Contrastive Collocation Analysis – a Comparison of Association Measures across Three Different Languages Using Dependency-Parsed Corpora

Stefan Evert\textsuperscript{1}, Thomas Proisl\textsuperscript{1}, Peter Uhrig\textsuperscript{1}, Maria Khokhlova\textsuperscript{2}

\textsuperscript{1} University of Erlangen-Nuremberg
\textsuperscript{2} St. Petersburg State University
Approaches to Collocation Extraction

Collocation:
combination of two lexical items as listed in collocations or explanatory dictionaries.

Evert 2004:
• segment-based co-occurrences;
• distance-based co-occurrences;
• relational co-occurrences.
Analysis

- English, German, Russian;
- dependency-annotated corpora;
- binary collocations (adj-noun);
- association measures
Data

• DECOW16A (Schäfer & Bildhauer, 2012);
• ENCOW16A (Schäfer, 2015);
• Araneum Russicum II Maximum (Benko, Zakharov, 2016);
• dependency parsed corpora:
  – English: spaCy;
  – German: mate-tools
  – Russian: UDPipe 1.20
Gold Standards

• English: Oxford Collocations Dictionary for Students of English, 2\textsuperscript{nd} ed. 2009 (OCD2);
• German: Wörterbuch der Kollokationen im Deutschen, 2011 (WdK)
Collocations from Gold Standards

- 86,563 adj-noun pairs extracted from the English gold standard;
  - Examples:
    - heavy smoker;
    - Western democracy;
    - inner demon;
- 105,872 adj-noun pairs extracted from the German gold standard;
  - Examples
    - farbig + Abbildung
    - radioaktiv + Abfall
    - stattlich + Abfindung
Collocations from Gold Standards

• 3,387 adj-noun pairs extracted from the English gold standard;

• Examples:
  – академический год
  – большой палец
  – декретный отпуск
  – медовый месяц
  – периодическая система
  – сухой закон
Association Measures

• Main cue: co-occurrence frequency, quantified by various statistical association measures.
• 84 measures (Pecina 2005); 47 measures (Wiechmann 2008).
• Evaluated by (Evert et al., 2017) on a smaller English gold standard:
  – MI;
  – MI²;
  – t-score;
  – Dice;
  – X²;
  – log-likelihood
Evaluation

• precision-recall curves;
• precision (P): the percentage of true positives among the n candidates;
• recall (R): the percentage of all true positives in the gold standard found in the n-best list;
• the “higher” a P/R graph is located in the plot, the better the ranking achieved by the corresponding association measure;
• AP50: a composite measure for average precision, recall = 50%
Results for English Adj-Noun Collocations
Results for English Adj-Noun Collocations

- the best overall measure: log-likelihood;
- coverage 78.5%

<table>
<thead>
<tr>
<th>l1</th>
<th>l2</th>
<th>b.TP</th>
<th>G^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>year_N</td>
<td>last_J</td>
<td>TRUE</td>
<td>7115806,250</td>
</tr>
<tr>
<td>device_N</td>
<td>electronic_J</td>
<td>TRUE</td>
<td>4454791,250</td>
</tr>
<tr>
<td>range_N</td>
<td>wide_J</td>
<td>TRUE</td>
<td>3604840,760</td>
</tr>
<tr>
<td>information_N</td>
<td>more_J</td>
<td>FALSE</td>
<td>3527524,720</td>
</tr>
<tr>
<td>week_N</td>
<td>last_J</td>
<td>TRUE</td>
<td>3012003,080</td>
</tr>
<tr>
<td>party_N</td>
<td>third_J</td>
<td>FALSE</td>
<td>2734559,170</td>
</tr>
<tr>
<td>navigation_N</td>
<td>usual_J</td>
<td>FALSE</td>
<td>2657799,180</td>
</tr>
</tbody>
</table>
Results for German Adj-Noun Collocations
Results for German Adj-Noun Collocations

• the best overall measure: Dice / log-likelihood;
• coverage 95.5%

<table>
<thead>
<tr>
<th>l1</th>
<th>l2</th>
<th>b.TP</th>
<th>G²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jahr_N</td>
<td>vergangen_J</td>
<td>TRUE</td>
<td>3524676,622</td>
</tr>
<tr>
<td>Zeit_N</td>
<td>kurz_J</td>
<td>TRUE</td>
<td>2728721,674</td>
</tr>
<tr>
<td>Zeit_N</td>
<td>lang_J</td>
<td>TRUE</td>
<td>2341170,448</td>
</tr>
<tr>
<td>Jahr_N</td>
<td>nah_J</td>
<td>FALSE</td>
<td>2029583,119</td>
</tr>
<tr>
<td>Stunde_N</td>
<td>halb_J</td>
<td>FALSE</td>
<td>1819704,642</td>
</tr>
<tr>
<td>Energie_N</td>
<td>erneuerbar_J</td>
<td>TRUE</td>
<td>1685377,270</td>
</tr>
<tr>
<td>Tag_N</td>
<td>nah_J</td>
<td>FALSE</td>
<td>1555725,796</td>
</tr>
</tbody>
</table>
Results for Russian Adj-Noun Collocations
Results for Russian Adj-Noun Collocations

- the best overall measure: $MI^2 / X^2$;
- coverage 90.2%

<table>
<thead>
<tr>
<th>l1</th>
<th>l2</th>
<th>b.TP</th>
<th>G²</th>
</tr>
</thead>
<tbody>
<tr>
<td>дело Н</td>
<td>самый J</td>
<td>TRUE</td>
<td>10097423,630</td>
</tr>
<tr>
<td>день Н</td>
<td>сегодняшний J</td>
<td>TRUE</td>
<td>8772634,170</td>
</tr>
<tr>
<td>директор Н</td>
<td>генеральный J</td>
<td>FALSE</td>
<td>7090687,977</td>
</tr>
<tr>
<td>год Н</td>
<td>прошлый J</td>
<td>FALSE</td>
<td>6815974,206</td>
</tr>
<tr>
<td>участок Н</td>
<td>земельный J</td>
<td>FALSE</td>
<td>6805317,506</td>
</tr>
<tr>
<td>плата Н</td>
<td>заработный J</td>
<td>TRUE</td>
<td>5713262,133</td>
</tr>
<tr>
<td>сад Н</td>
<td>детский J</td>
<td>TRUE</td>
<td>5176036,660</td>
</tr>
</tbody>
</table>
Average Precision at 50% (AP50)
Conclusions and Further Work

• Log-likelihood is the best association measure for this type of collocation for English and German but not for Russian.
• For some specific types other measures can perform even better.
• Carefully sampled and balanced corpora seem to have considerable advantages in precision.
• Larger, less balanced corpora (such as the web corpora) can be more useful for total coverage.
• Further enlargement of Russian gold standard.
Thank you!