A Hybrid Multilingual Method to Extract Collocations from Corpora

Amalia Todiraşcu¹

¹ LILPA, Université de Strasbourg, France
todiras@unistra.fr

WG 3: Statistical, hybrid and multilingual extraction methods

The project

• A hybrid system for collocation extraction, based on contextual morpho-syntactic properties (Heid & Ritz 2005), (Oodjik, 2013), (Nissim, Zanninello, 2013), (Krenn, 2000):

1. Statistical extraction from tagged and lemmatized corpora

2. Linguistic analysis of morpho-syntactic properties and semantic properties (Functional Systemic Grammar (Halliday, 1985))

* System available for several languages (French, Romanian)

Collocations

• Definition: multiword expressions, not necessarily continuous, with specific syntactic and semantic behaviour (Gledhill, 2007) (Williams, 2003)

• Several criteria:
  A) statistic criteria: collocations are word cooccurrences (frequent word associations) (Sinclair, 1991);
  B) linguistic criteria: the words composing the collocations are linked by syntactic relations (Hausmann, 2004) (Tutin, 2010);
  C) pragmatic: collocations are used in appropriate contexts

Contextual properties

- collocation property: type of construction: V+NP, V+PP
- noun properties: determination, number, modifiability
- verb properties: preference for certain forms, for voice

V-lemma N-lemma Det Case Prep Nb Voice Process
faire 'make' Objet 'subject' definite - - sg active Range
Aduce make Atingere 'touch' null acc - sg active Range
Mette 'put' Jeu 'game' null - En'into' sg active Range
Lua 'take' Decizie 'decision' Definite, indefinite, null acc - sg, pl Active, passive Mental
Satisfaire 'satisfy' Contrainte 'constraint' Definite, indefinite - - sg, pl Active, passive Mental

Monolingual extraction (Todirascu, et al. 2009)

- statistical language-independent extraction module + filtering module (Romanian, French)
- Statistical extraction of candidates from tagged corpora using the following properties
  - stable distance
  - cooccurrence significance : Loglikelihood (Dunning, 1990) ;
- Language-specific filters eliminating invalid candidates, based on contextual morpho-syntactic properties ;

Filters

- Romanian:[pos=V*] [pos=NSRY] [{lemma=*}] [pos=NxN|NxOx]
- NSRY = def. noun (sg), acc./nom.;  NxF = noun; {} – option; NxF = dative ;
- Examples: a face objectul"be subject of", a ține seamă"take into account";
- French: [pos=Vmn]Vmn(|is)p-- [pos=Da---] [pos=Ne-] [{pos=A---}]
- NSRY = definite noun (sg), acc./nom.;  NxF = noun; {} - optional
- Examples: prendre des mesures"take measures", donner un avis"give an opinion";

Evaluation

- Comparison with existing collocation dictionaries :
  Romanian : Multilingual collocation dictionary (Todirascu et al, 2008)
  84 % of candidates
- Comparison between several corpora from various domains (for Verbal+Noun collocations)

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Law texts</th>
<th>Law texts</th>
<th>News</th>
<th>News</th>
<th>Med</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FR</td>
<td>RO</td>
<td>FR</td>
<td>RO</td>
<td>FR</td>
</tr>
<tr>
<td>collocations</td>
<td>7.8%</td>
<td>10.4%</td>
<td>9.1%</td>
<td>11.3%</td>
<td>8.7%</td>
</tr>
<tr>
<td>V+complement</td>
<td>22%</td>
<td>23.9%</td>
<td>37.01%</td>
<td>29.08%</td>
<td>28.19%</td>
</tr>
<tr>
<td>S+P, N+VPP</td>
<td>32.1%</td>
<td>25.4%</td>
<td>21.42%</td>
<td>13.49%</td>
<td>24.5%</td>
</tr>
<tr>
<td>Noise</td>
<td>38.1%</td>
<td>33.3%</td>
<td>32.47%</td>
<td>46.13%</td>
<td>38.61%</td>
</tr>
</tbody>
</table>

Frequent Errors

- complements of the multiword expressions wrongly identified as MWE parts: informeze Comisia cu privire la ... ("to inform the Commission concerning...");
- subject+predicate combinations: La Comission, propose, la modification des dispositions legales... ("the Commission proposes the following change of the law provisions...");
- predicate+adjunct combinations: la surveillance de ces patients, est faite, sur plusieurs années ("The survey of these patients has been done for several years") articolul a fost modificat, ultima data, ("The article has been modified last time...", correct modifica+articol);
- mistagging adjectives as verb participles: ces patients, immunodéprimés, ("these imunodepressed patients") discuta,plăcută, "the pleased discussion";