A Collocation Extraction Tool for Romanian

Violeta Seretan, Eric Wehrli, Luka Nerima (University of Geneva),
Amalia Todirascu (University of Strasbourg)

- Aims : corpus-based extraction of collocations for the Romanian language to complete an existing dictionary
  - Method : syntactically-bound lexical combinations extracted from parsed data and statistical measures
    - FIPS parser (Wehrli et Nerima, 2015) for Romanian (Seretan *et al.*, 2010)
    - Extension of the multilingual collocation extraction tool, FipsCo (Seretan, 2008) to Romanian, applied to Europarl corpus (2011)

sentences	54453	*aduce + atingere,* lit. *bring touch*
words	1233996	*atinge + obiectiv*, lit. *touch objective*
average sentence length	22.7 words	*gaz + cu + efect de seră*, lit. *gas with effect of*
fully parsed sentences (%)	6169 (11.3%)	*greenhouse*
candidates: types; in lexicon	82829; 19.6%	*încheia + accord*, lit. *end agreement*
candidates: tokens; in lexicon	194654; 0.2%	*întâmpina + dificultate*, lit. *welcome difficulty*
candidates (LLR > 10): tokens	96225	*motiv + întemeiat*, lit. *valid reason*
candidates (LLR > 10): types	13844	*om + de + ştiință,* lit. *man of science*
valid candidates in top 1000 (%)	677 (67.7%)	*trage + semnal de alarmă,* lit. *pull a signal of alarm*

- Evaluation : 1 annotator, top 1000, precision 67.7%
- Comparison with a Romanian collocation dictionary (Todirascu *et al,* 2008)