Aligning Word Senses in GermaNet and the DWDS Dictionary of the German Language

Verena Henrich, Erhard Hinrichs, Reinhild Barkey

Department of Linguistics
University of Tübingen, Germany

Global WordNet Conference 2014
Identification and differentiation of word senses is one of the harder tasks that lexicographers have to face.

As a result, lexical resources display considerable variation in the number of word senses.

Lexicographic practice has undertaken considerable efforts to find external knowledge sources that can aid in distinguishing and identifying word senses:

- Very large electronic corpora
- Comparison with another semantic dictionary that has been constructed independently
Benefits of Aligning Lexical Resources

• For all sense distinctions that are completely parallel in two resources, such an alignment provides
  → Supporting evidence for the validity of sense distinction
  → Enriching word senses by information from another resource

• For all non-matching sense distinctions
  → Reason for revisiting and possibly revising the lexical entries
  → Suggestions for potentially missing senses
Different Methods for Constructing Word Meanings

- In common: both are long-term lexicographic projects aiming at a comprehensive coverage of contemporary standard German

- German wordnet (synsets, lexical units, relations)

- Based on three pre-existing dictionaries
- Revised and amended by information harvested from large electronic corpora
- Lexical entries are structured by the number of senses which may be further differentiated by an enumeration of subsenses
- Senses are accompanied by examples
Example: *Bau* (in DWDS)

**main senses**

- *Stelle, wo gebaut wird* ‘location where construction takes place’
- *Art, wie etwas gebaut ist, Gliederung, Struktur* ‘manner of how something is built, outline, structure’
- *das Bauen, Errichten* ‘the act of building, constructing’
- *das Gebaute, Errichtete* ‘the building, construction’

**sub-senses**

- i. *Gebäude* ‘building’
- ii. *Behausung von Säugetieren* ‘housing of mammals’
- iii. *Arrest* ‘imprisonment’
- iv. *künstlich hergestellter, unterirdisch verlaufender Hohlraum in der festen Erdrinde (Bergmannssprache)* ‘artificially constructed, subterranean space in the Earth’s solid crust (mining terminology)’
Example: *Bau* (in DWDS and GermaNet)

<table>
<thead>
<tr>
<th>GermaNet</th>
<th>DWDS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bau</strong></td>
<td></td>
</tr>
<tr>
<td>‘construction site’</td>
<td></td>
</tr>
<tr>
<td><strong>Bau</strong></td>
<td></td>
</tr>
<tr>
<td>‘act of building or constructing sth.’</td>
<td></td>
</tr>
<tr>
<td><strong>Bau</strong></td>
<td></td>
</tr>
<tr>
<td>‘building’</td>
<td></td>
</tr>
<tr>
<td><strong>Bau</strong></td>
<td></td>
</tr>
<tr>
<td>‘(animal) burrow’</td>
<td></td>
</tr>
<tr>
<td><strong>Bau</strong></td>
<td></td>
</tr>
<tr>
<td>‘prison’</td>
<td></td>
</tr>
</tbody>
</table>

- **Stelle, wo gebaut wird** ‘location where construction takes place’
- **Art, wie etwas gebaut ist, Gliederung, Struktur** ‘manner of how something is built, outline, structure’
- **das Bauen, Errichten** ‘the act of building, constructing’
- **das Gebaute, Errichtete** ‘the building, construction’
  - **a** **Gebäude** ‘building’
  - **b** **Behausung von Säugetieren** ‘housing of mammals’
  - **c** **Arrest** ‘imprisonment’
  - **d** **künstlich hergestellter, unterirdisch verlaufender Hohlraum in der festen Erdrinde (Bergmannssprache)** ‘artificially constructed, subterranean space in the Earth’s solid crust (mining terminology)’
Survey of the Overlapping Coverage

- 48,036 lemmas in both resources
  - 34,366 nouns
  - 7,735 verbs
  - 6,211 adjectives

- Explanations for apparently low overlap:
  - The history of the two resources
  - Different guidelines, for example, concerning the inclusion of regional, obsolete, technical, and colloquial terms as well as most recent contemporary language
  - The question of which compounds to include in a lexical resource is not trivial to answer
What is the Right Level of Senses and Subsenses?

• For the 48,036 lemmas that the two resources share
  - GermaNet distinguishes 59,495 senses
  - DWDS distinguishes 61,053 main senses

• The variability of how good the senses can be matched leads to a division into four classes (descending order according to their alignment appropriateness)

  1. Class 1: exact match of main senses
  2. Class 2: exact match of subsenses
  3. Class 3: partly overlapping coverage and different distinctions
  4. Class 4: distinct coverage
Class 1 Exact Match of Main Senses

- Pferd 1
  - ‘horse (animal)’
  - ‘riding animal with short ears and an already from the root long haired tail’
  - ‘gymnastics apparatus made from padded leather on four legs with two removable handles’
  - ‘chess: piece with stylized horse head, jumper, knight’

- Only exact matches occur for this lemma
Class 1 Exact Match of Main Senses

Pferd ‘horse (animal)’

[...] Reittier und Zugtier, das durch kurze Ohren und den schon von der Wurzel an lang behaarten Schwanz gekennzeichnet ist
‘riding animal with short ears and an already from the root long haired tail’

Examples:
- das Pferd hat den Reiter abgeworfen (‘the horse threw off its rider’)
- ein wildes, gezähmtes, dressiertes Pferd (‘a wild, tamed, trained horse’)
- es ist ein gutes, schnelles Pferd (‘it is a good, fast horse’)
- das Pferd geht im Schritt, trabt, galoppiert (‘the horse walks, trots, gallops’)
- die Pferde füttern, tränken, putzen, striegeln (‘feed, water, clean, groom the horses’)

• Only exact matches occur for this lemma
• All example sentences match
Class 2 Exact Match of Subsenses

- The overall coverage for these senses is the same
- The granularity level of the sense distinctions differs

**gebogenes Gerät** ‘curved device’

**Musik:** elastischer, mit [...] Haaren bespannter Holzstab, mit dem die Saiten der Streichinstrumente gespielt werden
‘music: flexible, wooden stick with hair stretched along it, for playing the strings of stringed instruments’

**Schusswaffe, die aus einem federnden Holzstab und einer straff gespannten Sehne besteht und mit der Pfeile abgeschossen werden**
‘flexible, wooden weapon with a bowstring for firing arrows’
Class 3 Partly Overlapping Coverage and Different Sense Distinctions

Example sentences:

endlose Wälder (‘endless forests’)  
der endlose Raum (‘the endless space’)  
der Krieg hatte endloses Leid gebracht (‘the war brought endless misery’)  
das endlose Gerede (‘the endless talk’)  
ein endloses Hin und Her (‘an endless back and forth’)  
ihr endloses Schweigen (‘her endless silence’)  
endlos scheinende Stunden (‘endlessly seeming hours’)  

...
Class 3 Partly Overlapping Coverage and Different Sense Distinctions

- **den Tod herbeiführend, zur Folge habend, übertragen: sehr groß, äußerst**
  - ‘to cause death, to have death as a consequence, figurative: very much, extremely’
  - Examples:
    - eine tödliche Krankheit (‘a deadly illness’)
    - ein tödlicher Unfall (‘a deadly accident’)
    - ein tödliches Gift (‘a deadly poison’)
    - tödlich verwundet werden (‘to be lethally wounded’)
    - tödliche Langeweile (‘deathly boring’)
    - tödlicher Ernst (‘deadly serious’)
    - mit tödlicher Sicherheit (‘with absolute certainty’)
    …

- In this example, the core meaning of the two senses is the same, but they are not completely identical in their coverage.
In this example, there is a sense in each resource that does not have a corresponding entry in the other resource.
Several Classes Can Occur for One Lemma

1. **Bau**
   - ‘construction site’

2. **Bau**
   - ‘act of building or constructing sth.’
   - ‘building’
   - ‘(animal) burrow’
   - ‘prison’

3. **Stelle, wo gebaut wird**
   ‘location where construction takes place’

4. **Art, wie etwas gebaut ist, Gliederung, Struktur**
   ‘manner of how something is built, outline, structure’

5. **das Bauen, Errichten**
   ‘the act of building, constructing’

6. **das Gebaute, Errichtete**
   ‘the building, construction’
   - **Gebäude**
     ‘building’
   - **Behausung von Säugetieren**
     ‘housing of mammals’
   - **Arrest**
     ‘imprisonment’

7. **künstlich hergestellter, unterirdisch verlaufender Hohlraum in der festen Erdrinde (Bergmannssprache)**
   ‘artificially constructed, subterranean space in the Earth’s solid crust (mining terminology)’
## Manual Alignment of 470 Lemmas (1,517 Senses)

<table>
<thead>
<tr>
<th>Class</th>
<th>Adj.</th>
<th>Nouns</th>
<th>Verbs</th>
<th>All POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1: main senses</td>
<td>126 (47%)</td>
<td>335 (53%)</td>
<td>250 (40%)</td>
<td>711 (47%)</td>
</tr>
<tr>
<td>Class 2: subsenses</td>
<td>36 (13%)</td>
<td>60 (10%)</td>
<td>110 (18%)</td>
<td>206 (14%)</td>
</tr>
<tr>
<td>Class 3: partly</td>
<td>92 (34%)</td>
<td>153 (24%)</td>
<td>220 (36%)</td>
<td>465 (31%)</td>
</tr>
<tr>
<td>Class 4: distinct</td>
<td>16 (7%)</td>
<td>81 (13%)</td>
<td>38 (6%)</td>
<td>135 (9%)</td>
</tr>
</tbody>
</table>

- Senses: 270, 629, 618, 1,517
- Senses/Lemma: 2.4, 3.1, 4.0, 3.2
- Lemmas: 113, 203, 154, 470

- Manual alignment allows to classify senses according to their alignment appropriateness (classes 1 to 4)
Discussion of the Results I

<table>
<thead>
<tr>
<th>Class</th>
<th>Adj.</th>
<th>Nouns</th>
<th>Verbs</th>
<th>All POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1: main senses</td>
<td>126 (47%)</td>
<td>335 (53%)</td>
<td>250 (40%)</td>
<td>711 (47%)</td>
</tr>
<tr>
<td>Class 2: subsenses</td>
<td>36 (13%)</td>
<td>60 (10%)</td>
<td>110 (18%)</td>
<td>206 (14%)</td>
</tr>
<tr>
<td>Class 3: partly</td>
<td>92 (34%)</td>
<td>153 (24%)</td>
<td>220 (36%)</td>
<td>465 (31%)</td>
</tr>
<tr>
<td>Class 4: distinct</td>
<td>16 (7%)</td>
<td>81 (13%)</td>
<td>38 (6%)</td>
<td>135 (9%)</td>
</tr>
</tbody>
</table>

- Classes 1 and 2 together arise in 61% of all cases
  → for three out of five word senses from GermaNet there is a matching sense in the DWDS

- Class 1 arises much more frequently than all other classes
  → The fact that class 1 outnumbers class 2 confirms the conception of word senses on the same granularity level
## Discussion of the Results II

<table>
<thead>
<tr>
<th>Class</th>
<th>Adj.</th>
<th>Nouns</th>
<th>Verbs</th>
<th>All POS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Class 1: main senses</strong></td>
<td>126 (47%)</td>
<td>335 (53%)</td>
<td>250 (40%)</td>
<td>711 (47%)</td>
</tr>
<tr>
<td><strong>Class 2: subsenses</strong></td>
<td>36 (13%)</td>
<td>60 (10%)</td>
<td>110 (18%)</td>
<td>206 (14%)</td>
</tr>
<tr>
<td><strong>Class 3: partly</strong></td>
<td>92 (34%)</td>
<td>153 (24%)</td>
<td>220 (36%)</td>
<td>465 (31%)</td>
</tr>
<tr>
<td><strong>Class 4: distinct</strong></td>
<td>16 (7%)</td>
<td>81 (13%)</td>
<td>38 (6%)</td>
<td>135 (9%)</td>
</tr>
</tbody>
</table>

- Both classes 3 and 4 reveal differences that prevent a straightforward sense alignment
  - Class 3: the lexicographers pursue different guidelines, e.g. with respect to the sense granularity
  - Class 4: indicates a distinct coverage
Conclusions and Future Work

• For all matching sense distinctions, the alignment
  → Provides supporting evidence for the validity of sense distinctions
  → Allows the enrichment of GermaNet senses by sense definitions
    and example sentences from the DWDS
  → Allows the enrichment of DWDS senses by lexical information
    about related words from GermaNet

• For all non-matching sense distinctions
  → Reason for revisiting and possibly revising the lexical entries
  → Suggestions for potentially missing senses

• Future work:
  - Automatic word sense alignment algorithm
  - Intelligent semantic search for the DWDS
Thank you.

Verena Henrich,
Erhard Hinrichs, and
Reinhild Barkey

Department of Linguistics
University of Tübingen

Wilhelmstr. 19
72074 Tübingen
Germany

verena.henrich@uni-tuebingen.de
http://www.verenahenrich.de