A Hybrid Multilingual Method to Extract Collocations from Corpora

Amalia Todirascu

FDT, LILPA, Université de Strasbourg, France

todiras@unistra.fr

WG3: Statistical, Hybrid and Multilingual Processing of MWEs

Goals

- **A hybrid method to extract collocations from monolingual corpora combining statistical methods and linguistic filters (Todirascu et al, 2009)**
  - **Available for several languages (French, Romanian), easily adaptable for other languages**
  - **Tagged and lemmatized corpora with TTL (Ion, 2007) (Todirascu et al, 2011)**
- **Collocations :**
  - **composed of two or several lexical units**
  - **specific syntactic and semantic behaviour**
- Several criteria
  - statistical criteria : frequent word cooccurrences (Sinclair,1991);
  - linguistic criteria: morpho-syntactic and syntactic properties (Haussmann, 2004) (Tutin, 2010) (M.Gross, 1993);
  - pragmatic : collocations are used in appropriate context.

The Method

Hybrid methods (Nissim, Zanninello, 2013) (Krenn, 2000), (Smadja, 1993)

1) Statistical extraction

- Stable distance
- Frequent word association (Loglikelihood (Dunning 1990))

2) Filters based on linguistic analysis (Halliday, 1985)

- Fixedness (preference for number, gender, voice)
- Semantic criteria (process)

Experiments

- Several Verb+Noun candidates extracted from large parallel and comparable corpora
  - Various classes of candidates

	Parallel law corpora		Comparable newspaper corpora		Medicine
	FR	RO	FR	RO	FR
colloc.	7,8%	10,4%	9,1%	11,3%	8,7%
V+comp.	22%	23,9%	37,01%	29,08%	28,19%
S+V, NP +VPP	32,1%	25,4%	21,42%	13,49%	24,5%
Other	38,1 %	33,3%	32,74%	46,13%	38,61%

- Comparing the results with existing dictionaries
  - French (BLF(Verlinde *et al*, 2003), LADL tables (M.Gross, 1993, Laporte, 2008)), Romanian (Todirascu et al, 2008)