Comparing Recurring Lexico-Syntactic Trees (RLTs) and Ngram Techniques for Extended Phraseology Extraction: a Corpus-based Study on French Scientific Articles

Agnès Tutin and **Olivier Kraif**
Univ. Grenoble Alpes, LIDILEM
CS40700
38058 Grenoble cedex 9, France
agnes.tutin,olivier.kraif@univ-grenoble-alpes.fr

Abstract

This paper aims at assessing to what extent a syntax-based method (Recurring Lexico-syntactic Trees (RLT) extraction) allows us to extract large phraseological units such as prefabricated routines, e.g. *as previously said* or *as far as we/I know* in scientific writing. In order to evaluate this method, we compare it to the classical ngram extraction technique, on a subset of recurring segments including speech verbs in a French corpus of scientific writing. Results show that the RLT extraction technique is far more accurate for extended MWEs such as routines or collocations but performs more poorly for surface phenomena such as syntactic constructions or fully frozen expressions.

1 Introduction

Multiword expressions are diverse. They include frozen expressions such as grammatical words (e.g. *as far as*, *in order to*), non compositional idioms (e.g. *kick the bucket*), but also less frozen expressions which belong to the "extended phraseology": collocations (e.g. *pay attention*), pragmatemes (e.g. *see you later*, *how do you do*?) or clichés and routines (*as far as I know*, *as previously said* in scientific writing). Given this diversity, we think that MWE extraction techniques should be tuned according to specific kinds of MWEs. Syntax-based MWE extraction techniques produce very interesting results for collocation extraction (e.g. (Evert, 2008), (Seretan, 2011)) and are now widely used in NLP, in particular to deal with binary collocations such as *pay attention* or *widely used*. In this paper, we wish to assess to what extent a syntax-based method (Recurring Lexico-syntactic Trees (RLT) extraction) is accurate to extract larger phraseological units such as prefabricated routines. In order to evaluate this method, we compare it to the classical ngram extraction technique on a subset of recurring segments including speech verbs in a French corpus of scientific writing. We will first present the syntax-based extraction technique and will present the methodology (corpus and linguistic typology). We will then provide some first results on a quantitative and a qualitative analysis.

2 Recurring Lexico-syntactic Trees: a syntax-based extraction technique for extended MWEs

In a dependency parsed treebank, one may be interested in identifying recurring sub-trees. From a sequence of words, it is easy to extract all the subsequences of 2..n words (for a given value of n, e.g. 8), with their frequencies (what (Salem, 1987) calls "repeated segments", also called "ngrams"). Similarly, it is possible to extract from a treebank all the sub-trees containing 2..n nodes. But combinatorics is much more larger in the case of trees: theoretically, for a tree that includes t nodes, one may have up to

$$\sum_{k=2}^{n} \binom{t-1}{k}$$

subtrees with 2..n nodes (Corman, 2012). For instance, with a sentence of 20 tokens we obtain a total of 54 ngrams of length 2 to 4, and up to 704 subtrees of 2 to 4 nodes (ibid.). To solve the computational problem due to this combinatorial explosion, we simplify it by focusing on the binary co-occurrences between nodes connected by syntactic relations (in this case dependency relations). The RLT method was developed within a software architecture centered on the notion of "syntactic co-occurrence", in the words of (Evert, 2008),

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 176–180,
Valencia, Spain, April 4.

which characterizes a significant statistical association between two words syntactically related, for example (play-OBJ->role). We used a tool called Lexicoscope ((Kraif and Diwersy, 2012); (Kraif and Diwersy, 2014)), which extracts, for a given node-word, a table that records its most significant syntactic collocates (for all or only a subset of syntactic relationships). This table is called *lexicogram*, and presents significant collocates in a way analogous to the *Sketch Engine* ((Kilgarriff and Tugwell, 2001)), except that all the involved relationships are merged into a single table. Including frequency statistics and association measures, this lexicogram contains information about the syntactic relations, and about the *dispersion*, which indicates the number of sub-corpora where the co-occurrence has been identified. This latter clue is useful to highlight general phenomena, shared by all the sub-corpora, because some recurring associations may be very prominent locally, in a small part of the corpus (even in a single document), without having general scope. The architecture of Lexicoscope allows to study the collocates for simple node-words, but also for trees, comparable to what (Rainsford and Heiden, 2014) call *keynodes*. As an example, for the subtree <présenter+article>we obtain the collocates of Figure 1:

We see that these collocates, when clustered two by two, may be used to reconstruct the full tree of the routine <nous + proposer + dans + cet + article>. Starting from these binary co-occurrence scheme, including a sub-tree and a single word, we developed an iterative method to extract complete recurring trees with an arbitrary number of nodes. This method is fully automated, and operates in the following manner:

1. start from an initial keynode (single word or subtree) ;
2. extract the lexicogram ;
3. expand the keynode with any collocate that exceed a given threshold of association measure ;
4. repeat step 2 for all the newly expanded keynodes.

The process is repeated as long as there are new collocates that exceed the significance threshold, and until the extracted trees have not exceeded a certain length (in the following, the maximum length will be set to 8 elements). We call "Recurring Lexico-syntactic Trees" (RLT) the recurring trees yielded by this process. These steps are illustrated in Figure 2, for the RLT corresponding to <proposer + dans + ce + article>:

This method assumes that most interesting recurring expressions have at least two adjacent nodes that are strongly associated, which allows to start the iterative process. Once the first two nodes are merged into one tree, the association measure with other nodes is usually high, even though the pairwise association measure between words is initially low (because the frequency of the initial subtree is generally much lower than the frequency of its individual words). The analysis of the results in a corpus-based study will make it possible to determine whether this hypothesis is valid.

3 Comparison of Ngrams and RLTs of Speech Verbs in Scientific Writing

3.1 Aims of the study

This study aims at comparing through concrete examples different kinds of segments extracted by the syntax-based *RLT method* and a conventional method widely used in phraseology and stylistics, the *repeated segments method* (or *n-grams*) which identify recurrent sequences of words, lemmas or contiguous punctuation ((Salem, 1987), (Biber et al., 2004)). We focused on particular recurring segments associated with 25 speech verbs, selected among several semantic subfields[1] and used to extract segments such as *comme on l'a dit* ('as previously said') or *article propose* (lit. 'article proposes'). Among these segments, the routines associated with the rhetorical and discourse functions in scientific writing are of particular interest (see also (Teufel and Moens, 2002); (Sándor, 2007); (Tutin and Kraif, 2016)). The corpus used for this experiment includes 500 scientific articles of about 5 million words in 10 fields of human science, syntactically annotated using the XIP dependency parser ((Aït-Mokhtar et al., 2002)). We evaluated qualitatively and quantitatively the segments extracted with both methods.

[1] e.g. 'mention', 'emphasis', 'discussion', 'formulation'...

Lexicogramme | Graphiques

Show 25 entries Search:

l1	l2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.log.likelihood
proposer_VERB article_NOUN	dans_PREP	PREPOBJ#2	19	98	40434	9736620	6	112,6651	1
proposer_VERB article_NOUN	ce_DET	DETERM#2	12	98	39649	9736620	5	59,9231	2
proposer_VERB article_NOUN	nous_PRON	SUBJ#1	8	98	12180	9736620	4	51,7553	3

Figure 1: Extracting a lexicogram for a given subtree (<proposer+article>))

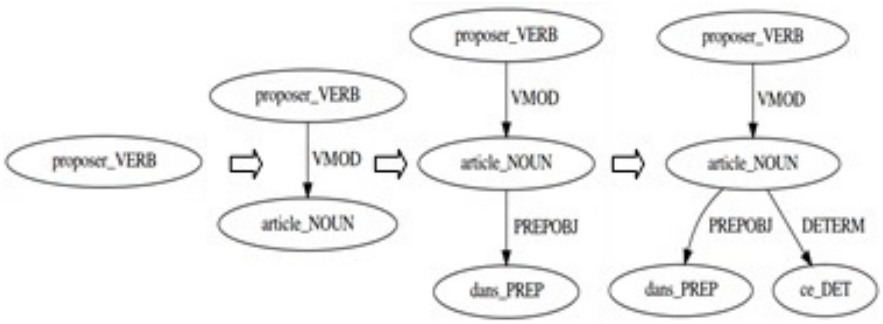

Figure 2: A three steps extraction to get the RLT <proposer + dans + ce + article>)

3.2 Extraction methods and linguistic typology of segments

Both extraction methods use the lemmatized corpus. Ngrams were extracted with the help of a homemade script, which identifies contiguous words and punctuation marks (essentially commas) occurring at least 8 times in at least 3 disciplines, and including at least 3 words. Similarly, we extracted RLTs occurring at least 8 times at each iteration (with a likelihood ratio >10.81) in at least three disciplines, including at least 3 words. The dispersion measure has proved useful for targeting cross-disciplinary expressions, and therefore the routines specific within the genre of scientific articles rather than within a specific discipline. We further characterized the extracted segments, relying on a linguistic typology in order to better understand the complementarity of both methods. A close look at the text was often necessary in order to characterize the segments more accurately.

a. Routines are sentence patterns which fulfill a rhetorical function in scientific writing, such as performing a demonstration, providing a proof, guiding the reader, etc. The following segments are routines: *comme nous le avoir souligner*(lit. 'as we have pointed it out'), *il falloir dire que* (lit. 'it must be said').

b. Collocations, unlike routines, are considered as plain binary recurring associations (cf. (Hausmann, 1989)), as in *formuler le hypothèse* (lit. formulate a hypothesis).

c. Specific syntactic constructions deal with specific alternations, e.g. passive constructions, impersonal or modal constructions, which are often characteristic of the scientific genre, e.g. *avoir être souligner* (lit. 'have been pointed out'), *permettre de préciser* (lit. 'allows to specify')

d. Frozen expressions include non compositional multiword expressions, close to idioms (see (Sag et al., 2002)), e.g. *c'est-à-dire* ('that is to say'), or *cela va sans dire* ('it goes without saying').

e. Non relevant expressions are segments which do not belong to the previous typology and are considered as irrelevant since they have no phraseological function, e.g. *avoir dire que il* (lit. 'have say that he/it'), *dire que ce łtre* (lit. 'say what this be').

4 Results

4.1 Quantitative comparison

The extractions performed with the ngram techniques produced a large set of sequences. To limit noise, we removed ngrams ending with a determiner (which proved to be redundant with segments without determiners). After filtering, there is a total of 435 ngrams to be examined. Extrcated RLTs are much less numerous (276 elements), slightly more than half of the ngrams. 124 segments are extracted by both techniques (45 % of extracted RLTs also extracted with ngram techniques). In order to assess the interest of both methods, we considered the relevance of the extracted segments according to the above linguistic typology. Figure 3 shows the results of this analysis, using raw data, while Figure 4 and Figure 5 show the relative distribution for each method.

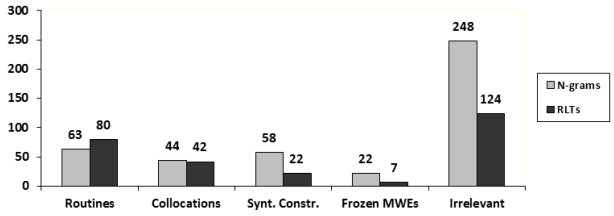

Figure 3: Comparison of results by type (raw data)

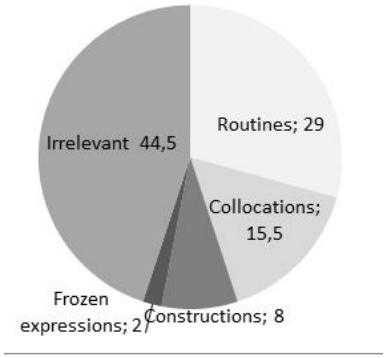

Figure 4: Distribution of results for RLTs (in %)

In general, the results broadly confirm our expectations. Regarding raw results, the RLT technique extracts less elements than the ngram technique, but a larger number of routines and a comparable number of collocations. On the other hand, for fixed expressions and constructions, which can be considered as surface phenomena among multiword expressions, the recall of the ngram technique is better. The contrast between

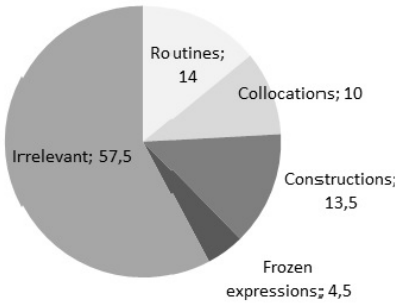

Figure 5: Distribution of results for ngrams (in %)

both approaches is even more striking when looking at the distribution of the linguistic MWE types in percentage terms (see Figures 4 and 5). The RLT technique undoubtedly produces more satisfactory results for the "extended" phraseological phenomena, such as collocations or routines, since almost half results fall into these two categories, but proves to be disappointing for fixed expressions and constructions. As regards precision rate now, the overall precision rate of the RLT technique is 55.5 %, 13 points ahead of ngram techniques, but given the complexity of RLT method, we expected a better accuracy.

4.2 Qualitative comparison

A qualitative comparison is essential to better understand the specificity of both approaches. The observation of **routines** extracted by both methods shows that expressions with contiguous elements are unsurprisingly well identified by both techniques, but frequencies are in general higher with the RLT method. Among the routines only identified by the RLT technique, we observed routines whose elements are often distant, occur in syntactic alternations or have variable determiners. Interestingly, some routines were best identified by ngram techniques than by RLT extraction techniques, e.g. routines such as 'ce + article + se + proposer + de' ('this article aims at'), due to the fact that in the dependency syntactic model used, prepositions and conjunctions are not directly related to the verb but to their arguments. This information could, however, be integrated within the RLTs with a syntactic post-treatment. Concerning **collocations**, both methods appear to be complementary. While the RLT method is more accurate with variable determiners in Verb Prep N structures (e.g. *insister sur aspect* 'insist on aspect'), it often fails to detect verb-adverb collocations due

to parsing errors (e.g. *voir plus haut/plus bas* 'see above/below'. Surface phenomena (**syntactic constructions** and **fully frozen MWEs** are better extracted by ngram techniques. Again, these poor results appear to be partly related to syntactic analysis, since some dependency relations do not relate adjacent words. For example, in an expression such as *s'exprimer par*, *par* ('lit. to be expressed with'), the preposition *par* is not attached to the verb, but to the noun which is the prepositional complement of the verb. This kind of syntactic representation is however not specific to XIP parser and is very common among dependency models.

5 Conclusion

Our comparison of RLT and ngram extraction techniques shows clearly that the first method is more suited to extract sentence patterns and routines, which have a hierarchical structure rather than a sequential nature. The RLT technique also performs well on collocation extraction, but does not produce good results on surface phenomena such as syntactic constructions or fully frozen MWEs, where grammatical words (preposition, conjunctions, adverbs) are not sufficiently taken into account. In future work, we would like to develop the multidimensional aspect of the LRT method, by using morphosyntactic categories or semantic classes rather than lexical units. The hierarchical representation makes it possible to substitute the lemmas to more general classes, more likely to explain the abstract structure of many linguistic patterns.

References

Salah Aït-Mokhtar, J-P Chanod, and Claude Roux. 2002. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(2-3):121–144.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3):371–405.

Julien Corman. 2012. *Extraction d'expressions polylexicales sur corpus arboré*. Mémoire de master recherche Industries de la langue, Univ. Stendhal Grenoble 3.

Stefan Evert. 2008. Corpora and collocations. *Corpus linguistics. An international handbook*, 2:1212–1248.

Franz Josef Hausmann. 1989. Le dictionnaire de collocations. *Wörterbücher, Dictionaries, Dictionnaires*, 1:1010–1019.

Adam Kilgarriff and David Tugwell. 2001. Word sketch: Extraction and display of significant collocations for lexicography.

Olivier Kraif and Sascha Diwersy. 2012. Le lexicoscope: un outil pour l'étude de profls combinatoires et l'extraction de constructions lexico-syntaxiques. In *Actes de la conférence TALN 2012*, pages 399–406.

Olivier Kraif and Sascha Diwersy. 2014. Exploring combinatorial profiles using lexicograms on a parsed corpus: a case study in the lexical field of emotions. *Blumenthal P., Novakova I., Siepmann D.(éd). Les émotions dans le discours. Emotions in discourse. Peter Lang*, pages 381–394.

Thomas M Rainsford and Serge Heiden. 2014. Key node in context (knic) concordances: Improving usability of an old french treebank. In *SHS Web of Conferences*, volume 8, pages 2707–2718. EDP Sciences.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer.

André Salem. 1987. Pratique des segments répétés. essai de statistique textuelle. *Lexicométrie et textes politiques*.

Ágnes Sándor. 2007. Modeling metadiscourse conveying the authors rhetorical strategy in biomedical research abstracts. *Revue française de linguistique appliquée*, 200(2):97–109.

Violeta Seretan. 2011. *Syntax-based collocation extraction*, volume 44. Springer Science & Business Media.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Agnès Tutin and Olivier Kraif. 2016. Routines sémantico-rhétoriques dans lécrit scientifique de sciences humaines: lapport des arbres lexico-syntaxiques récurrents. *Lidil. Revue de linguistique et de didactique des langues*, (53):119–141.