Automatic recognition and extraction of multiword nominal expressions from corpora

Angeliki Fotopoulou, Giorgos Giannopoulos, Maria Zourari, and Marianna Mini

Institute for Language and Speech Processing, “Athena” RIC

afotop@ilsp.athena-innovation.gr giann@imis.athena-innovation.gr

Cost PARSEME

WP2
Method

Nominal multiwords: παιδική χαρά [lit. “kids’ joy”, meaning “playground”], ψήφος εμπιστοσύνης [“vote of confidence”]

Tools
Corpus: ~142,000,000 words (930 MB)
ILSP morphological electronic dictionary
Ngram Statistics Package tool (Pedersen et al., 2006)

Method
(a) Six (6) general grammar rules are applied to a grammatically tagged corpus
(b) Results are filtered using more specific grammar rules and filters (word lists)
(c) Results are evaluated using statistical methods
(d) A linguist/encoder makes the final selection and MWE are stored in a Database
## Results

<table>
<thead>
<tr>
<th>Step Description</th>
<th>Adj + N</th>
<th>N + N(gen.)</th>
<th>N + N(gen.)</th>
<th>N + - + N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Step: Application of 6 general rules</td>
<td>1.068.000</td>
<td>242.000</td>
<td>619.000</td>
<td>120.000</td>
</tr>
<tr>
<td>2nd Step: Application of specific rules filters</td>
<td>570.000</td>
<td>189.000</td>
<td>347.000</td>
<td>12.000</td>
</tr>
<tr>
<td>3rd Step: Application of statistical metrics</td>
<td>68.000</td>
<td>22.000</td>
<td>31.000</td>
<td>400</td>
</tr>
</tbody>
</table>
Database for MWEs

[Image of a software interface for a database]