Extraction of potential multi-word expressions in a parallel corpus of their translations from a multilingual corpus

WG 2: Parsing Techniques for MWEs

Katerina Zdravkova, University Sts Cyril and Methodius, Skopje
Aleksandar Petrovski, International Slavic University, Sveti Nikole

Basic setting

Orwell’s 1984 parallel corpora
Resources of 16 languages belonging to Multext-East V.4. (courtesy of Institute Jozef Stefan, Ljubljana)

Orwl-mk.txt

Orwl-mk.xml

Main module

The core part

Separation of all the repeated blocks of words

15463 blocks of words (potential MWEs)
They appeared in total 53246 times
Average length: 3.44 words / block

Generated smaller blocks

<table>
<thead>
<tr>
<th>length of the block</th>
<th>frequency</th>
<th>generated smaller blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>39 words</td>
<td>2</td>
<td>741</td>
</tr>
<tr>
<td>24 words</td>
<td>2</td>
<td>276</td>
</tr>
<tr>
<td>20 words</td>
<td>2</td>
<td>190</td>
</tr>
<tr>
<td>17 words</td>
<td>2</td>
<td>136</td>
</tr>
<tr>
<td>12 words</td>
<td>2</td>
<td>66</td>
</tr>
<tr>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
</tbody>
</table>

Removing the blocks generated by longer blocks

4690 blocks of words (potential MWEs)
They appeared in total 7621 times
Average length: 2.07 words / block

Part of extracted MWEs

Frequency of extracted MWEs

Filtering of nominal phrases

Adj N (безнадежна любов - beznadezhna ljubov = hopeless love)
Adj Adj Noun (друго човечко существо - drugo chovecko sushestvo = another human being)
Adj N N (мал број људи - mal broj ludje = very few people)
N N (безумие безумие - bezumie bezumie = folly folly)
Adj N Prep N (бунтовно око на полицијата - bunтовno oko na policijata = under the eyes of the police)
Adj N Adv (неколку минути подоцна - nekolku minuti podocna = after only a few minutes' delay)
Adj N Prep Adj N (дневната заповед на големия брат - dnevnta zapoved na metalni brat = big brother's order for the day)
Adv N Adv N (повече храна повече облека - povrce hrana povrce obleka = more food more clothes)

Blocks that contain a verb