Phrase processing for detecting collocations with KoKS*

- *Korpusbasierte Kollokationssuche (corpus based search for collocations)

**Used bilingual corpora**
- DE-News
  - radio news broadcast
  - translated by volunteers
- EU-publications
  - press releases
  - political documents
  - contracts
- the four Harry Potter books

**Alignment of sentences**

1/2

- distance measure
  - bilingual dictionaries
  - character trigram to identify cognats
  - sentence length

2/2

*It stared back.*

*Die Katze starrte zurück.*

- translation found in the dictionary

- open class words

- bilingual dictionaries
- character trigram to identify cognats
- sentence length

**System overview**

- contents
  - detection of phrases
  - identifications of collocations
  - evaluation (results)
detecting phrase correspondences 3/3
- POS tags sequences
  - extracted from chunk-parsed monolingual corpora
  - distinguished by syntactic category
- pair matching phrases
- example:

results (phrase detection) 1/3
- so far, we processed
  - all sentences with at most 19 words
  - approx. 70,000 sentence pairs
- next table shows examples
  - ordered by frequency (f)
results (phrase detection)  2/3

<table>
<thead>
<tr>
<th>rank</th>
<th>f</th>
<th>German</th>
<th>English</th>
<th>correspondence</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>0</td>
<td>be for the door</td>
<td>near</td>
<td>0.00</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>Professor</td>
<td>Professor Lupin</td>
<td>near</td>
</tr>
<tr>
<td>00</td>
<td>0</td>
<td>none</td>
<td>good</td>
<td>0.00</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>Harry</td>
<td>good</td>
<td>0.00</td>
</tr>
<tr>
<td>00</td>
<td>0</td>
<td>all</td>
<td>near</td>
<td>0.00</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>entire</td>
<td>seem to be</td>
<td>0.00</td>
</tr>
<tr>
<td>00</td>
<td>0</td>
<td>all</td>
<td>know</td>
<td>0.00</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>sight (when)</td>
<td>near</td>
<td>0.00</td>
</tr>
<tr>
<td>00</td>
<td>0</td>
<td>nomenclature</td>
<td>meaningless</td>
<td>0.00</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>German</td>
<td>good</td>
<td>0.00</td>
</tr>
</tbody>
</table>

results (phrase detection)  3/3

• candidate set with $f > 6$
  - does not contain any collocations according to Breidt (human annotators)
  - a lot of compositional compounds
  - only a few non-compositional translations
• useless to apply collocativity measure

results (collocativity measure)  1/6

• manually aligned phrase pairs
  - 250 phrase pairs
  - 83 with non-compositional translation
  - 45 with non-compositional semantics (Breidt’s definition of collocation)
  - agreement of two annotators
  - 31 unresolved disagreements

results (collocativity measure)  2/6

<table>
<thead>
<tr>
<th>variant</th>
<th>ignores words with high $f$</th>
<th>uses length of phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>no</td>
<td>only if very different</td>
</tr>
<tr>
<td>01</td>
<td>no</td>
<td>always</td>
</tr>
<tr>
<td>10</td>
<td>yes</td>
<td>only if very different</td>
</tr>
<tr>
<td>11</td>
<td>yes</td>
<td>always</td>
</tr>
</tbody>
</table>

results (collocativity measure)  3/6

<table>
<thead>
<tr>
<th>precision (compositional translation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>measure 00</td>
</tr>
<tr>
<td>measure 01</td>
</tr>
<tr>
<td>measure 10</td>
</tr>
</tbody>
</table>

results (collocativity measure)  4/6

<table>
<thead>
<tr>
<th>recall (compositional translation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>measure 00</td>
</tr>
<tr>
<td>measure 01</td>
</tr>
<tr>
<td>measure 10</td>
</tr>
</tbody>
</table>
results (collocativity measure) 5/6

University of Osnabrück (Germany): oKS-Project

outlook 1/2

• improve phrase correspondences
  - use proper chunking to find phrases
  - use word alignment
  - weight phrase pairs according to their correspondence probability
  - replace simple counts with advanced statistics (associations measure)
  - exploit substring relations among phrases

discussion / questions / contact

• Norman Kummer, norman@VauDePe.de
• Joachim Wagner, jowagner@uos.de

University of Osnabrück
Institute of Cognitive Science
49078 Osnabrück
Germany

Link:
http://www.cl-ki.uos.de/~koks/
alignment of sentences (extra 1)

It stared back.

University of Osnabrück (Germany): oKS-Project

alignment of sentences (extra 2)

Die Katze starre zurück.

University of Osnabrück (Germany): oKS-Project

system overview

University of Osnabrück (Germany): oKS-Project

application

• CALL-context
• provides help to L2 learner in text understanding
• web based interface

University of Osnabrück (Germany): oKS-Project

other possible applications

• intelligent lexicon lookup (iKoKS)
• translation memory in CAT (computer assisted translation)
• full text search based on the lemmas

University of Osnabrück (Germany): oKS-Project

application

• current KoKS demo application (screen-shot)

University of Osnabrück (Germany): oKS-Project

• provides help to L2 learner in text understanding

University of Osnabrück (Germany): oKS-Project

• intelligent lexicon lookup (iKoKS)
• translation memory in CAT (computer assisted translation)
• full text search based on the lemmas

University of Osnabrück (Germany): oKS-Project