Consistency of Manual Sense Annotation and Integration into the TüBa-D/Z Treebank

Verena Henrich and Erhard Hinrichs
Department of Linguistics
University of Tübingen
Treebanks and Linguistic Theories (TLT13)
December 12, 2014
Introductory example

Sentence from TüBa-D/Z treebank
Introductory example

Sentence from TüBa-D/Z treebank

Target word *trauen* (two relevant word senses)

*trauen* – sense 1 ‘to trust someone’

*trauen* – sense 2 ‘to dare to do sth.’
Introductory example

The people trust me as monk

Sentence from TüBa-D/Z treebank

Target word *trauen*
(two relevant word senses)

*trauen* – sense 1
‘to trust someone’

*trauen* – sense 2
‘to dare to do sth.’
Introductory example

Sentence from TüBa-D/Z treebank

Target word *trauen* (two relevant word senses)

*trauen* – sense 1 ‘to trust someone’

*trauen* – sense 2 ‘to dare to do sth.’
Introductory example

Sentence from TüBa-D/Z treebank

because we us not into the place dare #2

Target word *trauen* (two relevant word senses)

*trauen* – sense 1
‘to trust someone’

*trauen* – sense 2
‘to dare to do sth.’
Introductory example

Sentence from TüBa-D/Z treebank

trauen – sense 1
‘to trust someone’

trauen – sense 2
‘to dare to do sth.’

because we us not into the place
dare

Senses are taken from GermaNet
(= sense inventory)
Why sense-annotating?

→ Disambiguating between word senses is important, for example, for machine translation (as the example shows)
Why sense-annotating ...the TüBa-D/Z?

→ Many existing linguistic annotations: POS, syntax, morphology, etc.
Why sense-annotating ...the TüBa-D/Z ...with senses from GermaNet?

• Sense inventory is necessary for word sense disambiguation

• Using a wordnet as the sense inventory is standard practice for other languages

• GermaNet contains many additional information
  - Relations to other word senses
  - Sense hierarchy
  - Translations (links to Princeton WordNet)
  - Links to Wiktionary
  - Compounds
  - Verbal frames
  - Example sentences
Which words to annotate? → Two main approaches

All words:
- All or nearly all words
- Limited size of running text

Lexical sample:
- All occurrences of a selected set of ambiguous word lemmas
- Chosen in advance
Which words to annotate? → Two main approaches

All words:
• All or nearly all words
• Limited size of running text

Lexical sample:
• All occurrences of a selected set of ambiguous word lemmas
• Chosen in advance

• Not enough training material for supervised machine learning methods

• Provides enough training material for supervised machine learning methods
Select word lemmas to be sense-annotated

• Criteria:
  Words have to be...
  - ambiguous $\rightarrow$ at least 2 senses in GermaNet
  - frequent $\rightarrow$ at least 21 occurrences in the TüBa-D/Z

<table>
<thead>
<tr>
<th></th>
<th>nouns</th>
<th>verbs</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of annotated word lemmas</td>
<td>30</td>
<td>79</td>
<td>109</td>
</tr>
<tr>
<td>Total number of tagged word tokens</td>
<td>8 803</td>
<td>9 107</td>
<td>17 910</td>
</tr>
<tr>
<td>Average frequency (occurrences/lemma)</td>
<td>293</td>
<td>115</td>
<td>164</td>
</tr>
<tr>
<td>Average polysemy (senses/lemma)</td>
<td>4.1</td>
<td>2.8</td>
<td>3.2</td>
</tr>
</tbody>
</table>
## 30 sense-annotated nouns

<table>
<thead>
<tr>
<th>nouns</th>
<th>Freq.</th>
<th>Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Frau</em>: ‘woman’, ‘wife’, ‘Ms’</td>
<td>1699</td>
<td>3</td>
</tr>
<tr>
<td><em>Mann</em>: ‘man’, ‘husband’, ‘manning level’</td>
<td>1114</td>
<td>3</td>
</tr>
<tr>
<td><em>Partei</em>: 3x ‘party’ (social, political, legal)</td>
<td>811</td>
<td>3</td>
</tr>
<tr>
<td><em>Stimme</em>: 2x ‘voice’ (human, opinion), ‘vote’, ‘pitch’</td>
<td>289</td>
<td>4</td>
</tr>
<tr>
<td><em>Mal</em>: ‘time’, ‘mark’</td>
<td>284</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
79 sense-annotated verbs

<table>
<thead>
<tr>
<th>verbs</th>
<th>Freq.</th>
<th>Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>heissen</em>: ‘call’, ‘intend’, ‘mean’, ‘tell’</td>
<td>801</td>
<td>4</td>
</tr>
<tr>
<td><em>sterben</em>: ‘to die’, ‘to decease’</td>
<td>220</td>
<td>2</td>
</tr>
<tr>
<td><em>unterstüztzen</em>: ‘support’, ‘back up’</td>
<td>188</td>
<td>2</td>
</tr>
<tr>
<td><em>bedeuten</em>: ‘mean’, ‘signify’, ‘indicate’</td>
<td>187</td>
<td>3</td>
</tr>
<tr>
<td><em>kündigen</em>: ‘to dismiss’, ‘to terminate’</td>
<td>39</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>verlesen</em>: ‘to read out’, ‘to read wrong’</td>
<td>21</td>
<td>2</td>
</tr>
</tbody>
</table>
Annotation process

• Initial sense annotation independently by 2 annotators:
  - Native Germans
  - Lemma-by-lemma
  - Pick one sense, if possible
  - Possibility to indicate problematic word occurrence
  - Supervised by experienced lexicographer
Annotation process

• Initial sense annotation independently by 2 annotators:
  - Native Germans
  - Lemma-by-lemma
  - Pick one sense, if possible
  - Possibility to indicate problematic word occurrence
  - Supervised by experienced lexicographer

• Adjudication by experienced lexicographer:
  - Native German
  - Goes through all occurrences
    ▪ where annotators do not agree
    ▪ at least one annotator had a comment
  - Resolves disagreements
Inter-annotator agreement (IAA)

<table>
<thead>
<tr>
<th></th>
<th>nouns</th>
<th>verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average inter-annotator agreement</td>
<td>96.4%</td>
<td>93.7%</td>
</tr>
<tr>
<td>Average Cohen’s kappa</td>
<td>85.4</td>
<td>82.4</td>
</tr>
</tbody>
</table>

- Recurring problems:
  - Sense distinction/granularity
  - Subsequent restructuring of the sense inventory
  - Transferred and idiomatic usages of the words
Problem: distinction of senses

Two word senses for *kündigen* in GermaNet:

- *kündigen* – sense 1
  ‘to dismiss someone’
  ‘to quit a job’

- *kündigen* – sense 2
  ‘to terminate/cancel a contract’

Example sentences:

*Der Chef *kündigt* dem Angestellten.*
‘The boss dismisses the employee.’

*Der Angestellte *kündigt* seinen Job.*
‘The employee quits his job.’

*Der Mieter *kündigt* den Vertrag.*
‘The tenant terminates the contract.’

*Der Angestellte *kündigt* seinen Vertrag.*
‘The employee terminates his contract.’
Problem: distinction of senses

Two word senses for *kündigen* in GermaNet:

- **kündigen** – sense 1
  - ‘to dismiss someone’
  - ‘to quit a job’

- **kündigen** – sense 2
  - ‘to terminate/cancel a contract’

Example sentences:

- Der Chef *kündigt* dem Angestellten.
  - ‘The boss dismisses the employee.’

- Der Angestellte *kündigt* seinen Job.
  - ‘The employee quits his job.’

- Der Mieter *kündigt* den Vertrag.
  - ‘The tenant terminates the contract.’

- Der Angestellte *kündigt* seinen Vertrag.
  - ‘The employee terminates his contract.’
Technical integration into the treebank

• TüBa-D/Z Treebank is released in several data formats

• We included the sense annotation into:
  - NeGra export format
  - Export XML format
  - CoNLL 2012

TüBa-D/Z release 9.1, December 2014, contains sense annotation
Technical integration into the treebank (NeGra)

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Structural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>ART np*</td>
</tr>
<tr>
<td>Leute</td>
<td>NN np* HD</td>
</tr>
<tr>
<td>trauen</td>
<td>VVFIN 3pis HD</td>
</tr>
<tr>
<td>mir</td>
<td>PPER ds*1 HD</td>
</tr>
<tr>
<td>als</td>
<td>KOKOM --</td>
</tr>
<tr>
<td>Mönch</td>
<td>NN dsm HD</td>
</tr>
<tr>
<td>.</td>
<td>$. --</td>
</tr>
<tr>
<td>#500</td>
<td>-- NX</td>
</tr>
<tr>
<td>#501</td>
<td>-- VXFIN</td>
</tr>
<tr>
<td>#502</td>
<td>-- NX</td>
</tr>
<tr>
<td>#503</td>
<td>-- NX</td>
</tr>
<tr>
<td>#504</td>
<td>-- VF</td>
</tr>
<tr>
<td>#505</td>
<td>-- LK</td>
</tr>
<tr>
<td>#506</td>
<td>-- NX</td>
</tr>
<tr>
<td>#507</td>
<td>-- MF</td>
</tr>
<tr>
<td>#508</td>
<td>-- SIMPX</td>
</tr>
</tbody>
</table>

#EOS 73702
Technical integration into the treebank (Export XML)

```xml
<sentence id="s73702">
  <node cat="SIMPX" func="--">
    <node cat="VF" func="-">
      <node cat="NX" func="ON">
        <word form="Die" pos="ART" morph="np*" lemma="die" func="-"/>
        <word form="Leute" pos="NN" morph="np*" lemma="Leute" func="HD"/>
      </node>
    </node>
  </node>
  <node cat="LK" func="-">
    <node cat="VXFIN" func="HD">
      <word form="trauen" pos="VVFIN" morph="3pis" lemma="trauen" func="HD" wsd-lexunits="77233"/>
    </node>
  </node>
  <node cat="MF" func="-">
    <node cat="NX" func="OD">
      <node cat="NX" func="HD">
        <word form="mir" pos="PPER" morph="ds*1" lemma="ich" func="HD"/>
      </node>
    </node>
  </node>
  ...
</sentence>
```
Technical integration into the treebank (CoNLL 2012)

<table>
<thead>
<tr>
<th>doc. ID</th>
<th>word no.</th>
<th>token</th>
<th>POS</th>
<th>parse bit</th>
<th>word sense</th>
<th>coref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>T890421</td>
<td>1</td>
<td>Die</td>
<td>ART</td>
<td>(VROOT:</td>
<td>--(SIMPX:</td>
<td>--(VF:-(NX:ON*</td>
</tr>
<tr>
<td>T890421</td>
<td>2</td>
<td>Leute</td>
<td>NN</td>
<td>*)</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>T890421</td>
<td>3</td>
<td>trauen</td>
<td>VVFIN</td>
<td>(LK:-(VXFIN:HD*))</td>
<td>77233</td>
<td>–</td>
</tr>
<tr>
<td>T890421</td>
<td>4</td>
<td>mir</td>
<td>PPER</td>
<td>(MF:-(NX:OD(NX:HD*))</td>
<td>–</td>
<td>(0)</td>
</tr>
<tr>
<td>T890421</td>
<td>5</td>
<td>als</td>
<td>KOKOM</td>
<td>(NX:*</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>T890421</td>
<td>6</td>
<td>Mönch</td>
<td>NN</td>
<td>*(</td>
<td>–</td>
<td>0)</td>
</tr>
<tr>
<td>T890421</td>
<td>7</td>
<td>.</td>
<td>$.</td>
<td>*</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Acknowledgements

• Reinhild Barkey and Valentin Deyringer for their help with the annotations

• Marie Hinrichs and Jianqiang Ma for their help with the technical integration into the TüBa-D/Z

• Yannick Versley for his support with the sense-annotation tool

• Scott Martens and Christian M. Meyer for their input on how to calculate inter-annotator agreement
Thank you.

Verena Henrich and Erhard Hinrichs

Department of Linguistics
University of Tübingen

Wilhelmstr. 19
72074 Tübingen
Germany

verena.henrich@uni-tuebingen.de
erhard.hinrichs@uni-tuebingen.de