A Collocation Extraction Tool for Romanian

Violeta Seretan, Eric Wehrli, Luka Nerima

UNIVERSITÉ DE GENÈVE — LATL — LABORATOIRE D'ANALYSE ET DE TECHNOLOGIE DU LANGAGE

Amalia Todirascu

UNIVERSITÉ DE STRASBOURG

PARSEME 5th general meeting, 23-24 September 2015, Iași, Romania

IC1207 COST Action PARSEME (PARSing and Multi-word Expressions): Towards linguistic precision and computational efficiency in natural language processing

WG1 Lexicon-Grammar Interface

WG2 Parsing Techniques for MWEs

WG3 Statistical, Hybrid and Multilingual Processing of MWEs

Background

ROMANIAN – LIMBA ROMÂNĂ

Limba română este o limbă romanică, din grupul italic al familiei de
Language Romanian is a language Romance, from group$_{+Def}$ Italic of family$_{+Dat}$ of

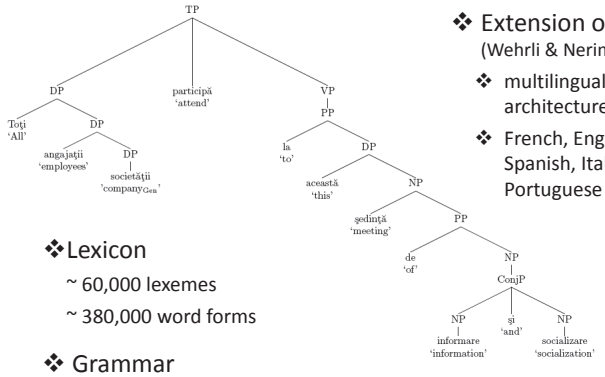
limbi indo-europene, prezentând multe similarități cu limbile
languages Indo-European, presenting many similarities with languages$_{+Def}$

franceză, italiană, spaniolă, portugheză, catalană și reto-romană.
French, Italian, Spanish, Portuguese, Catalan and Romansh.

https://ro.wikipedia.org/

- Phonemic orthography (Latin alphabet; diacritics: ă/ə, â, î/ɨ; cedilla: ş/ʃ, ţ/ʦ)
- Rich morphology and relatively free word order

FIPSROMANIAN – PARSER PROTOTYPE (Seretan et al. 2010)



- Extension of Fips (Wehrli & Nerima 2015)
  - multilingual parsing architecture
  - French, English, German, Spanish, Italian, Greek, Portuguese …
- Lexicon
  - ~ 60,000 lexemes
  - ~ 380,000 word forms
- Grammar
  - ~ 100 rules specified
  - implementation progress: ~ 50%

Try it at:

http://latlapps.unige.ch/Parser

SYNTAX-BASED COLLOCATION EXTRACTION (Seretan 2011)

Valid Syntactic Configurations

Adjective-Noun	*wide range*
Noun-Adjective	*work concerned*
Noun-Noun	*food chain*
Noun-Preposition-Noun	*fight against terrorism*
Noun-Verb	*rule apply*
Verb-Noun	*strike balance*
Verb-Preposition-Noun	*bring to justice*
Verb-Adverb	*desperately need*
Verb-Preposition	*point out*
Adjective-Adverb	*highly controversial*
Adjective-Preposition	*concerned about*
…	

Lexical Association Measures

	$Y = v$	$Y = \neg v$
$X = u$	a	b
$X = \neg u$	c	d

(u, v) – candidate

Chi-square	$\frac{(a+b+c+d)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$
Log-likelihood ratios (LLR)	$2(a\log a + b\log b + c\log c + d\log d - (a+b)\log(a+b) - (a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) + (a+b+c+d)\log(a+b+c+d))$
Mutual information (MI)	$\log_2 \frac{a(a+b+c+d)}{(a+b)(a+c)}$
Salience	$\log_2 \frac{a(a+b+c+d)}{(a+b)(a+c)} \log_2 a$
t-score	$\frac{a(a+b+c+d) - (a+b)(a+c)}{(a+b+c+d)\sqrt{a}}$
z-score	$\frac{a(a+b+c+d) - (a+b)(a+c)}{\sqrt{a+b+c+d}\sqrt{(a+b)(a+c)}}$
…	

Try it at:

http://latlapps.unige.ch:81/Colloc

Experiment & Results

DATA

Europarl 2011 version — http://www.statmt.org/europarl/v7/ro-en.tgz

sentences	54453
words	1233996
average sentence length	22.7 words

Sample sentence (**MWEs**): *În **calitate de** politicieni responsabili, trebuie să fim pregătiţi pentru a **contribui în mod decisiv** la găsirea rapidă a unei **soluţii de durată** pentru această **spirală a violenţei**.*

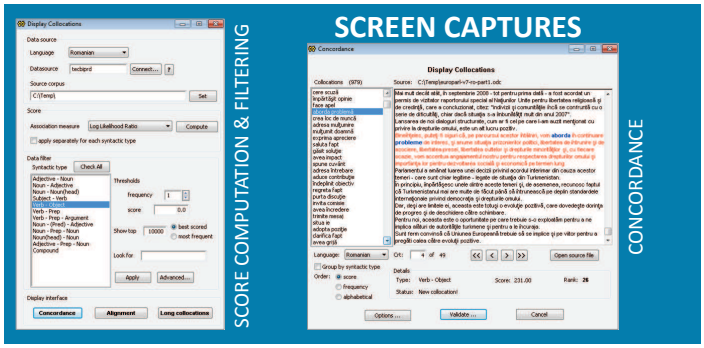
PARSING & EXTRACTION RESULTS

processing time	01:47:08
processing speed	216 tokens/s
fully parsed sentences	6169 (11.3%)
partially parsed sentences	48284
unknown words	69269

all valid combinations	
total (tokens)	194654
distinct (types)	82829
combinations (LLR > 10)	
total (tokens)	96225
distinct (types)	13844

Configuration	Tokens	Types	Example	Gloss
Noun + Adjective	48975	18264	*motiv + întemeiat*	reason valid
Noun + Prep. + Noun	39499	22045	*om + de + știință*	man of science
Verb + Object	38333	19283	*atinge + obiectiv*	touch objective
Subject + Verb	19191	10839	*timp + trece*	time pass
Compound	11529	83	*prin urmare*	through consequence
Adjective + Prep.	11438	3549	*responsabil + pentru*	responsible for
Verb + Adjective	8503	1383	*face + public*	make public
Adverb + Adjective	6862	1769	*deplin + conștient*	wholly conscious
Verb + Prep. + Noun	3653	2575	*pune + la+ îndoială*	put to doubt
Adjective + Noun	3356	1090	*cald + mulțumire*	warm thank
Verb + Prep.	1769	1089	*vota + împotriva*	vote against
Noun + Pred. + Adj.	1144	822	*lucru + este + cert*	thing is certain
Verb + Adverb	402	247	*ști + bine*	know well

SCREEN CAPTURES



EVALUATION

- Annotation
  - one judge
  - lexicographic criterion
  - test set: top 1000 results
- Precision: 67.7%

Dataset & Annotations:

http://tinyurl.com/ro-collocations

IMPORTANCE

- Extension of FipsRomanian lexical coverage
- Integration with parsing (Wehrli et al. 2010)
  - lexical disambiguation
  - collocation-driven syntactic attachment
- Exploitation for collocation dictionary building (Todirascu 2014)

References

Seretan, Violeta. *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology, Vol. 44, Springer, 2011.

Seretan, Violeta, Eric Wehrli, Luka Nerima, and Gabriela Soare. FipsRomanian: Towards a Romanian version of the Fips syntactic parser. In *Proc. of LREC'10*, Valletta, Malta, 2010.

Todirascu, Amalia. A hybrid multilingual method to extract collocations from corpora. http://typo.uni-konstanz.de/parseme/images/Meeting/2014-03-11-Athens-meeting/PosterAbstracts/todirascu-abstract.pdf, 2014 (retrieved June, 2015).

Wehrli, Eric and Luka Nerima. The Fips multilingual parser. In *Language Production, Cognition, and the Lexicon*, Gala et al. (Eds.) Text, Speech and Language Technology, Vol. 48, Springer, 2015.

Wehrli, Eric, Violeta Seretan, and Luka Nerima. Sentence analysis and collocation identification. In *Proc. of MWE 2010*, pages 27–35, Beijing, China, 2010.

PARSEME — cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY — European Science Foundation