A Hybrid Multilingual Method to Extract Collocations from Corpora

Amalia Todiraşcu[1]

[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 (Haussmann, 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
Mettre '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

Frequent Errors

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

Methodology

- Preprocess monolingual corpora (tagger and lemmatizer (TTL (Ion, 2007), (Todirascu, *et al*, 2011)))
  - French corpora (medicine, law, newspapers)
  - Romanian corpora (newspapers, law)
- Statistical extraction of candidates
- Linguistic analysis of collocations properties (Verb+Noun collocations)
- Filter definition on the basis of linguistic analysis
- Manual evaluation (semantic criteria)

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.; NxN = noun; {} – option; NxOx = dative ;
  - Examples: a face obiectul/'be subject of', a ţine seama/'take into account';
- **French: [pos=Vmn|Vm(i|s)p--] [pos=Da---] [pos=Nc--] {[pos=A----]}**
  - NSRY = definite noun (sg), acc./nom.; NxN = noun; {} - optional
  - Examples: prendre des mesures/'take measures', donner un avis/'give an opinion' ;

Evaluation

- Comparison with existing collocation dictionaries :

French : BLF (Verlinde, et al, 2003) 93 % of candidates, LADL tables (Gross, 1993, Laporte et al, 1998) 96 % of candidates

Romanian : Multilingual collocation dictionary (Todirascu *et al*, 2008) 84 % of candidates

- Comparison between several corpora from various domains (for Verb+Noun collocations)

	Law texts FR	Law texts RO	News FR	News RO	Med FR
collocations	7,8%	10,4%	9,1%	11,3%	8,7%
V+complement	22 %	23,9%	37,01%	29,08%	28,19%
S+P, N+VPP	32,1%	25,4%	21,42%	13,49%	24,5%
Noise	38,1 %	33,3 %	32,47%	46,13%	38,61%