Worth its Weight in Gold or Yet Another Resource

A Comparative Study of Wiktionary, OpenThesaurus and GermaNet

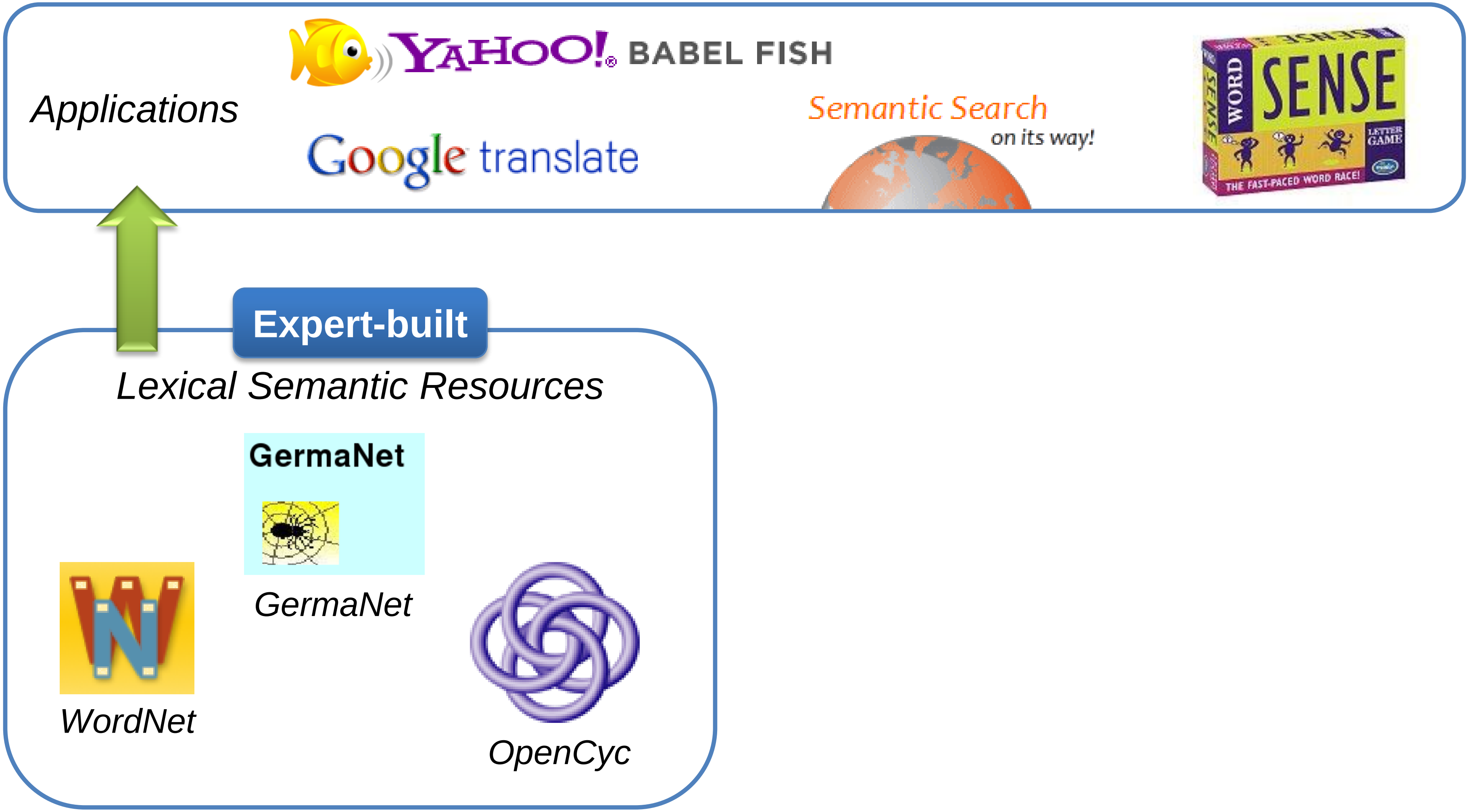
Christian M. Meyer and Iryna Gurevych

First Workshop on Automated Knowledge Base Construction (AKBC), Grenoble, France, May 2010.

Previously published in:
Lecture Notes in Computer Science, Vol. 6008, p. 38-49.
http://dx.doi.org/10.1007/978-3-642-12116-6_4

Motivation

NLP Tasks and Lexical Semantic Knowledge



Applications

Expert-built

Lexical Semantic Resources

GermaNet

WordNet

OpenCyc

Motivation

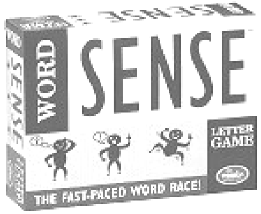
Expert-built Lexical Semantic Resources

Applications

YAHOO! BABEL FISH

Google translate

Semantic Search on its way!



Expert-built

Lexical Semantic Resources

GermaNet

GermaNet

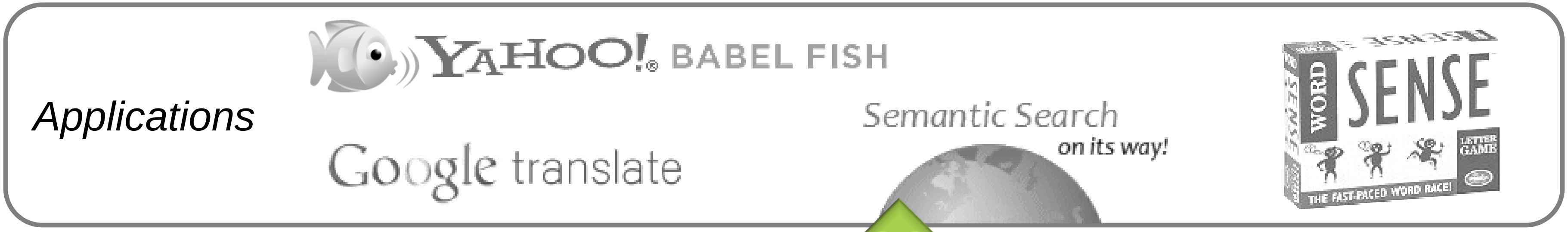
WordNet

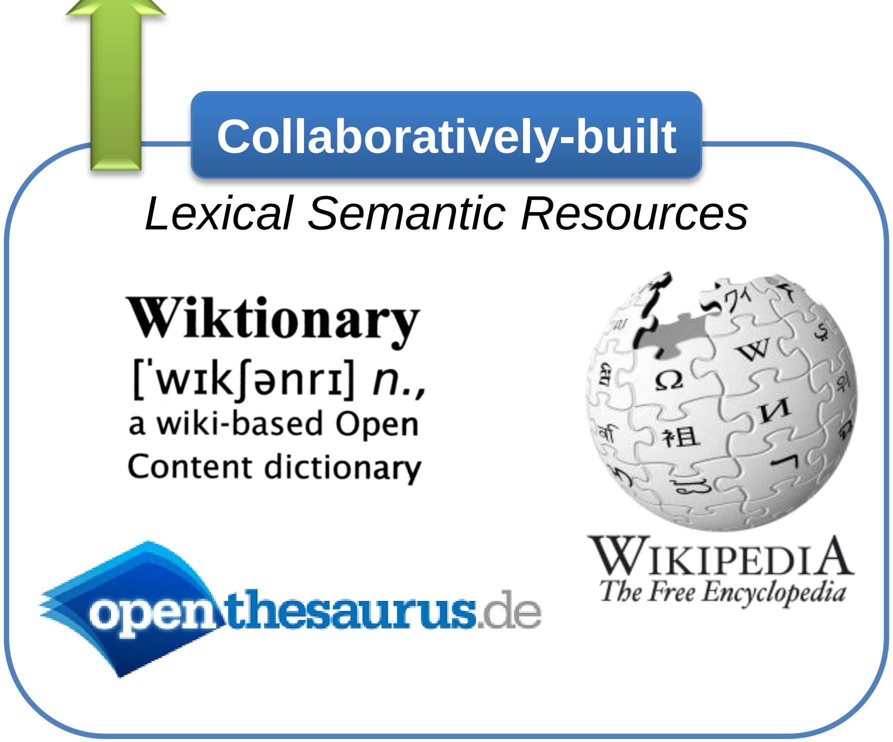
OpenCyc

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

Motivation

Collaboratively-built Lexical Semantic Resources





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

Motivation

Collaboratively-built Lexical Semantic Resources

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
openthesaurus.de

Motivation

Collaboratively-built Lexical Semantic Resources

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

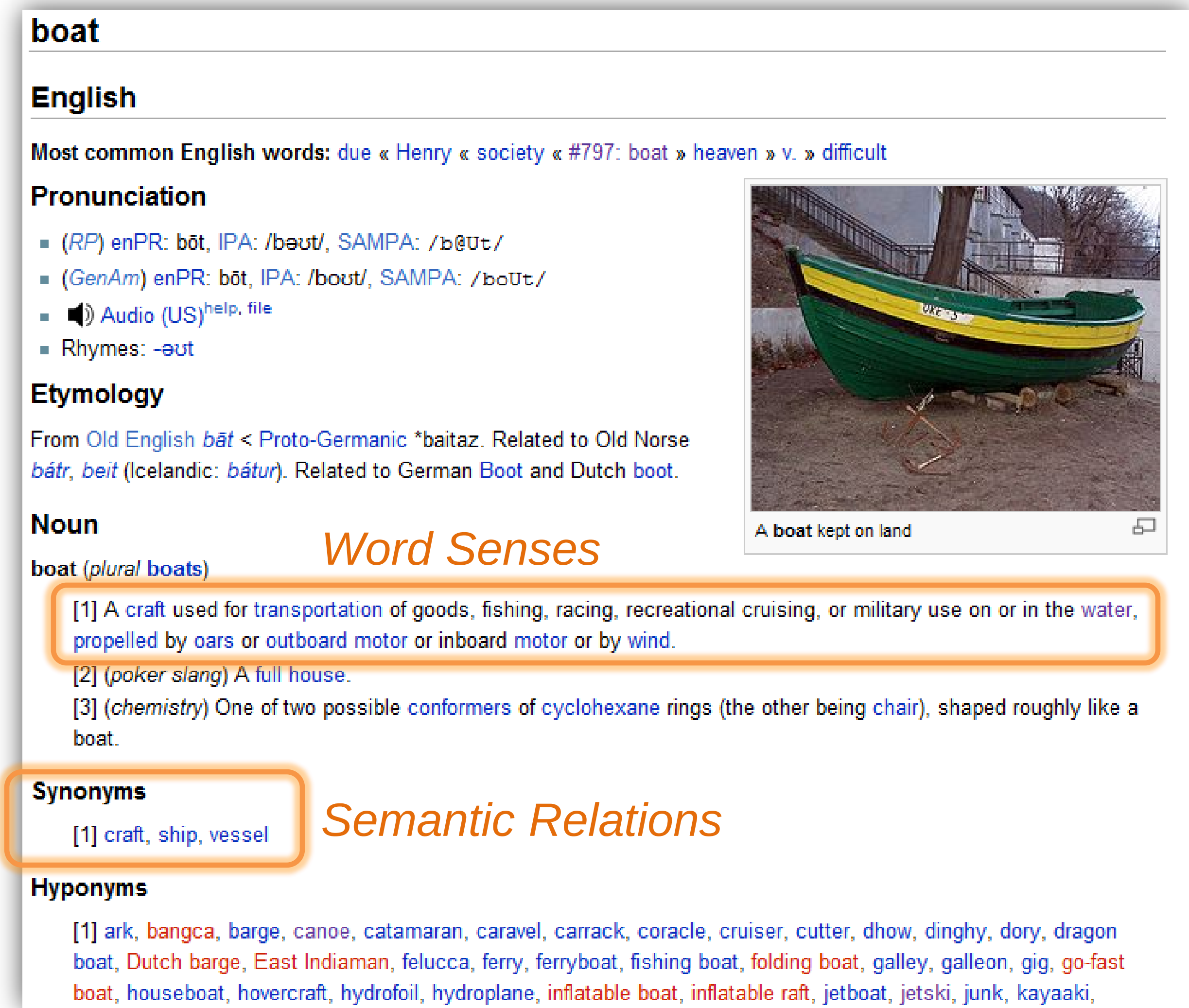
Lexical Semantic Resources

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

Wikipedia
The Free Encyclopedia

openthesaurus.de

Lexical Semantic Resources
Wiktionary



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 kept on land

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.

Synonyms

[1] craft, ship, vessel

Semantic Relations

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,

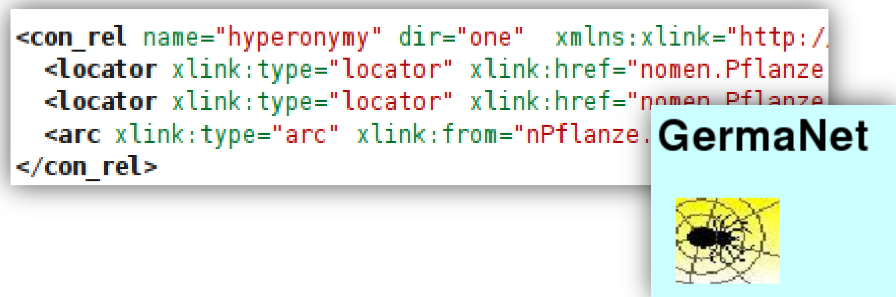
Collaboratively created online dictionary

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

Lexical Semantic Resources
GermaNet and OpenThesaurus

GermaNet

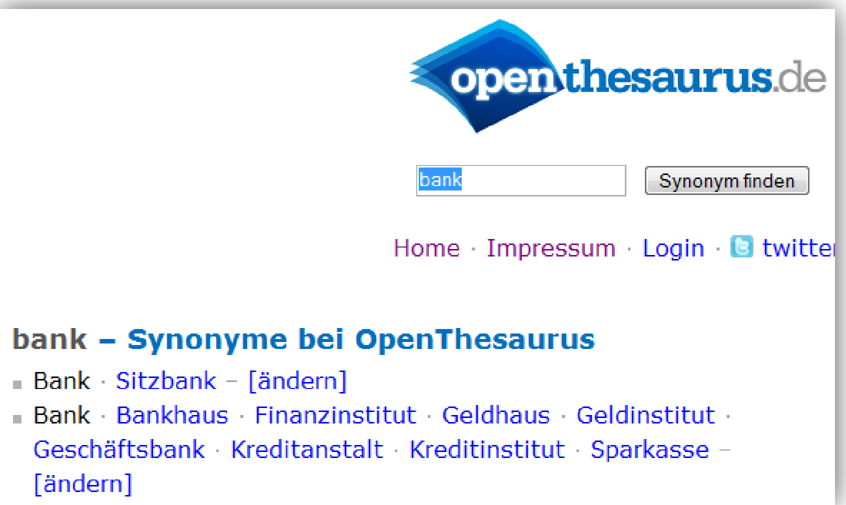
- Semantic Network for the German Language
- Created by lexicographers
- WordNet-like structure
- [Kunze and Lemnitzer, 2002]



```
<con_rel name="hyperonymy" dir="one"  xmlns:xlink="http:/
  <locator xlink:type="locator" xlink:href="nomen.Pflanze
  <locator xlink:type="locator" xlink:href="nomen.Pflanze
  <arc xlink:type="arc" xlink:from="nPflanze.
</con_rel>
```

OpenThesaurus

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



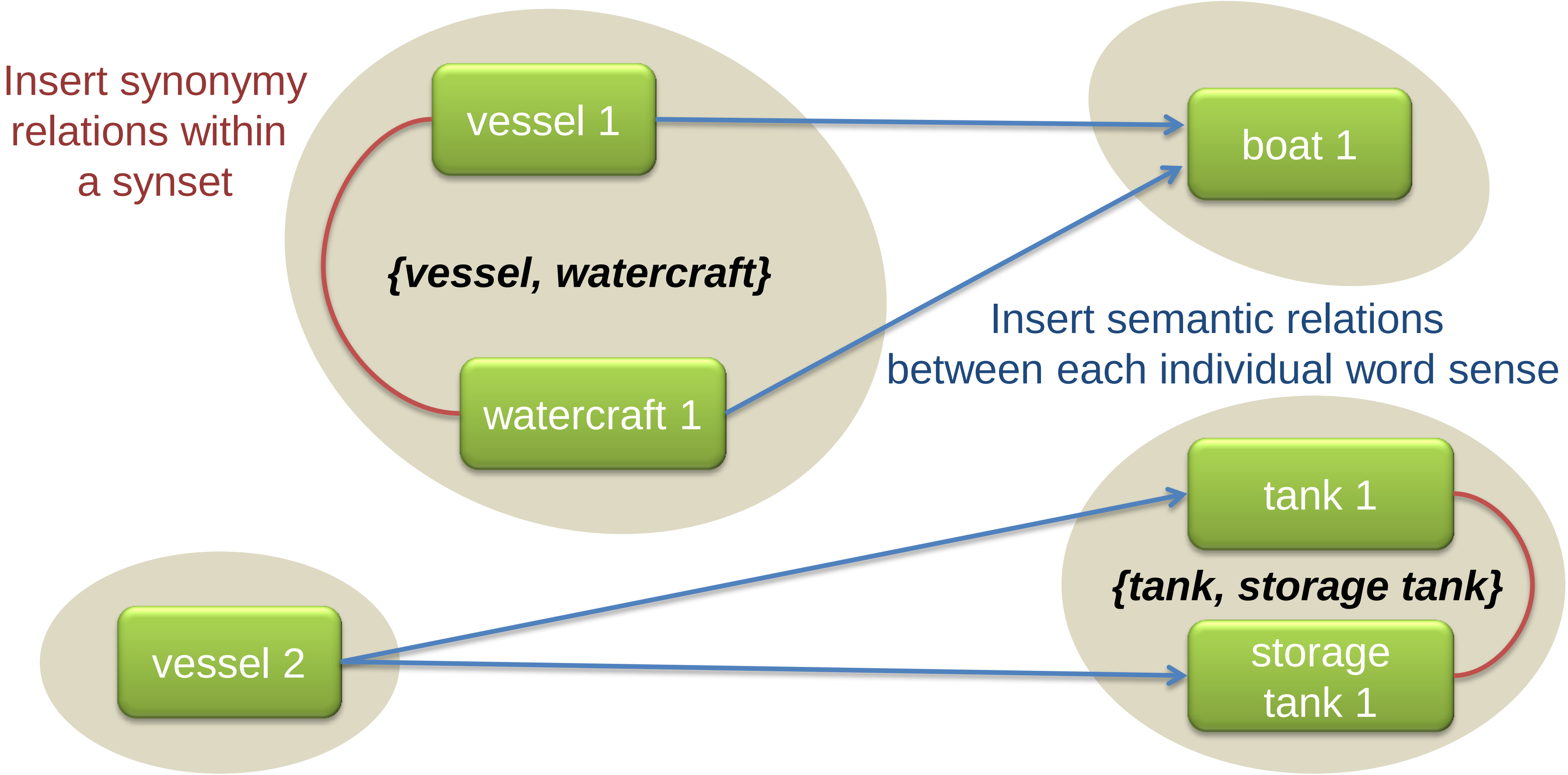
bank – Synonyme bei OpenThesaurus

- Bank · Sitzbank – [ändern]
- Bank · Bankhaus · Finanzinstitut · Geldhaus · Geldinstitut · Geschäftsbank · Kreditanstalt · Kreditinstitut · Sparkasse – [ändern]

A Uniform Representation of Resources

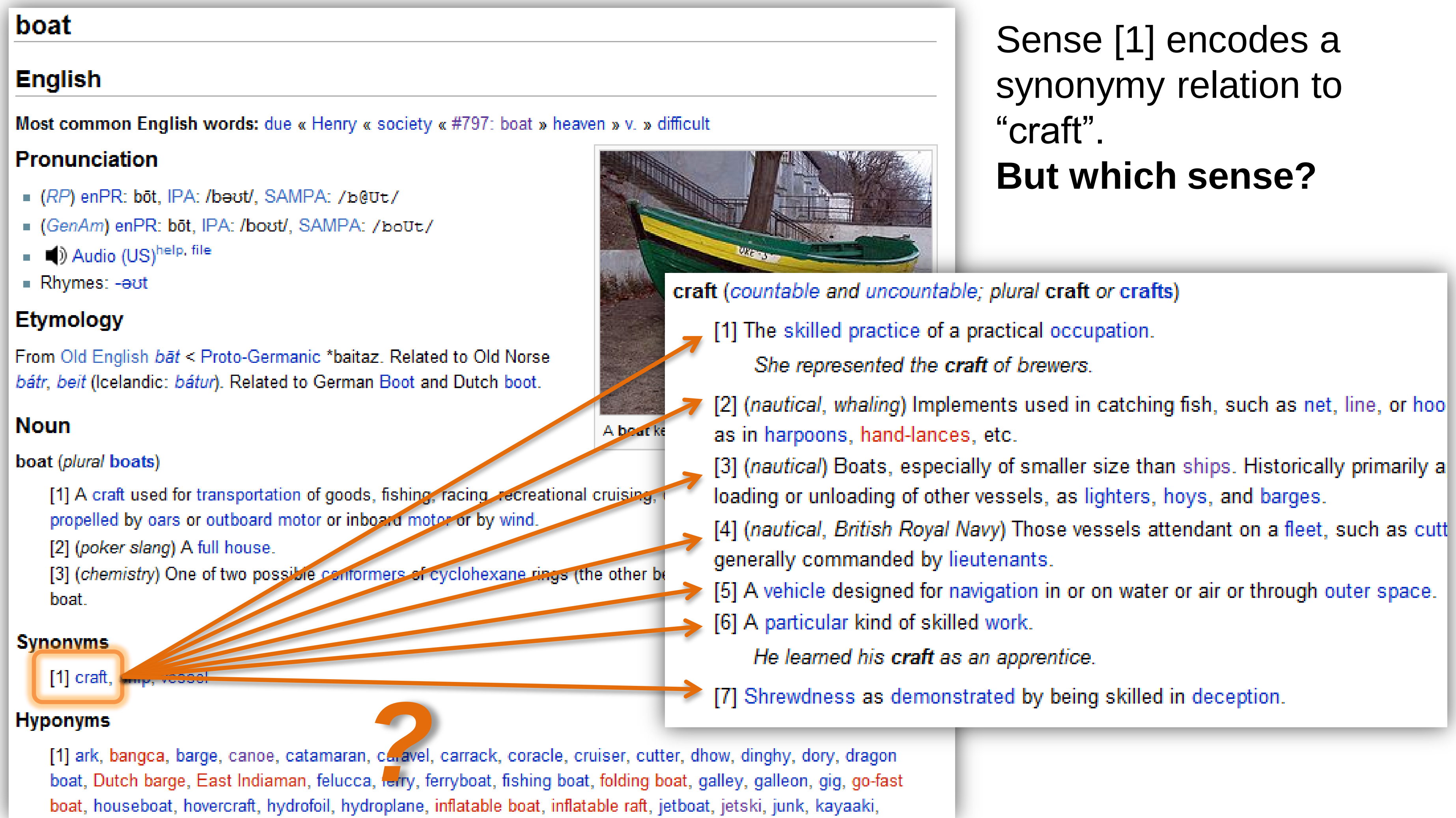
Splitting of Synsets

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


Insert synonymy relations within a synset
vessel 1
{vessel, watercraft}
watercraft 1
boat 1
Insert semantic relations between each individual word sense
tank 1
{tank, storage tank}
storage tank 1
vessel 2

Word Sense Disambiguation in Wiktionary

Finding the Correct Target Word Sense



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 be 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.

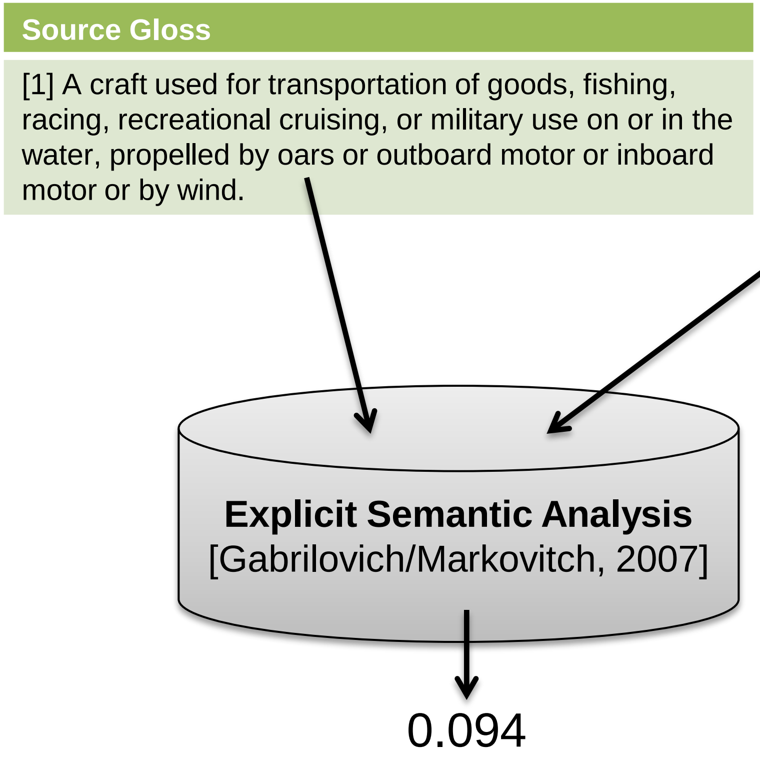
Sense [1] encodes a synonymy relation to “craft”.

But which sense?

Word Sense Disambiguation in Wiktionary

Finding the Correct Target Word Sense – Approach

- We apply a method based on semantic relatedness here.
- Finding the best method for this task is a subject of our current studies.



Source Gloss

[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.

Explicit Semantic Analysis
[Gabrilovich/Markovitch, 2007]

0.094

Score	Target Gloss
0.094	[1] The skilled practice of a practical occupation.
0.711	[2] (nautical, whaling) Implements used in catching fish, such as net, line, or hook. Modern use primarily in whaling, as in harpoons, hand-lances, etc.
0.766	[3] (nautical) Boats, especially of smaller size than ships. Historically primarily applied to vessels engaged in loading or unloading of other vessels, as lighters, hoys, and barges.
0.685	[4] (nautical, British Royal Navy) Those vessels attendant on a fleet, such as cutters, schooners, and gun-boats, generally commanded by lieutenants.
0.237	[5] A vehicle designed for navigation in or on water or air or through outer space.
0.170	[6] A particular kind of skilled work.
0.109	[7] Shrewdness as demonstrated by being skilled in deception.

Word Sense Disambiguation in Wiktionary

Finding the Correct Target Word Sense – Evaluation

- **Upper bound:** 2 human annotators A and B judged 250 randomly sampled relations (= 920 pairs of source and target candidate)
- **Lower bound:** always choose the first word sense (this is usually the most frequent one)

	Lower Bound		*Our Algorithm*		*Upper Bound*
	0–A	*0–B*	*M–A*	*M–B*	*A–B*
A_O	.791	.780	.820	.791	.886
κ	.498	.452	.567	.480	.728
α	.679	.620	.726	.649	.866

A_O: Percentage of agreement

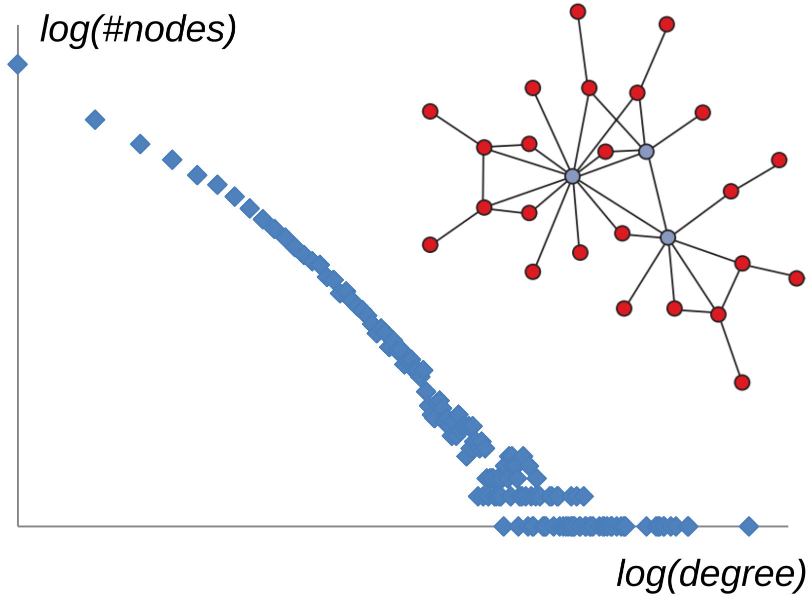
κ: Cohen’s Kappa (chance-corrected)

α: Krippendorff’s Alpha with set-valued distance function (MASI)

Structural Analysis
Topological Results

Analysis:

- Connectivity
- Degree distribution
- Network organization
- Cluster analysis



Results:

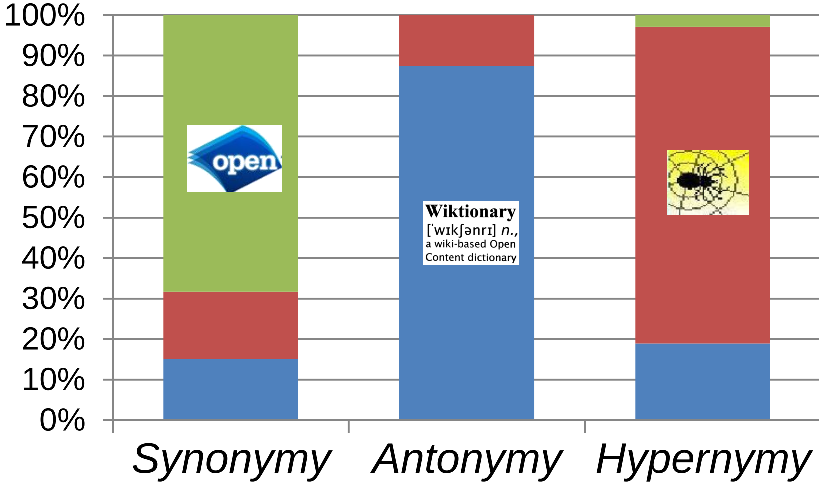
- The **largest connected component** contains the bulk of semantic knowledge
- The Wiktionary graph is **scale-free** and allows to predict analysis results to future (larger) versions
- All graphs are **small world graphs**; they show organizational patterns that significantly differ from random graphs

Content Analysis

Encoded Lexemes, Word Senses and Semantic Relations

Analysis:

- Resource Size
- Polysemy
- Relation Type
- Unidirectional Relations



Results:

- Wiktionary has most **lexemes** and **word senses**
- GermaNet has most **semantic relations**
- In general, more **polysemous lexemes** in Wiktionary
- Many **“dangling” articles** in Wiktionary
- Predominant **type of relation** for each resource
- Most Wiktionary relations are **unidirectional**

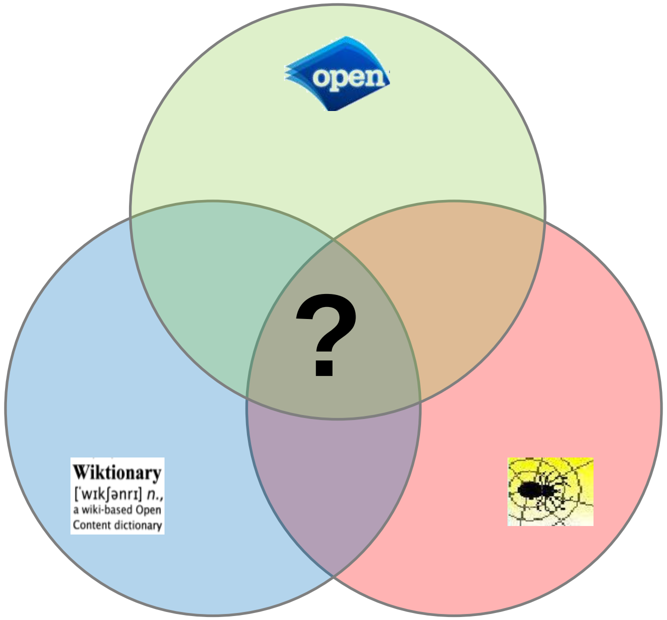
Conclusions
Take-home Message

- **How are the resources organized?**
  - Uniform representation needed
  - 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

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



Thank you for your attention!

Ubiquitous Knowledge Processing

KTS
Klaus Tschira Stiftung
gemeinnützige GmbH

e-learning
center of
research
excellence

Bundesministerium
für Wirtschaft
und Technologie

VolkswagenStiftung

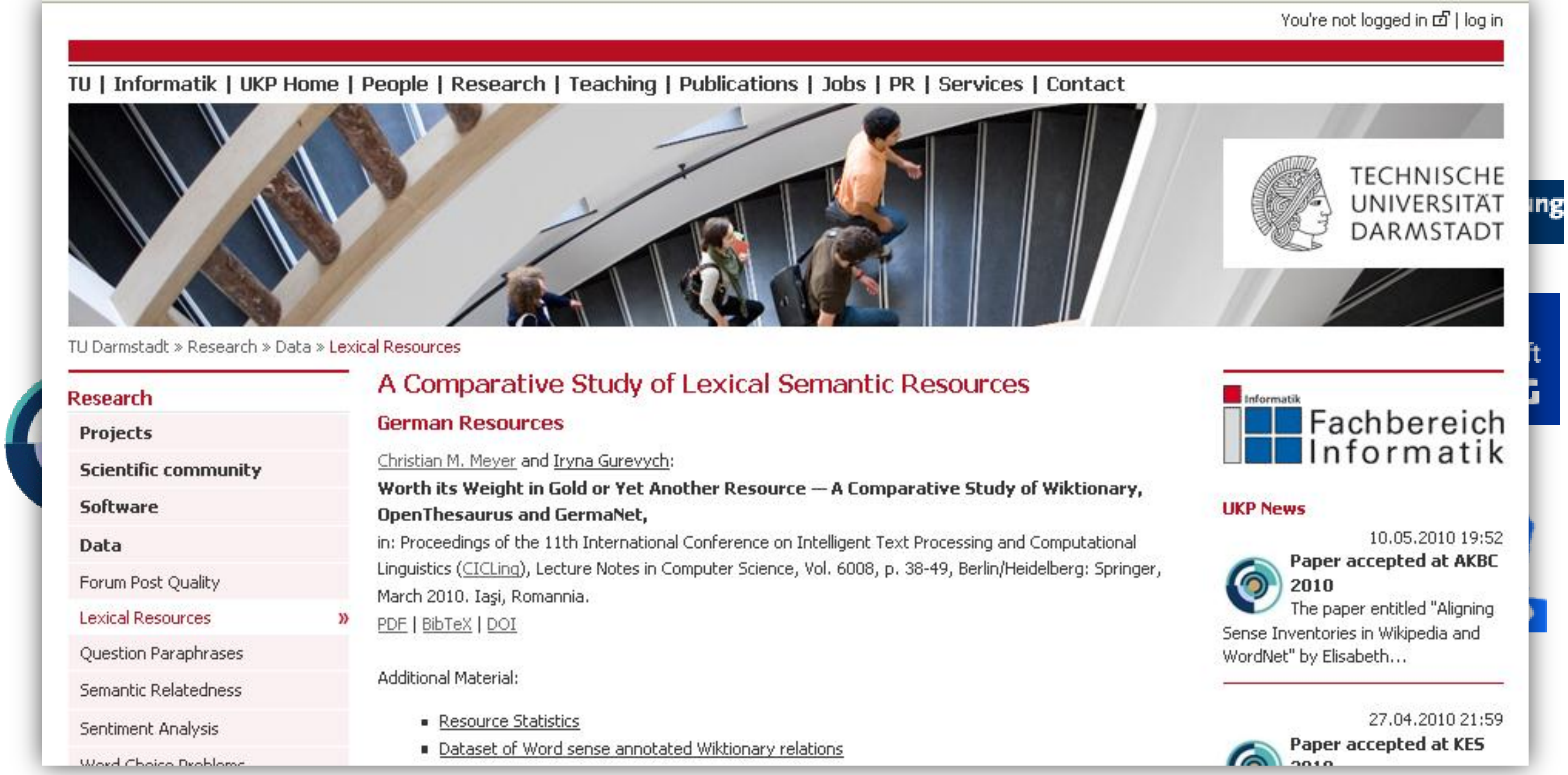
Deutsche
Forschungsgemeinschaft
DFG

Emmy
Noether-
Programm
Deutsche
Forschungsgemeinschaft
DFG

Additional Online Material:

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

Thank you for your attention!



You're not logged in | log in

TU | Informatik | UKP Home | People | Research | Teaching | Publications | Jobs | PR | Services | Contact

TU Darmstadt » Research » Data » Lexical Resources

Research
- Projects
- Scientific community
- Software
- Data
- Forum Post Quality
- Lexical Resources »
- Question Paraphrases
- Semantic Relatedness
- Sentiment Analysis
- Word Choice Problems

A Comparative Study of Lexical Semantic Resources

German Resources

Christian M. Meyer and Iryna Gurevych:

Worth its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet,

in: Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Lecture Notes in Computer Science, Vol. 6008, p. 38-49, Berlin/Heidelberg: Springer, March 2010. Iaşi, Romannia.

PDF | BibTeX | DOI

Additional Material:

- Resource Statistics
- Dataset of Word sense annotated Wiktionary relations

Informatik Fachbereich Informatik

UKP News

10.05.2010 19:52

Paper accepted at AKBC 2010

The paper entitled "Aligning Sense Inventories in Wikipedia and WordNet" by Elisabeth...

27.04.2010 21:59

Paper accepted at KES 2010

Additional Online Material:

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

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

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.