Poster abstract at the PARSEME 5th general meeting, 23-24 September 2015, Iași, Romania
Relevant Work Groups: WG 1-3

A Collocation Extraction Tool for Romanian

Violeta Seretan, Eric Wehrli, Luka Nerima — University of Geneva, Switzerland

Amalia Todirascu — University of Strasbourg, France

Background. Lexical knowledge, and in particular knowledge on multi-word expressions, is at the cornerstone of language applications such as syntactic parsing or machine translation. Corpus-driven lexical acquisition is one of the major means to create such knowledge, in order to build or consolidate dictionaries and similar types of lexical resources. We describe ongoing work devoted to the corpus-based extraction of multi-word expressions – in particular, collocations – for the Romanian language. Romanian is since 2002 one of the 23 official languages of the European Union; it is the native language or around 24 million people, and is currently ranked 8th in the list of most spoken European languages worldwide, after Spanish (405 million native speakers), English (360), Portuguese (215), German (89), French (74), Italian (59), and Polish (40)[1]. This high rank contrasts, however, with the relatively scarce development of language resources and tools compared to other languages.

Objectives. In order to advance the state of the art in Romanian language technologies, we developed a syntactic parser prototype for Romanian (Seretan et al., 2010), in the context of a larger, long-term project devoted to developing multilingual syntactic parsers, tools for extracting lexical knowledge, and machine translation systems. We relied on syntactic parsing to extraction collocations from Romanian corpora, and added around six hundred manually validated collocations in the lexical database of the parser. We also created Romanian-French bilingual entries for these collocations in the lexical database of our rule-based translation system. This bilingual resource will be used in the Romanian-French system we want to develop next. We build on this work to further develop the Romanian parser, to extend the lexical acquisition work by carrying out more experiments on new corpora that became available relatively recently, and, thus, to pave the way to the development of machine translation systems for Romanian, so that Romanian will be in a better position in the language technologies landscape.

Our specific research objectives include: the corpus-based analysis of morphosyntactic preferences exhibited by collocations as a whole and by collocation components (as in similar work by Todirascu, 2014a); the integration of collocations in syntactic parsing, to guide attachment decisions (as described in Wehrli et al., 2010); the use of syntactic information for lexical acquisition; and the comparison of syntax-based and window-based approaches to collocation extraction (Seretan, 2008). For the purpose of the present poster presentation, we focus on: *i)* the syntax-based extraction of collocation from new corpora, and *ii)* the evaluation of the extraction results by manual validation of top-scored collocation candidates.

Method. The Romanian parser prototype (Seretan et al., 2010) has been used to analyse a subpart of the Romanian version of Europarl corpus (the 2011 version[2]). During the parsing process, syntactically-bound lexical combinations in specific grammatical configurations –

[1] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers (Accessed: June, 2015).
[2] http://www.statmt.org/europarl/v7/ro-en.tgz (Accessed: May, 2015).

adjective-noun, noun-preposition-noun, subject-verb, verb-object, verb-preposition-noun etc. – have been stored in a database, then statistically analysed. Next, lexical association measures – e.g., log-likelihood ratio (LLR) – were computed in order to score the candidates.

Our multilingual collocation extraction tool, FipsCo (Seretan, 2008) has been easily extended to Romanian. Its interface allows, in particular, for collocation validation in context. The top 1000 extraction results were manually validated, and are now being added to the lexical database of the parser. The dataset and annotations are available online.[3]

Results. In Table 1, we report experimental results pertaining to parsing, extraction (including the recognition of collocations already in lexicon) and manual validation, and we provide sample extraction results.

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 + ştiinţă,* lit. *man of science*
valid candidates in top 1000 (%)	677 (67.7%)	*trage + semnal de alarmă,* lit. *pull a signal of alarm*

Table 1. Experimental results (parsing, extraction, validation) and sample collocations extracted.

The extraction precision is 67.7%, in line with results for other languages (Seretan, 2008). Previously, we achieved 30.3% precision for a corpus of journalistic text (Seretan et al., 2010). The better performance is explained by the nature and size of the present data.

Impact. Our work proved the feasibility of carrying out large-scale lexical acquisition work for Romanian using a parser prototype, and allowed us to double the coverage of our collocation database. In the future, we plan to assess the impact of the newly-extracted collocations on parsing; compare syntax-based and window-based approaches to collocation extraction for Romanian, as we did for other languages (Seretan, 2008)[4]; and look for synergies with work on creating a Romanian collocation dictionary (Todirascu 2014b).

References

Seretan, Violeta. Collocation extraction based on syntactic parsing. PhD thesis, University of Geneva, 2008.

Seretan, Violeta, Eric Wehrli, Luka Nerima, and Gabriela Soare. FipsRomanian: Towards a Romanian version of the Fips syntactic parser. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 2010.

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

Todirascu, Amalia (2014). An LMF model for a French-Romanian collocation dictionary. Poster abstract retrieved from: http://typo.uni-konstanz.de/parseme/images/Meeting/2014-09-08-Frankfurt-meeting/WG1-TODIRASCU-abstract.pdf (Accessed: June, 2015).

Wehrli, Eric, Violeta Seretan, and Luka Nerima. Sentence analysis and collocation identification. In Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), pages 27–35, Beijing, China, 2010.

[3] http://www.issco.unige.ch/en/staff/seretan/data/annot/RO-data-PARSEME-Iasi.xlsx (Accessed July, 2015).

[4] Parsing may make a huge difference for Romanian, since this language exhibits a very rich morphology; consequently, window-based extraction is affected by the high data fragmentation (i.e., the same collocation form, or *type*, being spread among many different forms, or *tokens)*.