

Worth its Weight in Gold or Yet Another Resource



TECHNISCHE
UNIVERSITÄT
DARMSTADT

*A Comparative Study of Wiktionary,
OpenThesaurus and GermaNet*

Christian M. Meyer and Iryna Gurevych

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

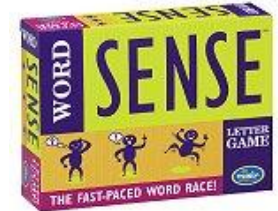
Motivation

NLP Tasks and Lexical Semantic Knowledge

Applications



Semantic Search
on its way!



Expert-built

Lexical Semantic Resources

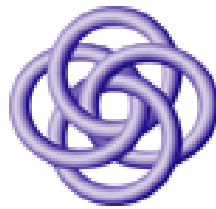
GermaNet



GermaNet



WordNet



OpenCyc

Motivation

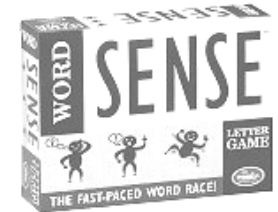
Expert-built Lexical Semantic Resources

Applications



Google translate

Semantic Search
on its way!



Expert-built

Lexical Semantic Resources

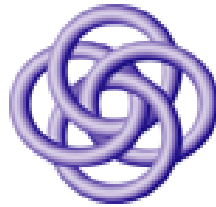
GermaNet



GermaNet



WordNet



OpenCyc

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

Motivation

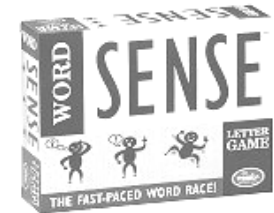
Collaboratively-built Lexical Semantic Resources

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



WIKIPEDIA
The Free Encyclopedia

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



WIKIPEDIA
The Free Encyclopedia



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

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



WIKIPEDIA
The Free Encyclopedia



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@ʊt/
- (GenAm) enPR: bōt, IPA: /boʊt/, SAMPA: /boʊt/
-  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](#).

Noun

boat (plural [boats](#))

Word Senses



[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](#), [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]

```
<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]

openthesaurus.de

bank

[Home](#) · [Impressum](#) · [Login](#) · [twitter](#)

bank – Synonyme bei OpenThesaurus

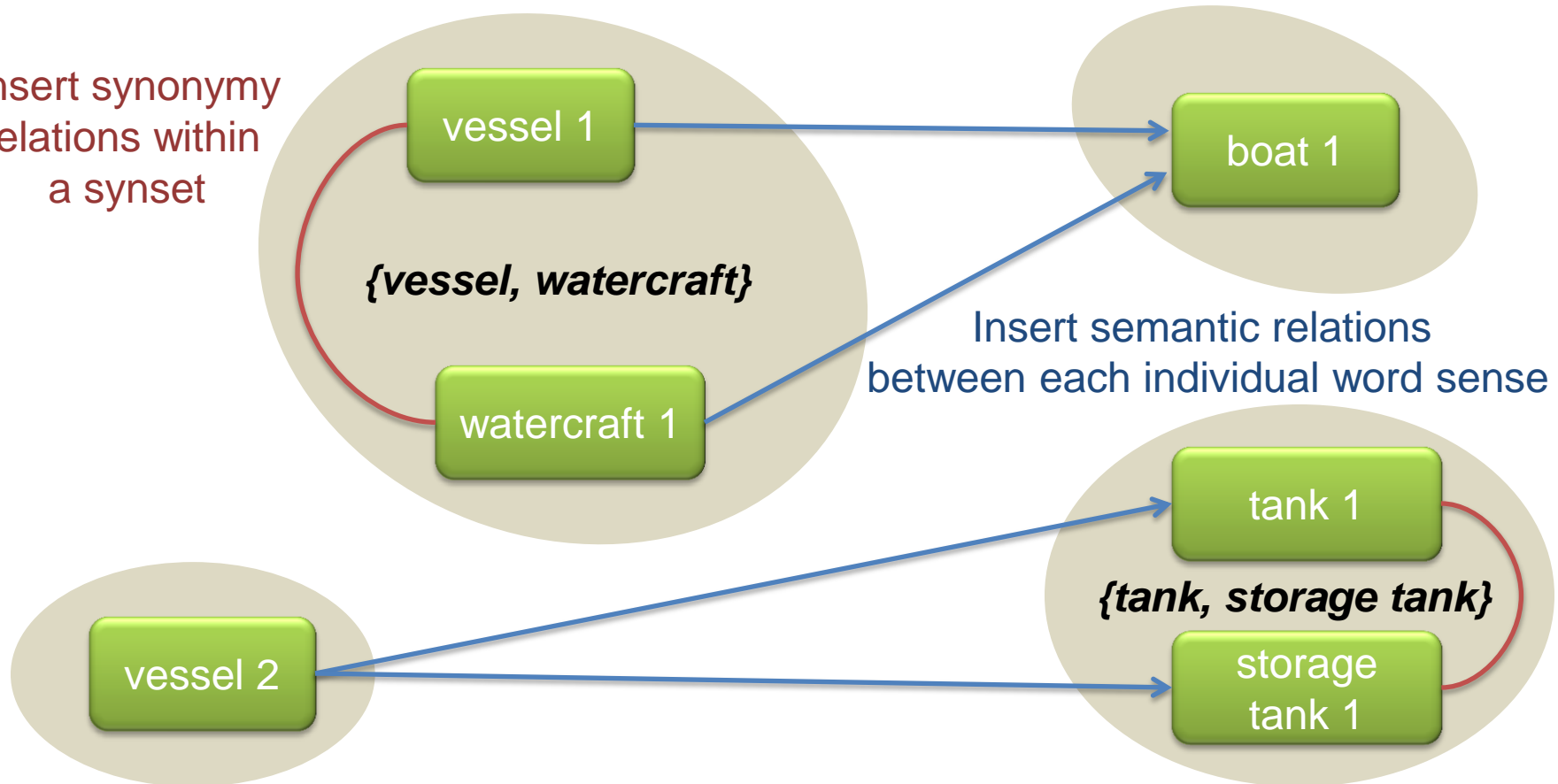
- Bank · [Sitzbank](#) – [[ändern](#)]
- Bank · [Bankhaus](#) · [Finanzinstitut](#) · [Geldhaus](#) · [Geldinstitut](#) · [Geschäftsbank](#) · [Kreditanstalt](#) · [Kreditinstitut](#) · [Sparkasse](#) – [[ändern](#)]

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



A Uniform Representation of Resources


Word Sense Disambiguation of Relations

boat

English

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

Pronunciation

- (RP) enPR: bōt, IPA: /bəʊt/, SAMPA: /b@ʊt/
- (GenAm) enPR: bōt, IPA: /boʊt/, SAMPA: /boʊt/
-  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*.

Noun

boat (plural **boats**)

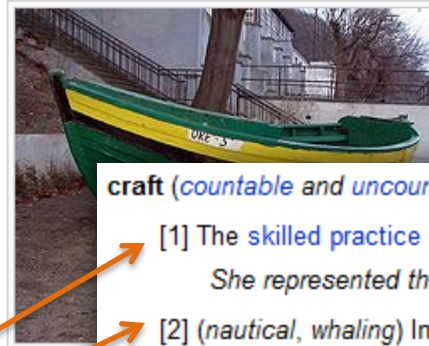
- [1] A craft used for transportation of goods, fishing, racing, recreational cruising, propelled by oars or outboard motor or inboard motor or by wind.
- [2] (*poker slang*) A full house.
- [3] (*chemistry*) One of two possible conformers of cyclohexane rings (the other is boat).

Synonyms

[1] craft, *ship*, *vessel*

Hyponyms

[1] ark, *bangca*, barge, canoe, catamaran, *caivel*, 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,



A boat ke

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 hook 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 cutters 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?



Structural Analysis

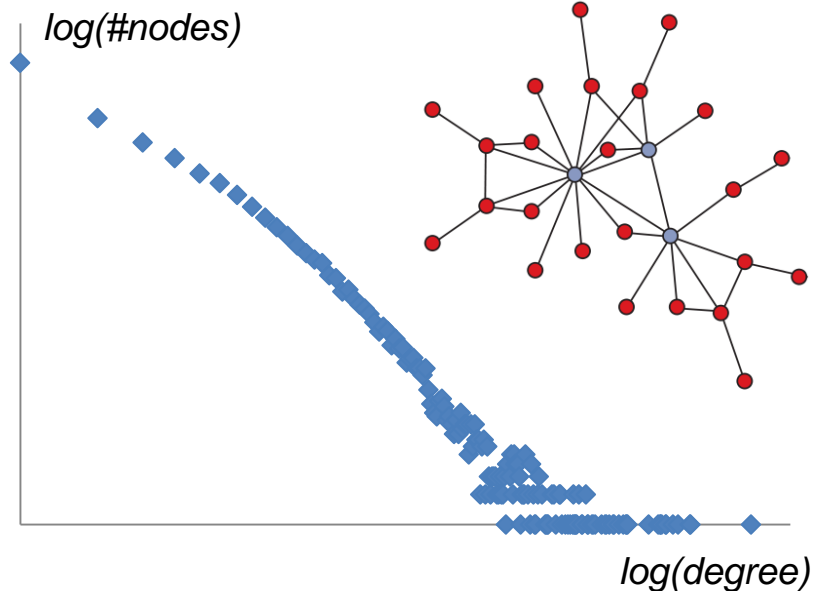
Topological Results

Analysis:

- Connectivity
- Degree distribution
- Network organization
- Cluster analysis

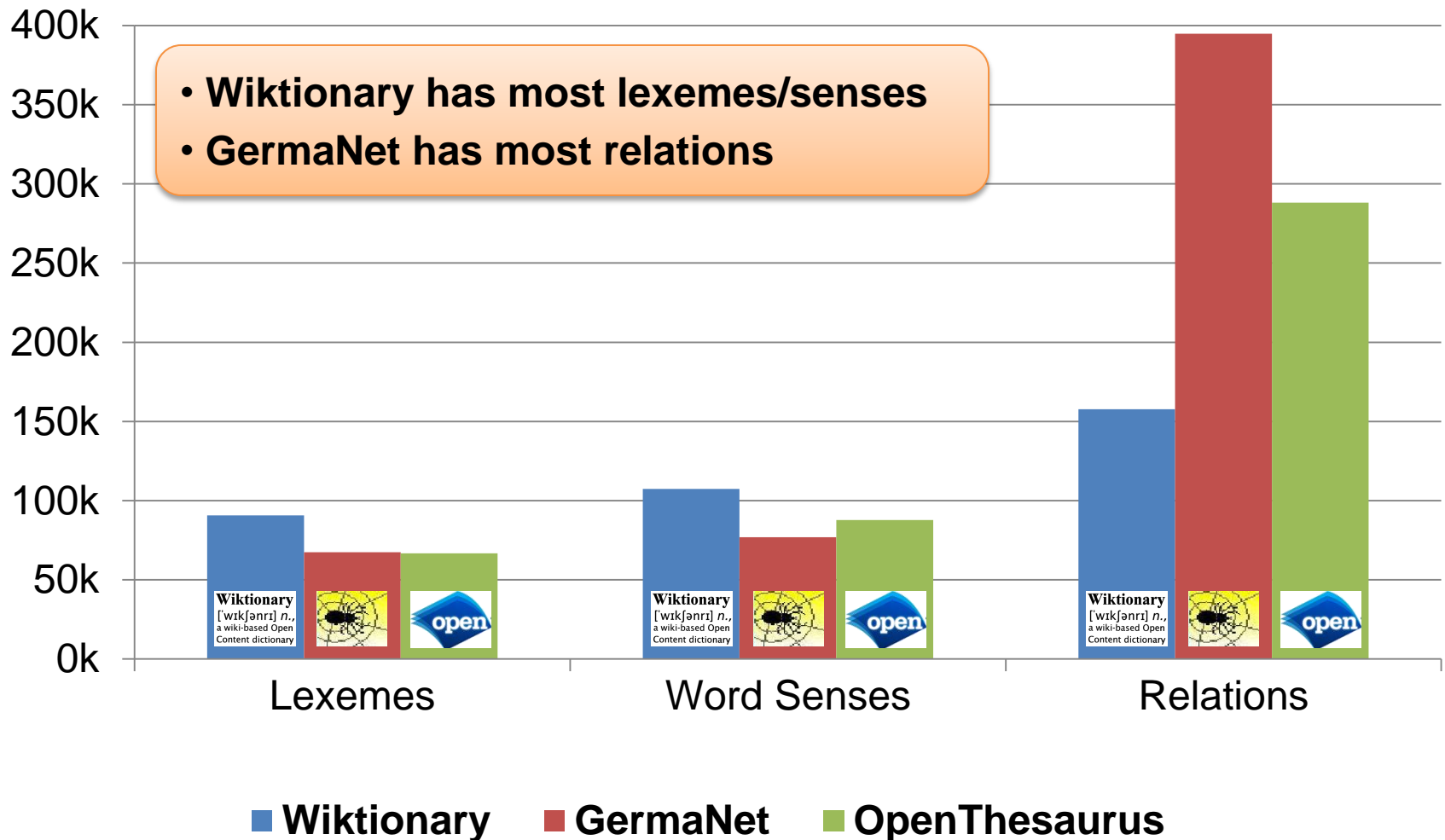
Results:

- 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

Resource Size



Content Analysis

Polysemy

	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

	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%	--		--	

boat

English

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

Pronunciation

- (RP) enPR: bōt, IPA: /bəʊt/, SAMPA: /bəʊt/
- (GenAm) enPR: bōt, IPA: /bəʊt/, SAMPA: /bəʊt/
- Audio (US)^{help}.file
- Rhymes: -əut


Etymology

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

Noun

Hyponyms

[1] ark, **bangca**, barge, canoe, boat, Dutch barge, East Indian boat, houseboat, hovercraft, hydrofoil, hydroplane, inflatable boat, inflatable raft, jetboat, jetski, junk, kyaaki,



Wiktionary

entry discussion citations create

Editing bangca

Wiktionary does not yet have an entry for bangca

- To start the entry, type in the box below and click visible immediately.
- If you are not sure how to format a new entry from templates to help you get started.
- If you are new to Wiktionary, please see [Help:Starting experiments](#). Also make sure your entry meets the

You are not logged in. If you save your edits to this page's history. For that and other reasons, you should

Content Analysis

Dangling Lexemes

	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...

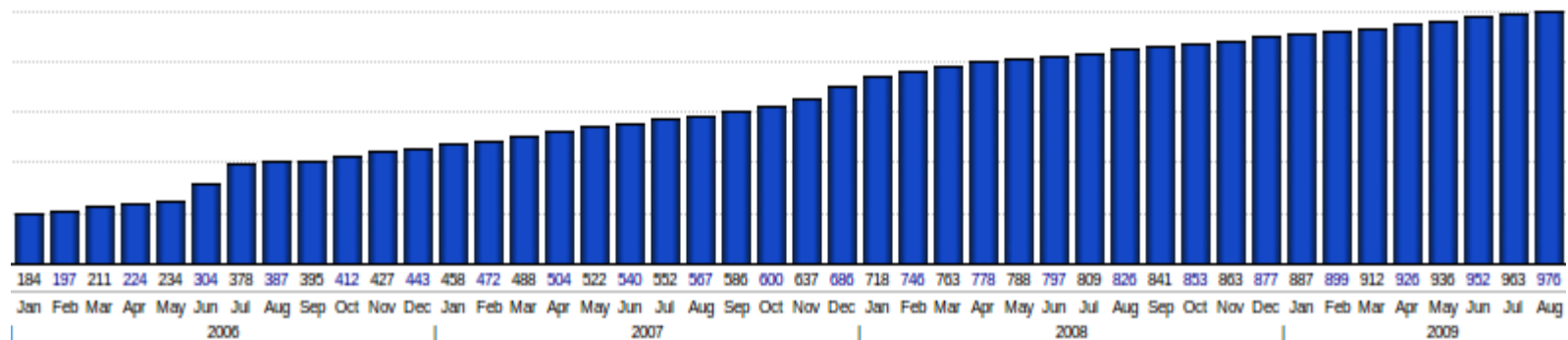
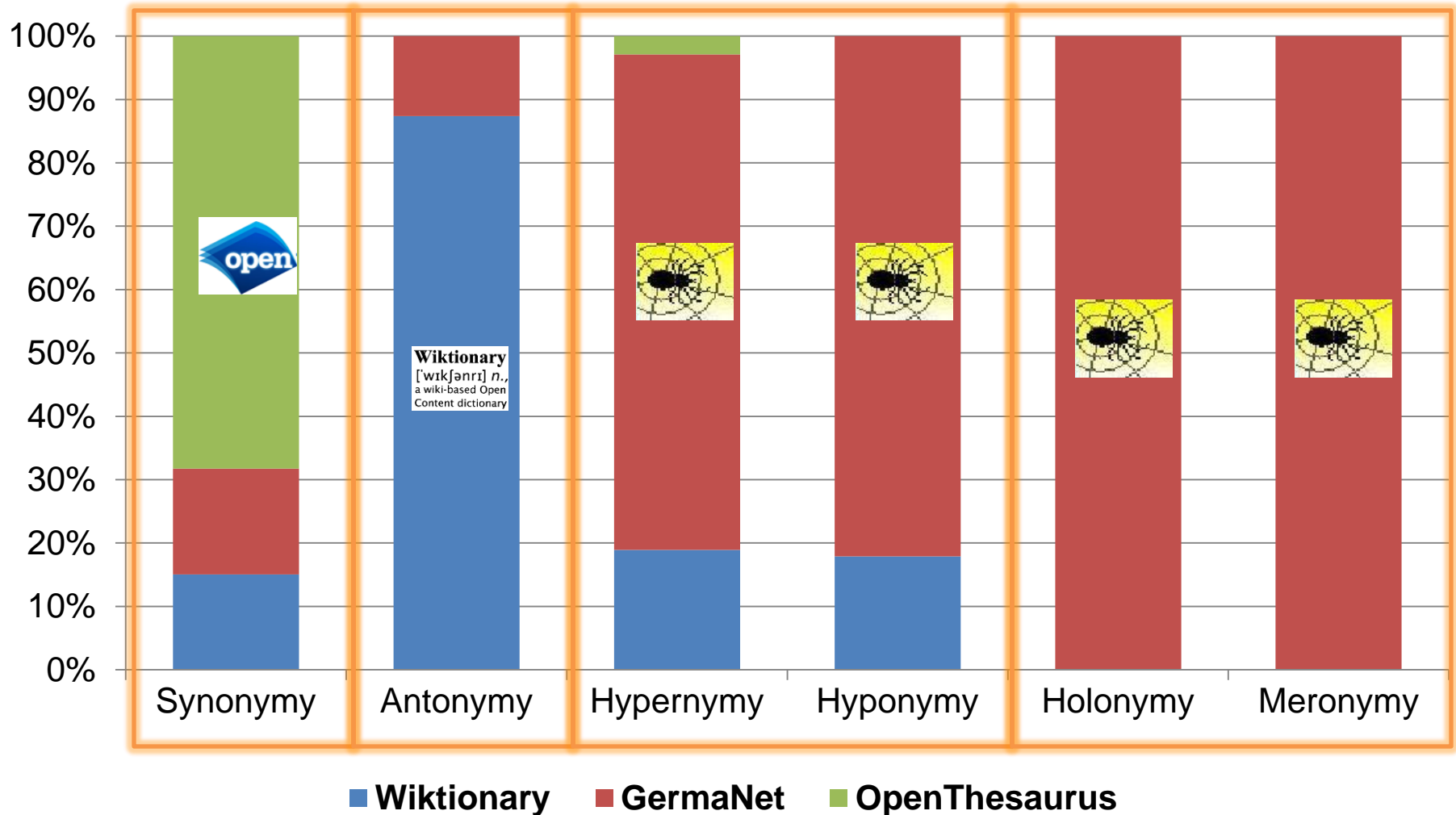


Fig. courtesy: <http://stats.wikimedia.org/wiktionary/EN/ChartsWikipediaDE.htm>

Content Analysis

Semantic Relations



Content Analysis

One Way Relations

	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

boat

English

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

Pronunciation

- (RP) enPR: bōt, IPA: /bəʊt/, SAMPA: /bəʊt/
- (GenAm) enPR: bōt, IPA: /boʊt/, SAMPA: /boʊt/
- Audio (US)^{help}·file
- Rhymes: -out


Etymology

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

Noun

Hyponyms

[1] ark, bangca, barge, canoe, boat, Dutch barge, East Indian boat, houseboat, hovercraft, hydrofoil, hydroplane, inflatable boat, inflatable raft, jetboat, jetski, junk, kyaaki,



canoe

See also canoeé

Contents [show]

English [edit]

Etymology [edit]

Adopted in 16th century from Spanish *canoa*, borrowed in turn by Columbus from Taino *kanoa* ("dugout canoe").

Pronunciation [edit]

- enPR: kə-ˈnoʊ, IPA: /kəˈnuː/, SAMPA: /kəˈnuː/
- Audio (US) (file)

Noun [edit]


Hyponyms

[1] boat

people (depending on the size of the canoe) sit in either the bow or the stern, in either the forward or the aft position, or kneeling on the bottom of the boat. Canoes are open on top, and pointed at both ends.

1. A small, narrow, pointed boat, usually made of wood, with a pointed prow and stern, used for transport or recreation. It is usually propelled by a single person (the paddler) sitting in the middle of the boat, using a double-bladed paddle. Canoes are open on top, and pointed at both ends.

2. (slang) An oversize, usually older, luxury car.



Conclusions

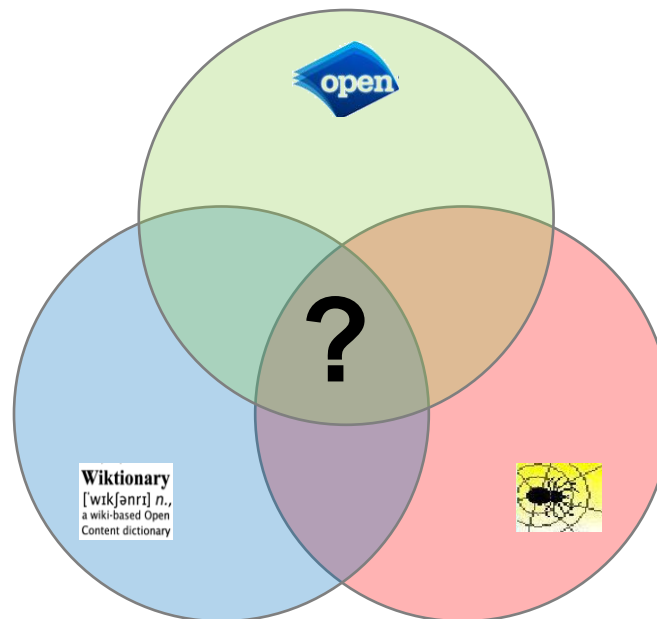
Take-home Message

- **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

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



Thank you for your attention!

Ubiquitous Knowledge Processing



KLAUS TSCHIRA STIFTUNG
GEMEINNÜTZIGE GMBH



e-learning
center of
research
excellence



Bundesministerium
für Wirtschaft
und Technologie



Additional Online Material:

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

Thank you for your attention!



TECHNISCHE
UNIVERSITÄT
DARMSTADT

UKP - Prof. Dr. Iryna Gurevych

People



**Ubiquitous
Knowledge
Processing**

Lexical Resources

UKP Lab > Data > Lexical Resources

[Home](#)
[e-NLP](#)
[NLP4Wikis](#)
[SIM](#)
[News](#)
[Partners](#)
[People](#)

Prof. Dr. Iryna Gurevych
Dr. Aljoscha Burchardt
Richard Eckart de Castilho
Dr. György Szarvas
Torsten Zesch
Daniel Bär
Joachim Caspar
Oliver Fersckke
Benjamin Herbert
Niklas Jakob

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. Iași, Romannia.
[PDF](#) | [BibTeX](#)

Additional Materials:

- [Resource Statistics](#)
- [Dataset of Word sense annotated Wiktionary relations](#)
- [Corresponding user guide for the annotators](#)

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, trademarks, 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.