Wiktionary
The Metalexicographic and the Natural Language Processing Perspective

Christian M. Meyer

Dissertation talk
boat

English

**Etymology**

From Middle English *boat*, *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 *bel†* ("boat").

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

**Pronunciation**

- (RP) enPR: bōt, IPA: /bɒt/, X-SAMPA: /b@t/
- Rhymes: -ōt
- (GenAm) enPR: bōt, IPA: /boʊt/, X-SAMPA: /boUt/
- Rhymes: -ōt
- Audio (US)

**Noun**

*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**

There's no explicit limit, but the word *boat* usually refers to a relatively small watercraft that is generally smaller than a "ship" and...
New lexicographic paradigm: collaborative lexicography
Motivation

Why care?

- The Web leads to quickly changing business models of dictionary publishers
  - unclear future of the Deutsches Wörterbuch (the “Grimm”)
  - downsizing of Duden incl. editorial staff

- Collaborative encyclopedias are about to replace expert-built ones
  - radical changes to Brockhaus encyclopedia

- Large amount of freely available information for language technology

New lexicographic paradigm: collaborative lexicography
The Metalexicographic Perspective

Research Question:
Implications of collaborative construction on lexicography

Wiktionary
[ˈwɪkʃənəri] n.,
a wiki-based Open Content dictionary
The Metalexicographic Perspective

Research Question:
Implications of collaborative construction on lexicography

Contributions:
- Comprehensive description of Wiktionary
- Analysis of the collaborative lexicographic process
- Quantitative comparison of dictionaries
The Natural Language Processing Perspective

Research Question:
Harvesting linguistic knowledge from Wiktionary

Wiktionary
[ˈwɪkʃənərɪ] n.,
a wiki-based Open Content dictionary
Research Question:
Harvesting linguistic knowledge from Wiktionary

Contributions:
- Extracting & disambiguating lexical-semantic information
- Alignment of Wiktionary and WordNet at sense level
- Standardized representation based on LMF
Research Goals

Metalexicographic Perspective

→ Dictionary development
→ New theoretical insights

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

Natural Language Processing Perspective

→ Resource development
→ Improved NLP systems
Interdisciplinary Relationships

Potential of collaboratively created dictionaries

Metalexicographic Perspective
- Dictionary development
- New theoretical insights

Natural Language Processing Perspective
- Resource development
- Improved NLP systems

Tools for qualitative and quantitative analysis
Agenda

Wiktionary
[ˈwɪkʃənri] n.,
a wiki-based Open
Content dictionary
Dictionary Analysis and Comparison

Multilingualism

Size
Dictionary Analysis and Comparison

Multilingualism

Size

Up-to-dateness
Dictionary Analysis and Comparison

Multilingualism

Size

Up-to-dateness

History
Dictionary Analysis and Comparison

### Sprachvarietäten
- **Dialekte:** Liste deutscher Dialekte • Hunsrücker Mundart • Kärntner Mundart • Wäller Platt
- **Außereuropäische Varietäten:** Außereuropäische Varietäten des Deutschen
- **Fachsprache:** Fliegensprache • Hip-Hop-Jargon • Jägersprache • Netzjargon Polizeijargon • Soldatensprache • Soldatensprache (Schweizer Armee)
- **Seemannssprache**

### Listen von Namen
- Liste von Staatenennamen
- Liste der Männlichen Vornamen
- Liste der Weiblichen Vornamen
- Liste der häufigsten weiblichen Vornamen Deutschlands
- Liste der häufigsten männlichen Vornamen Deutschlands
- Liste geschlechtsneutraler Vornamen
- Liste der Kosenamen und Kosewörter
- Liste der häufigsten Nachnamen Deutschlands / Österreichs

### Weitere Wortlisten
- Abkürzungen: Liste der Abkürzungen im Internet • ASCII-Smileys • Liste

### Wissenschaft
- Astronomie • Geowissenschaften • Informatik • Linguistik • Mathematik • Medizin • Pharmazie • Philosophie • Philosophie • Physik • Psychologie • Rechtswissenschaften
- **Biologie:** Ökologie • Zoologie • Zytologie und Genetik
- **Chemie:** Anorganische Stoffe • Organische Stoffe

### Gesellschaft
- Berufe • Politik
  - *Essen und Trinken:* Getränke • Speisen • Tätigkeiten und Eigenschaften • Deutschsprachige Küchenbegriffe nach Regionen • Küche und Küchengeräte • Weinbau
  - *Sport und Freizeit:* Sportarten • Fußball • Golf • Radsport • Schach • Tanzen
Overlap of Lexical Entries

surprisingly small overlap

Wiktionary
[ˈwɪkʃənəri] n.,
a wiki-based Open Content dictionary
Overlap of Lexical Entries

- Slang
- Dialect
- Natural sciences
- Computer science
- Neologisms
- Math
- Surprisingly small overlap
- Named entities
- Social sciences
- Humanities
- Biological taxonomy

Wiktionary [ˈwɪkʃənri] n., a wiki-based Open Content dictionary
But…

Wiktionary is far from being perfect

- Different developmental stages of articles
- Inconsistencies
- Spelling errors
- Old-fashioned and overly general definitions

No systematic corpus analysis

- Many references to expert-built dictionaries
- Sum of subjective opinions of the authors

A user suggests that this entry be cleaned up, giving the reason: “clean up music definitions”.

Please see the discussion on Requests for cleanup\(^{(+)}\) or the talk page for more information and remove this template after the problem has been dealt with.
Wiktionary
[ˈwɪkʃənri] n.,
a wiki-based Open Content dictionary
Harvesting Linguistic Knowledge from...

**Machine-readable Dictionaries?**
- Usually high quality
- Weakly structured; often not freely available

**Wordnets?**
- Commonly used; well-structured
- Small coverage, in particular multilingual ones
Harvesting Linguistic Knowledge from…

Machine-readable Dictionaries?
- Usually high quality
- Weakly structured; often not freely available

Wordnets?
- Commonly used; well-structured
- Small coverage, in particular multilingual ones

Wikipedia?
- High coverage of factual knowledge
- No linguistic knowledge (word usage, verbs, …)

Wiktionary!
- Freely available, high coverage, all parts of speech
Adding Structure to Wiktionary

Structure

Dictionary → Ontology

puppy

Contents [show]

English

Pronunciation
- Audio (US)
- Rhymes: -api

Noun
puppy (plural puppies)
1. A young dog.
2. A young rat.
3. A young seal.
4. (slang, usually in the plural) A woman's breast.
5. (informal) A (generic) thing: particularly something that
   I have another two dozen of these puppies to finish

{dog₁, hound₁, canine₁}

Concepts

{puppy₁, juvenile dog₁}

IS-A

Relation

OntoWiktionary

1. Concept formation
   - Start with Wiktionary senses
   - Create synonymy graph
   - Calculate the transitive hull
   - Synsets $\rightarrow$ concepts

2. Conceptual relations
   - From lexical relations to conceptual relations

3. Multilingual Lexicalizations
Disambiguation of Relations and Translations

**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, Nang)* A full house.
3. A *vehicle*, *utensil*, or *dish* somewhat resembling a boat in shape.
   - *a stone boat*; *a gravy boat*
4. *(chemistry)* One of two possible *conformations* of *cyclohexane* rings (the other being *chair*), shaped roughly like a boat.
5. *(Australia, politics, informal)* The refugee boats arriving in Australian waters, and by extension, refugees generally.

**Synonyms**

- *(craft on or in water)*: *craft*, *ship*, *vessel*

**Translations**

<table>
<thead>
<tr>
<th>Language</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrikaans</td>
<td><em>boot</em></td>
</tr>
<tr>
<td>Ainu</td>
<td>富士 (cip)</td>
</tr>
<tr>
<td>Albanian</td>
<td><em>varkë f</em></td>
</tr>
<tr>
<td>Lojban</td>
<td><em>bloti</em></td>
</tr>
<tr>
<td>Luxembourgish</td>
<td><em>please add this translation if you can</em></td>
</tr>
</tbody>
</table>
Disambiguation of Relations and Translations

**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. A **vehicle**, utensil, or dish somewhat resembling a **boat**; a **gravy** **boat**

4. *(chemistry)* One of two possible conformations shaped roughly like a boat.

5. *(Australia, politics, informal)* The refugee boat in shape.

**craft** *(plural craft or -)*

- The **skilled practice** of a practical **occupation**.
  
  *She represented the craft of brewers.*

- *(nautical, whaling)* Implements used in catching fish, such as **net**, **line**, or **hook**. Modern use primarily in whaling, as in **harpoons**, **hand-lances**, etc. [quotations ▼]

- *(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**.

- *(nautical, British Royal Navy)* Those vessels attendant on a **fleet**, such as **cutters**, **schooners**, and **gun-boats**, generally commanded by **lieutenants**.

- A **vehicle** designed for **navigation** in or on water or air or through outer space.

- A particular kind of skilled **work**.

---

**Synonyms**

- *(craft on or in water): craft, ship, vessel*

**Translations**

- **water craft**
  - Afrikaans: **boot**
  - Ainu: **チブ**
  - Albanian: **varkë f**
Disambiguation Features (Selection)

**Definition**

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

**Lemma of the Source Relation**

```
```

**Pragmatic Labels**

```
[2] (nautical, whaling) Implements used in catching fish, such as net, line, or hook as in harpoons, hand-liners, etc.
[3] (nautical) Boats that are primarily a loading or unloading of other vessels, as lighters, hoy, and barges.
[4] (nautical, British Royal Navy) Those vessels, such as cutters, generally commanded by lieutenants.
[5] A vehicle designed for navigation in or on water or air or through outer space.
```

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

```
[7] Shrewdness as demonstrated by being skilled in deception.
```
Disambiguation Features (Selection)

- **Definition**: The skilled practice of a practical occupation. She represented the craft of brewers.

- **Lemma of the source relation**: Boats

- **Pragmatic labels**: (nautical, whaling) Implements used in catching fish, such as net, line, or hook as in harpoons, hand-lines, etc.

- **Shared relations**: A vehicle designed for navigation in or on water or air or through outer space.

- **Inverse relation**: A particular kind of shrewdness.

- **Synonyms**: (skill at work): craftsmanship, workmanship
  - (nautical sense): boat
  - (vehicle):
    - (kind of skilled work): trade
    - (shrewdness): craftiness, cunning, foxiness, guile, slyness, wiliness

- **Hyponyms**
Evaluation Results

F1-score

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Random</th>
<th>MFS</th>
<th>Our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>en:en</td>
<td>0.65</td>
<td>0.74</td>
<td>0.82</td>
</tr>
<tr>
<td>de:de</td>
<td>0.71</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td>en:de</td>
<td>0.59</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>de:en</td>
<td>0.48</td>
<td>0.56</td>
<td>0.58</td>
</tr>
</tbody>
</table>

+0.05
+0.06
+0.02
+0.02
Overlap of Lexical Entries – Revisited

Wiktionary
[wɪkˈʃənri] n., a wiki-based Open Content dictionary
Agenda

Wiktionary
[ˈwɪkʃənəri] n., a wiki-based Open Content dictionary
**Word Sense Alignment**

**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
Aligning Wiktionary and WordNet

A two-step approach:
1. Candidate extraction
2. Candidate disambiguation

{plant, works, industrial plant}

WordNet synsets

plant (organism)
plant (factory)
plant (person)
reddish (color)
works (factory)
works (machine)
to fly (move)
bird (animal)

Wiktionary senses

[ˈwɪkʃənɔri] n., a wiki-based Open Content dictionary
Aligning Wiktionary and WordNet

A two-step approach:
1. Candidate extraction
2. Candidate disambiguation

WordNet synsets

{plant, works, industrial plant}

Wiktionary senses

plant (organism)

plant (factory)

plant (person)

works (factory)

works (machine)

bird (animal)

reddish (color)

to fly (move)

...
Aligning Wiktionary and WordNet

A two-step approach:
1. Candidate extraction
2. Candidate disambiguation

WordNet synsets

{plant, works, industrial plant}

Wiktionary senses

plant (organism)

bird (animal)
to fly (move)

reddish (color)

works (factory)

WordNet synsets

plant (factory)

plant (person)

works (factory)

WORKS (machine)
Candidate Disambiguation

- bag-of-words
- semantic relatedness measure
- bag-of-words

COS: Cosine similarity

PPR: Personalized PageRank

s < threshold → No alignment!

s ≥ threshold → Align this pair of WordNet synset and Wiktionary sense!
Alignment Evaluation

Accuracy: +.11
Precision: +.35
Recall: +.06
F$_1$ score: +.26
Integrated Resource UBY

Joint work of Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, Tri-Duc Nghiem
Integrated Resource UBY

Joint work of Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, Tri-Duc Nghiem
Agenda

Wiktionary
[ˈwɪkʃənəri] n., a wiki-based Open Content dictionary
Case Study: Detecting Marketing Blunders
Case Study: Detecting Marketing Blunders

Internationally advertised as “Silver Mist”
But: “Mist” means dung/manure in German!
Case Study: Detecting Marketing Blunders

Internationally advertised as “Silver Mist”
But: “Mist” means dung/manure in German!

Avoiding such blunders is challenging, since there are no tools and only few resources

Wiktionary covers many languages and language varieties (including slang, offensive and vulgar terms, etc.)
Case Study: Detecting Marketing Blunders

**Approach:**

- Index translations and pragmatic labels from OntoWiktionary
- Lookup word forms for carrying a problematic meaning in any language

<table>
<thead>
<tr>
<th>Mist</th>
<th>Animal excrement, especially that of common domestic farm animals [...]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mist</td>
<td>(uncountable) Manure; animal excrement.</td>
</tr>
<tr>
<td>Mist</td>
<td>(mildly vulgar, slang, in the plural) Nonsense.</td>
</tr>
<tr>
<td>mist</td>
<td>(uncountable) A thick cloud that forms near the ground [...]</td>
</tr>
<tr>
<td>mist</td>
<td>(countable) A layer of fine droplets or particles.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Case Study: Detecting Marketing Blunders

Approach:
- Index translations and pragmatic labels from OntoWiktionary
- Lookup word forms for carrying a problematic meaning in any language

<table>
<thead>
<tr>
<th>Mist</th>
<th>Animal excrement, especially that of common domestic farm animals [...]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mist</td>
<td>(uncountable) Manure; animal excrement.</td>
</tr>
<tr>
<td>Mist</td>
<td>(mildly vulgar, slang, in the plural) Nonsense.</td>
</tr>
<tr>
<td>mist</td>
<td>(uncountable) A thick cloud that forms near the ground [...]</td>
</tr>
<tr>
<td>mist</td>
<td>(countable) A layer of fine droplets or particles.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Evaluation:
- Copywriters become able to detect **35 out of the 45 marketing blunders**
  → Recall 0.78 (primary goal)
- They need to check **on average 12 clues per blunder**, of which 5 are relevant → Precision 0.41 (secondary goal)
Conclusion

Analyzing Wiktionary’s potential

- neologisms, translations, language varieties, language feeling
- subjective opinions rather than corpus evidence
Conclusion

Analyzing Wiktionary’s potential
- neologisms, translations, language varieties, language feeling
- subjective opinions rather than corpus evidence

Creating OntoWiktionary
- Automatic disambiguation of relations

Integrating language resources
- Aligning Wiktionary and WordNet
- Standardized representation

Detecting marketing blunders

Publicly available data and software:
- http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/
Future Outlook

Joint forces: experts and collaborative authors

Relation disambiguation:
- graph-based methods
- crowd-sourced disambiguation based on Wikidata

Word sense alignment:
- other information types
- more than two resources

Much potential for further NLP applications
Thank you for your attention!

Questions?
Kontakt / Contact

Christian M. Meyer
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

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.