Worth its Weight in Gold or Yet Another Resource

A Comparative Study of Wiktionary, OpenThesaurus and GermaNet

TECHNISCHE UNIVERSITÄT DARMSTADT

Christian M. Meyer and Iryna Gurevych

11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Iaşi, Romania, March 2010. Lecture Notes in Computer Science, Vol. 6008, p. 38-49.

Motivation

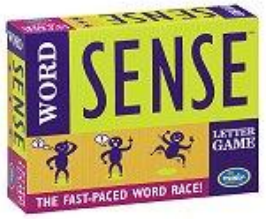
NLP Tasks and Lexical Semantic Knowledge

TECHNISCHE UNIVERSITÄT DARMSTADT

Applications

Google translate

Semantic Search on its way!



Expert-built

Lexical Semantic Resources

GermaNet

GermaNet

WordNet

OpenCyc

Ubiquitous Knowledge Processing

Motivation

Expert-built Lexical Semantic Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

Applications

YAHOO! BABEL FISH

Google translate

Semantic Search on its way!

WORD SENSE

Expert-built

Lexical Semantic Resources

GermaNet

GermaNet

WordNet

OpenCyc

- ✓ used for many years
- ✓ well studied
- ✗ high construction cost
- ✗ limited size
- ✗ hard to keep up-to-date

Ubiquitous Knowledge Processing

Motivation

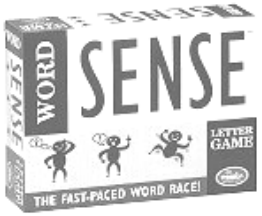
Collaboratively-built Lexical Semantic Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

Applications

Google translate

Semantic Search on its way!



- ✓ emerging
- ✓ freely available
- ✓ constantly updated
- ✓ competitive to expert-built
- ✗ **structure and content related properties are largely unknown**

Collaboratively-built

Lexical Semantic Resources

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

openthesaurus.de

Ubiquitous Knowledge Processing

Motivation

Collaboratively-built Lexical Semantic Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

Structure and content related properties of collaborative resources are largely unknown

- **How are the resources organized?**
- **Which kind of semantic knowledge is encoded?**
- **What are their strengths and drawbacks?**

Collaboratively-built

Lexical Semantic Resources

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

WIKIPEDIA
The Free Encyclopedia

Ubiquitous Knowledge Processing

Motivation

Collaboratively-built Lexical Semantic Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

Structure and content related properties of collaborative resources are largely unknown

→ Perform a comparative study of resources

Expert-built

Lexical Semantic Resources

GermaNet

GermaNet

WordNet

OpenCyc

Collaboratively-built

Lexical Semantic Resources

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

openthesaurus.de

WIKIPEDIA
The Free Encyclopedia

Ubiquitous Knowledge Processing

Lexical Semantic Resources
Wiktionary

TECHNISCHE UNIVERSITÄT DARMSTADT

boat

English

Most common English words: due « Henry « society « #797: boat » heaven » v. » difficult

Pronunciation

- (*RP*) enPR: bōt, IPA: /bəʊt/, SAMPA: /b@Ut/
- (*GenAm*) enPR: bōt, IPA: /boʊt/, SAMPA: /boUt/
- Audio (US)[help, file]
- Rhymes: -əʊt

A **boat** kept on land

Etymology

From Old English *bāt* < Proto-Germanic *baitaz. Related to Old Norse *bátr*, *beit* (Icelandic: *bátur*). Related to German Boot and Dutch boot.

Noun

Word Senses

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 conformers of cyclohexane rings (the other being chair), shaped roughly like a boat.

Semantic Relations

Synonyms

[1] craft, ship, vessel

Hyponyms

[1] ark, bangca, barge, canoe, catamaran, caravel, carrack, coracle, cruiser, cutter, dhow, dinghy, dory, dragon boat, Dutch barge, East Indiaman, felucca, ferry, ferryboat, fishing boat, folding boat, galley, galleon, gig, go-fast boat, houseboat, hovercraft, hydrofoil, hydroplane, inflatable boat, inflatable raft, jetboat, jetski, junk, kayaaki,

Collaboratively created online dictionary

- Language
- Etymology
- Pronunciation
- Part-of-speech
- Word senses
- Synonyms
- Derived Terms
- Translations
- …

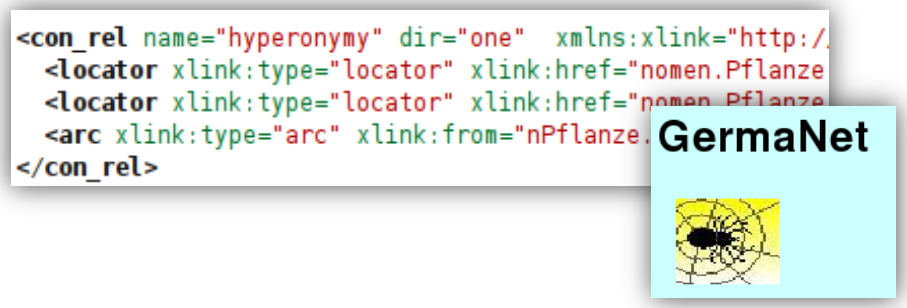
Lexical Semantic Resources

GermaNet and OpenThesaurus

TECHNISCHE UNIVERSITÄT DARMSTADT

GermaNet

- Semantic Network for the German Language
- Created by lexicographers
- WordNet-like structure
- [Kunze and Lemnitzer, 2002]



OpenThesaurus

- Collaborative (but moderated) collection of synonyms
- Used in OpenOffice
- [Naber, 2005]

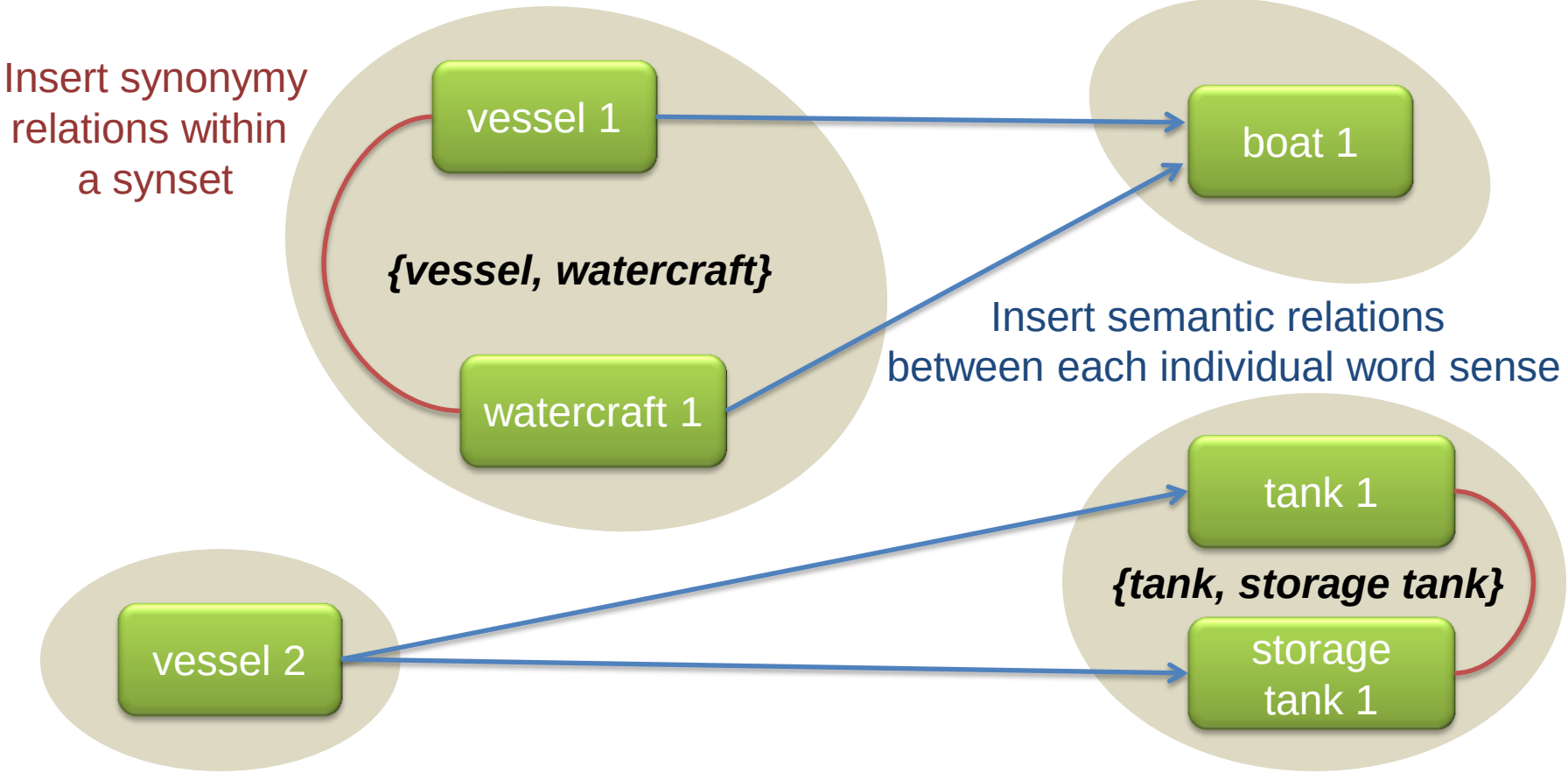
Ubiquitous Knowledge Processing

A Uniform Representation of Resources

Splitting of Synsets

TECHNISCHE UNIVERSITÄT DARMSTADT

{vessel, watercraft} is hypernym of {boat}
{vessel} is hypernym of {tank, storage tank}



Ubiquitous Knowledge Processing

A Uniform Representation of Resources

Word Sense Disambiguation of Relations

TECHNISCHE UNIVERSITÄT DARMSTADT

boat

English

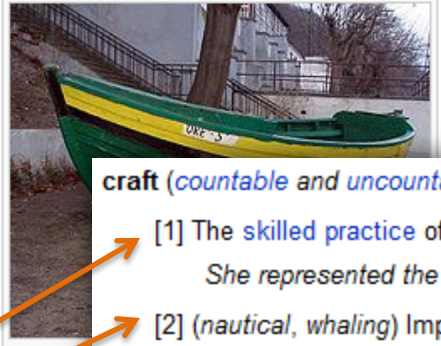
Most common English words: due « Henry « society « #797: boat » heaven » v. » difficult

Pronunciation

- (*RP*) enPR: bōt, IPA: /bəʊt/, SAMPA: /b@Ut/
- (*GenAm*) enPR: bōt, IPA: /boʊt/, SAMPA: /boUt/
- Audio (US)help, file
- Rhymes: -əʊt

Etymology

From Old English *bāt* < Proto-Germanic *baitaz. Related to Old Norse *bátr*, *beit* (Icelandic: *bátur*). Related to German Boot and Dutch boot.



A boat ke

Noun

boat (*plural* **boats**)

[1] A craft used for transportation of goods, fishing, racing, recreational cruising, (propelled by oars or outboard motor or inboard motor or by wind.

[2] (*poker slang*) A full house.

[3] (*chemistry*) One of two possible conformers of cyclohexane rings (the other b boat.

Synonyms

[1] craft, ship, vessel

Hyponyms

[1] ark, bangca, barge, canoe, catamaran, caravel, carrack, coracle, cruiser, cutter, dhow, dinghy, dory, dragon boat, Dutch barge, East Indiaman, felucca, ferry, ferryboat, fishing boat, folding boat, galley, galleon, gig, go-fast boat, houseboat, hovercraft, hydrofoil, hydroplane, inflatable boat, inflatable raft, jetboat, jetski, junk, kayaaki,

?

craft (*countable and uncountable; plural* **craft** *or* **crafts**)

[1] The skilled practice of a practical occupation.

*She represented the **craft** of brewers.*

[2] (*nautical*, *whaling*) Implements used in catching fish, such as net, line, or hoo as in harpoons, hand-lances, etc.

[3] (*nautical*) Boats, especially of smaller size than ships. Historically primarily a loading or unloading of other vessels, as lighters, hoys, and barges.

[4] (*nautical*, *British Royal Navy*) Those vessels attendant on a fleet, such as cutt generally commanded by lieutenants.

[5] A vehicle designed for navigation in or on water or air or through outer space.

[6] A particular kind of skilled work.

*He learned his **craft** as an apprentice.*

[7] Shrewdness as demonstrated by being skilled in deception.

Sense [1] encodes a synonymy relation to “craft”.

But which sense?

Ubiquitous Knowledge Processing

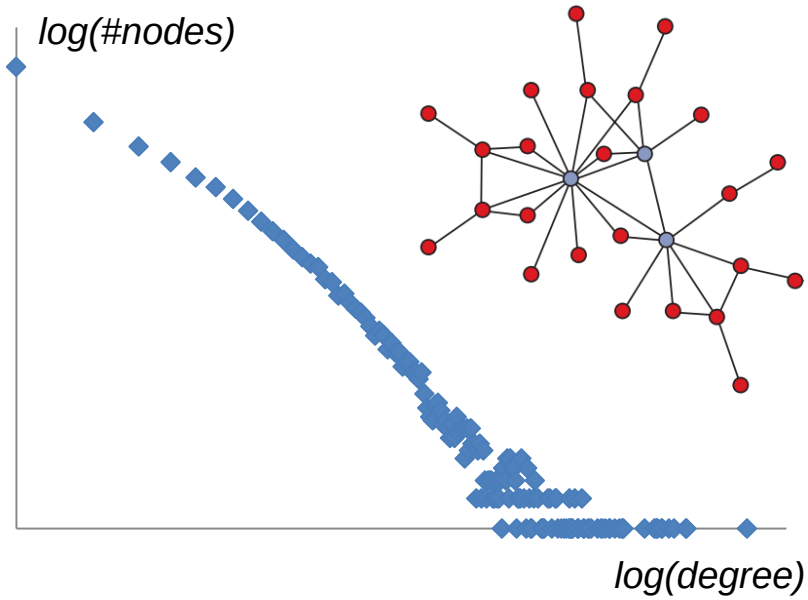
Structural Analysis

Topological Results

TECHNISCHE UNIVERSITÄT DARMSTADT

Analysis:

- Connectivity
- Degree distribution
- Network organization
- Cluster analysis



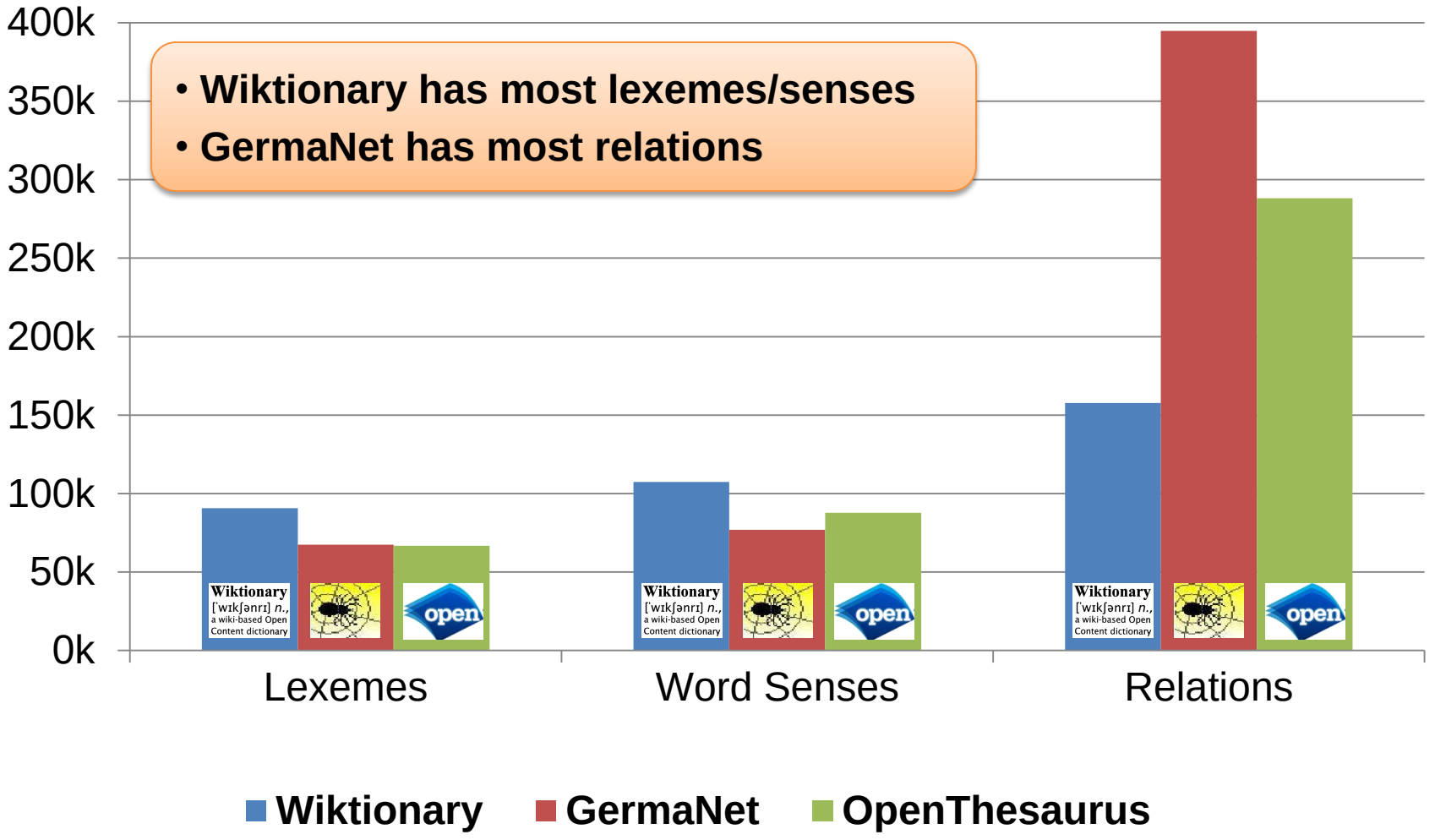
Results:

- The **largest connected component** contains the bulk of semantic knowledge
- The Wiktionary graph is **scale-free** and allows to predict analysis results to future (larger) versions
- All graphs are **small world graphs**; they show organizational patterns that significantly differ from random graphs

Content Analysis

Resource Size

TECHNISCHE UNIVERSITÄT DARMSTADT



Ubiquitous Knowledge Processing

Content Analysis
Polysemy

TECHNISCHE UNIVERSITÄT DARMSTADT

	Wiktionary		GermaNet		OpenThesaurus	
Number of Lexemes	90,611		67,402		66,754	
..Monosemous Lexemes	29,025	32.0%	61,129	90.6%	54,939	82.3%
..Polysemous Lexemes	10,643	26.8%	6,237	9.2%	11,815	17.6%
..Dangling Lexemes	50,943	56.2%	--		--	

What causes this difference? Possible explanations:

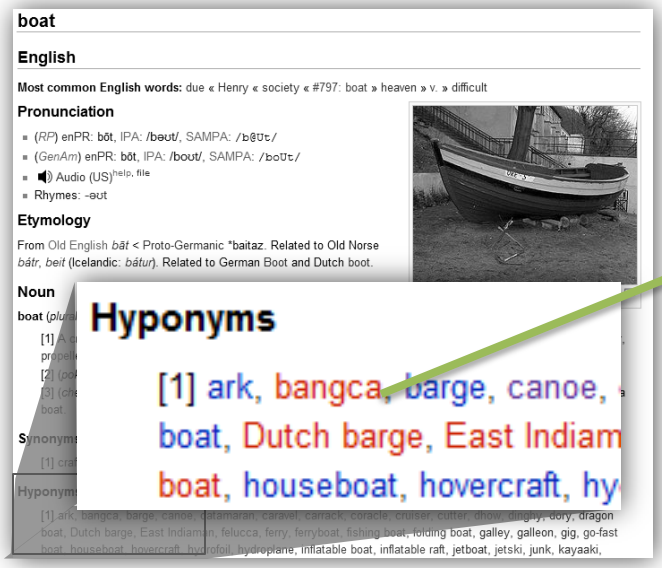
1. Wiktionary contains more word with a high frequency in language (known to be more ambiguous)
2. The community more likely creates articles for polysemous terms, since they might be more interesting to create
3. The coverage of Wiktionary senses is on average higher
4. Wiktionary word senses are more fine-grained

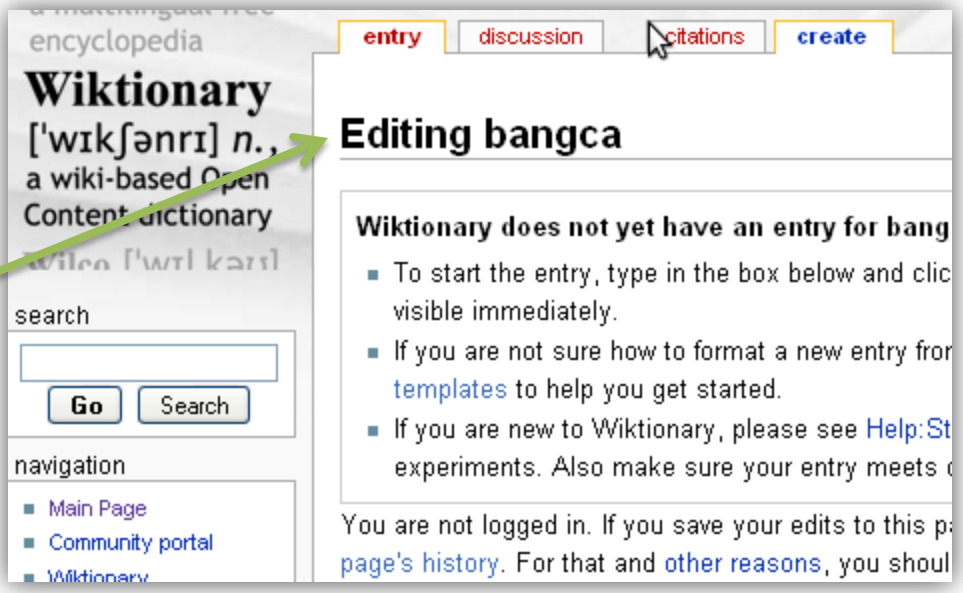
Subject of ongoing work

Content Analysis
Dangling Lexemes

TECHNISCHE UNIVERSITÄT DARMSTADT

	Wiktionary		GermaNet		OpenThesaurus	
Number of Lexemes	90,611		67,402		66,754	
..Monosemous Lexemes	29,025	32.0%	61,129	90.6%	54,939	82.3%
..Polysemous Lexemes	10,643	26.8%	6,237	9.2%	11,815	17.6%
..Dangling Lexemes	50,943	56.2%	--		--	





Ubiquitous Knowledge Processing

Content Analysis
Dangling Lexemes

TECHNISCHE UNIVERSITÄT DARMSTADT

	Wiktionary		GermaNet		OpenThesaurus	
Number of Lexemes	90,611		67,402		66,754	
..Monosemous Lexemes	29,025	32.0%	61,129	90.6%	54,939	82.3%
..Polysemous Lexemes	10,643	26.8%	6,237	9.2%	11,815	17.6%
..Dangling Lexemes	50,943	56.2%	--		--	

Wiktionary is still growing…

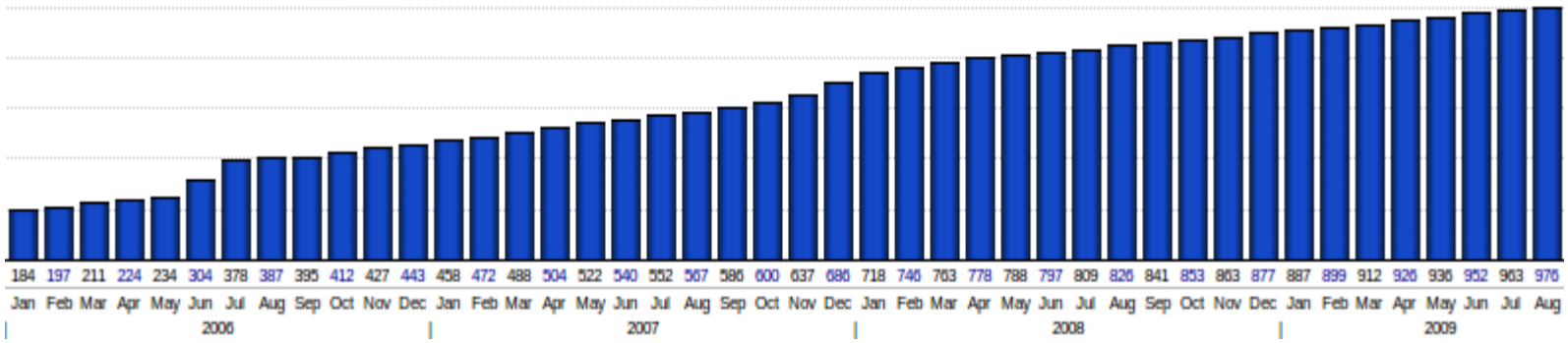


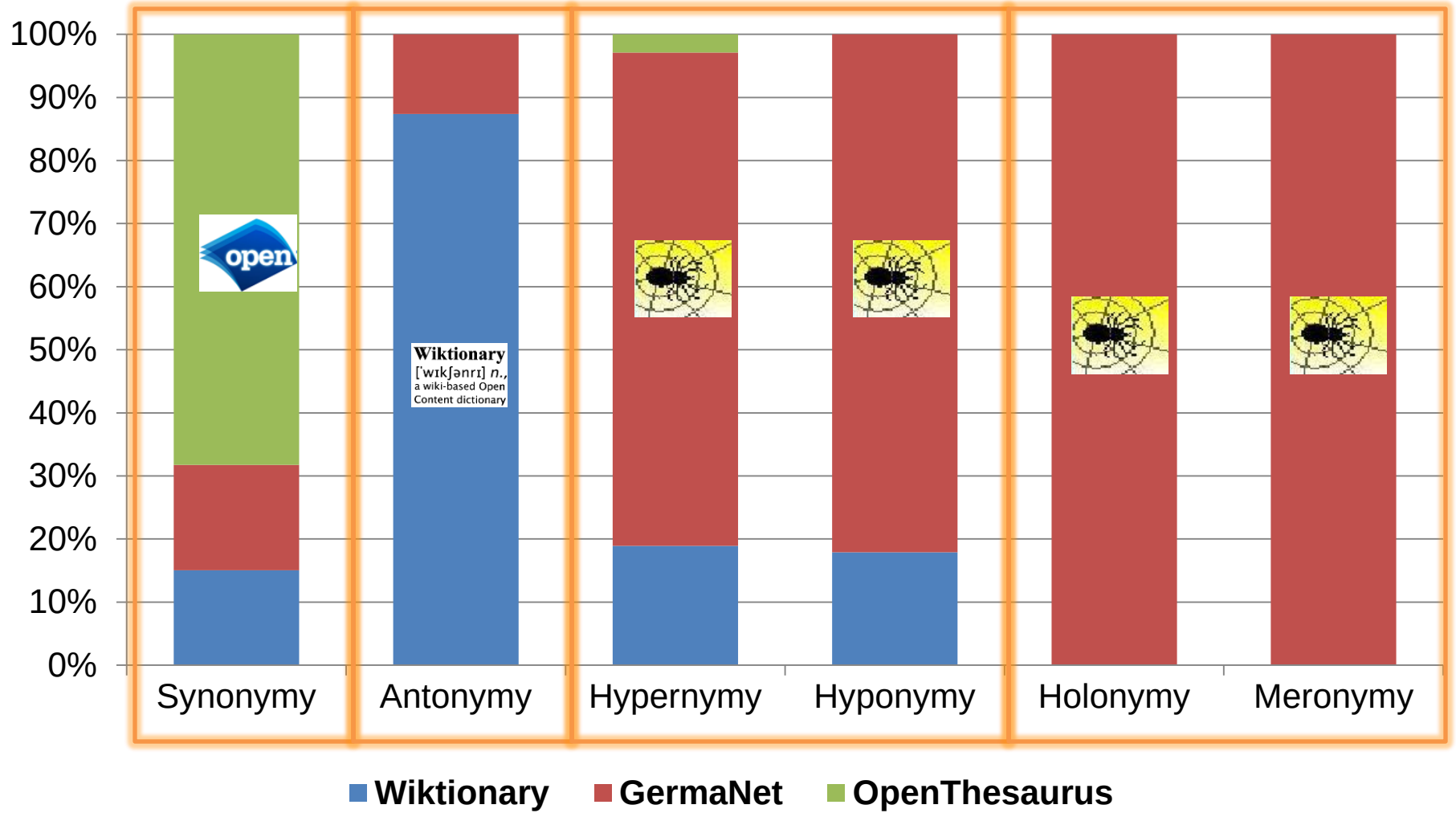
Fig. courtesy: http://stats.wikimedia.org/wiktionary/EN/ChartsWikipediaDE.htm

Ubiquitous Knowledge Processing

Content Analysis

Semantic Relations

TECHNISCHE UNIVERSITÄT DARMSTADT

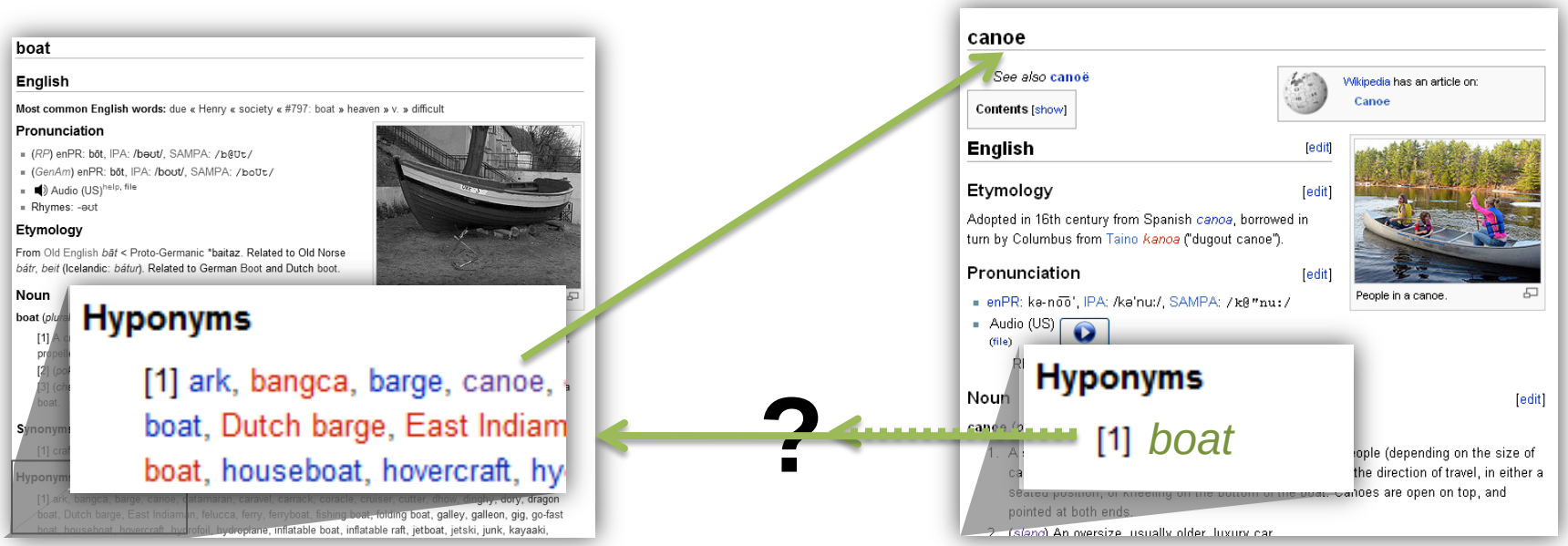


Ubiquitous Knowledge Processing

Content Analysis
One Way Relations

TECHNISCHE UNIVERSITÄT DARMSTADT

	Wiktionary	GermaNet	OpenThesaurus
Number of Relations	157,786	394,856	288,121
…One Way Relations	139,453	11,941	5,731
Resulting two way relations	297,120	406,328	293,846



Ubiquitous Knowledge Processing

Conclusions

Take-home Message

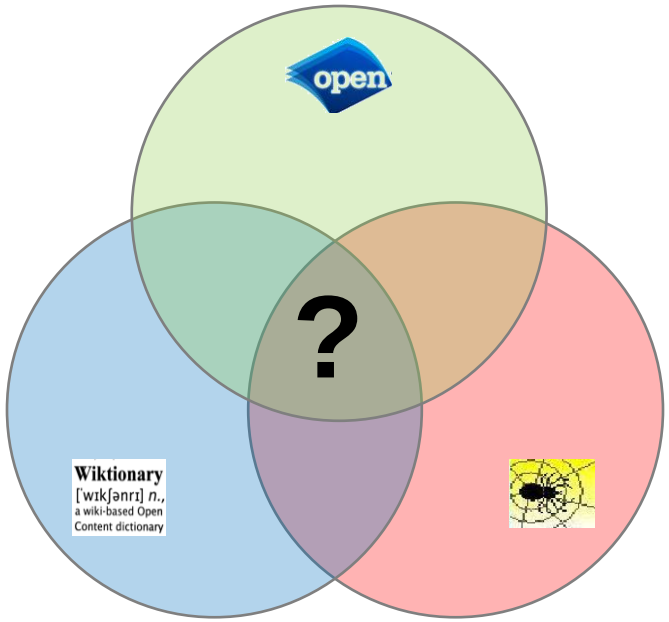
TECHNISCHE UNIVERSITÄT DARMSTADT

- **How are the resources organized?**
  - Largest connected component is sufficient
  - Small world property
- **Which kind of semantic knowledge is encoded?**
  - More polysemous lexemes in Wiktionary
  - Many dangling lexemes
- **What are their strengths and drawbacks?**
  - Predominant type of relation for each resource
  - Number of semantic relations can be increased in Wiktionary

Conclusions
Future Work

TECHNISCHE UNIVERSITÄT DARMSTADT

- Study English resources
- Improve word sense disambiguation in Wiktionary
- How large is the information overlap of the resources?
- Combine the resources



Ubiquitous Knowledge Processing

Thank you for your attention!

TECHNISCHE UNIVERSITÄT DARMSTADT

Ubiquitous Knowledge Processing

KTS
Klaus Tschira Stiftung gemeinnützige GmbH

e-learning center of research excellence

Bundesministerium für Wirtschaft und Technologie

VolkswagenStiftung

Deutsche Forschungsgemeinschaft DFG

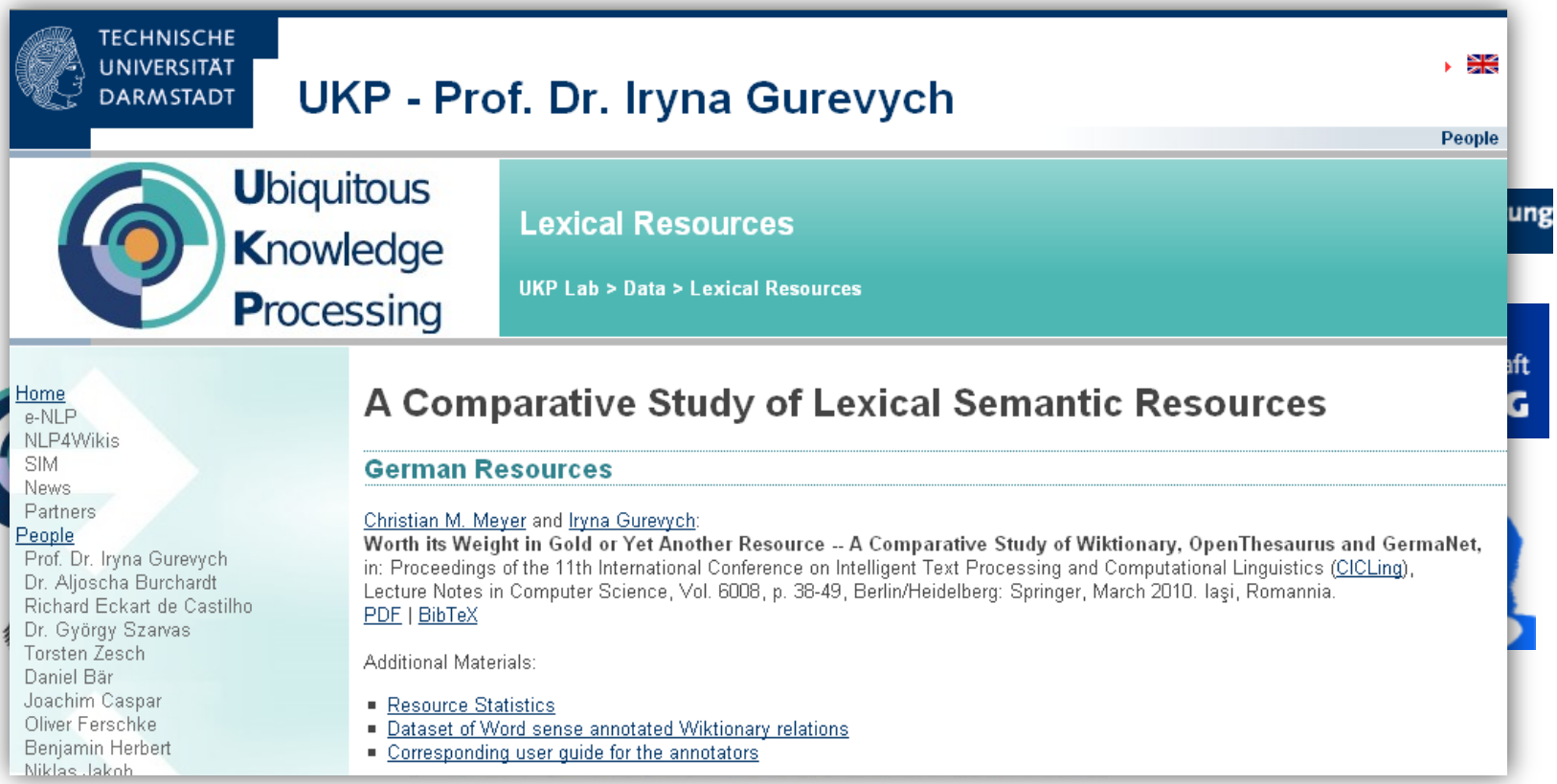
Emmy Noether-Programm
Deutsche Forschungsgemeinschaft DFG

Additional Online Material:

http://www.ukp.tu-darmstadt.de/data/lexical-resources/

Thank you for your attention!

TECHNISCHE UNIVERSITÄT DARMSTADT



Additional Online Material:

http://www.ukp.tu-darmstadt.de/data/lexical-resources/

TECHNISCHE UNIVERSITÄT DARMSTADT

Kontakt / Contact

Christian M. Meyer
Technische Universität Darmstadt
Ubiquitous Knowledge Processing Lab

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