit the sense inventory to a **few super senses**
- Typically: **WordNet lexicographer file names** (aka. semantic fields)
- Multiple other possibilities, e.g., **Wiktionary domain labels**

[Clara Harris]$_{\text{n.person}}$, one of the [guests]$_{\text{n.person}}$ in the [box]$_{\text{n.artifact}}$, [stood up]$_{\text{v.motion}}$ and [demanded]$_{\text{v.communication}}$ [water]$_{\text{n.substance}}$

(Ciaramita and Altun, 2006)

Verb Sense Labeling



(Cholakov, Eckle-Kohler, Gurevych, 2014)

Lexical Expansion

TECHNISCHE UNIVERSITÄT DARMSTADT

Lexical expansion can helps to **bridge the lexical gap** between texts

- Lexical chains for document summarization
- IR or question answering systems
- Recognizing textual entailment
- …

Approach

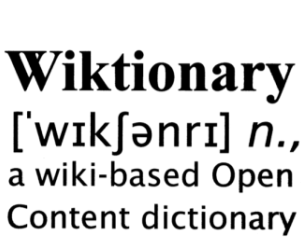
1. Apply a WSD strategy
2. Retrieve synonyms from SenseRelation and Synsets
3. Retrieve linked senses
4. Repeat steps 2 and 3 until stopping criterion is fulfilled

Enriched Sense Representation

TECHNISCHE UNIVERSITÄT DARMSTADT

Linked resources such as UBY allow for combining information types from heterogeneous resources

- Synonyms
- Subsumption hierarchy
- Sense gloss
- Example sentence
- …



- Pronunciation
- Word translations
- Inflected forms
- Etymology
- …

- Semantic predicates
- …

- Subcat frames
- …

UKP

Outlook: Deep Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Embeddings

- Low-dimensional vector representations of atomic symbols
- Typically: word embeddings

Structured embeddings (Bordes et al., 2011)

- Leverage the structure encoded in knowledge resources into statistical learning systems
- Use neural networks to represent the entries and relations of a lexical resource in a low-dimensional (≈50?) space
- Multiple relation types capturing semantics (e.g., hypernym-of), syntax (e.g., subject-of), morphology (e.g., perfect-of),…
- Generalize relations – allows queries such as $f(\texttt{car}_1$, has-part, ?)

Reading Suggestions

TECHNISCHE UNIVERSITÄT DARMSTADT

- **[DKPro Core]** R: Eckart de Castilho/I. Gurevych: A broad-coverage collection of portable NLP components for building shareable analysis pipelines, in *Proceedings of the COLING Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pp. 1–11, 2014. Dublin, Ireland.
- **[WSD]** M. Lesk: Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, in: *Proceedings of the 5th Annual International Conference on Systems Documentation* (SIGDOC), pp. 24–26, 1986. Toronto, Canada.
- **[WSD]** E. Agirre/A. Soroa: Personalizing PageRank for Word Sense Disambiguation, in: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), pp. 33–41, 2009. Athens, Greece.
- **[WSD]** R. Navigli/M. Lapata: An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (4): 678–692, 2010.
- **[Super sense tagging]** M. Ciaramita/Y. Altun: Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), pp. 594–602, 2006. Sydney, Australia.
- **[Verb sense labeling]** K. Cholakov/J. Eckle-Kohler/I. Gurevych: Automated Verb Sense Labelling Based on Linked Lexical Resources, in: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), pp. 68–77, 2014. Gothenburg, Sweden.

Reading Suggestions

TECHNISCHE UNIVERSITÄT DARMSTADT

- **[Deep learning]** A. Bordes/J. Weston/R. Collobert/Y. Bengio: Learning Structured Embeddings of Knowledge Bases, in: *Proceedings of the 25th Conference on Artificial Intelligence* (AAAI), pp. 301–306, 2011. San Francisco, CA, USA.
- **[Word relatedness]** A, Budanitsky/G. Hirst: Evaluating WordNet-based Measures of Lexical Semantic Relatedness, *Computational Linguistics* 32 (1): 13–47, 2006.
- **[Question answering]** D. Shen/M. Lapata: Using Semantic Roles to Improve Question Answering, in: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL), pp. 12–21, 2007. Prague, Czech.
- **[Machine translation]** L.T. Lim: Multilingual Lexicons for Machine Translation, in: Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services (iiWAS), pp. 734–738, 2009. Kuala Lumpur, Malaysia.
- **[Speech recognition]** M. Adda-Decker/L. Lamel: The Use of Lexica in Automatic Speech Recognition, in: Lexicon Development for Speech and Language Processing (= Text, Speech and Language Technology), pp 235–266, Amsterdam: Kluwer, 2000.
- **[Query expansion]** J. Zhang/B. Deng/X. Li: Concept Based Query Expansion Using WordNet, in: Proceedings of the International e-Conference on Advanced Science and Technology (AST), pp. 52–55, 2009. Daejeon, Korea.

Lexical Resources for NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

- Introduction
- Dictionaries
- Wordnets and Thesauri
- Multilingual and Aligned Resources

– Break –

- Deep Semantic Resources
- Syntactic Resources
- Lexical Resources in Action
- **Wrap-up**

Try it!

Try it!

Try it!

Try it!

http://pixabay.com/en/coffee-cup-drink-cafe-brown-mug-312521/ (CC0)

Lexical Resources are…

TECHNISCHE UNIVERSITÄT DARMSTADT

- different from **corpora**
- not limited to the **Princeton WordNet**
- very **heterogeneous**
  - expert vs. collaboratively created
  - different information types (e.g., focus on syntax)
  - fine-grained vs. coarse-grained

  → harmonizing and inter-linking approaches
- helpful for a **variety of tasks**
- getting a bit unpopular recently, although there's **much potential**
  - e.g., upcoming trends of **structured embeddings**
- **what WE make of them!**

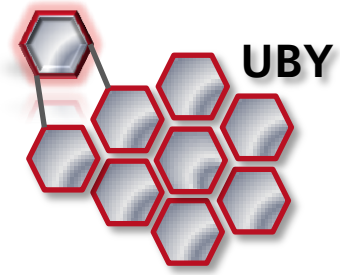
Obtaining Lexical Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

- **SIGLEX Online resources catalog**
  http://www.clres.com/siglex/ressearch.php
- **LRE map**
  http://www.resourcebook.eu/searchll.php
- **META-NET/META-SHARE**
  http://www.meta-share.eu/
- **Computerlinguistik.org Ressourcenkatalog**
  http://www.computerlinguistik.org/portal/portal.html?s=Ressourcen
- **OBELEX^{dict}**
  http://www.owid.de/obelex/dict
- **German online dictionaries**
  https://lexikographieblog.wordpress.com/deutsche-worterbucher-online/

Obtaining UBY

TECHNISCHE UNIVERSITÄT DARMSTADT



Web

- https://www.ukp.tu-darmstadt.de/uby
- https://dkpro.github.io/dkpro-uby/

Source and database dumps

- https://github.com/dkpro/dkpro-uby
- http://uby.ukp.informatik.tu-darmstadt.de/uby

Web interface

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

Mailing list and publications

- https://groups.google.com/forum/#!forum/uby-users
- https://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/publications

Acknowledgments

This tutorial would not have been possible without the help of the UBY team:

Yevgen Chebotar, Kostadin Cholakov, Richard Eckart de Castilho, Judith Eckle-Kohler, Iryna Gurevych, Than-Le Ha, Silvana Hartmann, Mohamed Khemakhem, Masoud Kiaeeha, Zijad Maksuti, Michael Matuschek, Tristan Miller, Tri-Duc Nghiem, Daniil Sorokin, Christian Wirth

Thank you for your attention!

(There's one more "try it" on the next slide in case you're not yet starving…)

Try it yourself! – Assignment 4

TECHNISCHE UNIVERSITÄT DARMSTADT

The code examples provide three *very simple* experiments to see UBY in action:

1) Run `org.dkpro.uby.examples.` **`Assignment4_SupersenseTagging`**
2) Run `org.dkpro.uby.examples.` **`Assignment4_WordSenseDisambiguation`**
3) Run `org.dkpro.uby.examples.` **`Assignment4_LexicalExpansion`**

The first run takes some time to download the preprocessing methods!

Mind that a all of these experiments can benefit from additional tuning and of course from more recent methods…

15 minutes

Too simple for you? Start your own project right away!

http://pixabay.com/en/stopwatch-racing-timer-timing-153398/ (CC0)

UKP

TECHNISCHE UNIVERSITÄT DARMSTADT

Kontakt / Contact

Christian M. Meyer and **Hatem Mousselly-Sergieh**
Technische Universität Darmstadt
Ubiquitous Knowledge Processing Lab

Hochschulstr. 10, 64289 Darmstadt, Germany
+49 (0)6151 16–5386
+49 (0)6151 16–5455
meyer (at) ukp.informatik.tu-darmstadt.de

Rechtliche Hinweise

Die Folien sind für den persönlichen Gebrauch der Vortragsteilnehmer gedacht. Im Vortrag verwendete Photographien, Illustrationen, Wort- und Bildmarken sind Eigentum der jeweiligen Rechteinhaber oder Lizenzgeber. Um Missverständnisse zu vermeiden, wäre eine kurze Kontaktaufnahme vor Weitergabe oder -nutzung der Vortragsmaterialien empfehlenswert. Sofern Sie Ihre Rechte verletzt sehen, bitte ich ebenfalls um Kontaktaufnahme zur Klärung der Sachlage.

Legal Issues

The slides are intended for personal use by the audience of the talk. Photographies, illustrations, tradedmarks, or logos are property of the holder of rights. To avoid any misconceptions, I would strongly recommend to get in touch before reusing or redistributing the slides or any additional material of the talk. The same applies if you consider your rights infringed – please let me know to initiate further clarification.