Combining Compound Recognition and PCFG-LA Parsing with Word Lattices and Conditional Random Fields[1]

MATTHIEU CONSTANT, Université Paris-Est, LIGM, CNRS
JOSEPH LE ROUX, Université Paris-Nord, LIPN, CNRS
ANTHONY SIGOGNE, Université Paris-Est, LIGM, CNRS

The integration of compounds in a parsing procedure has been shown to improve accuracy in an artificial context where such expressions have been perfectly pre-identified. This paper evaluates two empirical strategies to incorporate such multiword units in a real PCFG-LA parsing context: (1) the use of a grammar including compound recognition thanks to specialized annotation schemes for compounds; (2) the use of a state-of-the-art discriminative compound pre-recognizer integrating endogenous and exogenous features. We show how these two strategies can be combined with word lattices representing possible lexical analyses generated by the recognizer. The proposed systems display significant gains in terms of multiword recognition and often in terms of standard parsing accuracy. Moreover, we show through an oracle analysis that this combined strategy opens new promising research directions.

Categories and Subject Descriptors: I.2.7 [**ARTIFICIAL INTELLIGENCE**]: Natural Language Processing

General Terms: Experimentation,Languages

Additional Key Words and Phrases: Conditional Random Fields, Multiword Expressions, Parsing, Word Lattice

ACM Reference Format:
Constant, M., Le Roux, J. and Sigogne, A. 2013. Combining Compound Recognition and PCFG-LA Parsing with word lattices and Conditional Random Fields. ACM Trans. Speech Lang. Process. 9, 4, Article 39 (March 2013), 24 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Multiword Expressions (MWE) form complex lexical units that have the particularity of having a certain level of idiomaticity. Their recognition is therefore crucial for real-life applications and fundamental Natural Language Processing tools like part-of-speech taggers, syntactic and semantic parsers, etc. This paper is devoted to the integration of MWE recognition in empirical syntactic parsing, which has already been considered in several studies such as in [Nivre and Nilsson 2004] for dependency parsing and in [Arun and Keller 2005] in constituency parsing. Although these experiments always relied on a corpus where the MWEs were perfectly pre-identified, they showed that pre-grouping such expressions could significantly improve parsing accuracy. Recently, Green et al. [2011] proposed to integrate the multiword expressions directly in the grammar without pre-recognizing them. The grammar was trained with a reference treebank where MWEs were annotated with a specific non-terminal node.

In this paper, we further investigate these two empirical strategies in a realistic constituency parsing framework and show how to combine them. The first strategy is to extend the one described in [Green et al. 2011] by modifying the MWE annotation scheme in order to better guide MWE parsing. The second strategy is to chain a realistic MWE pre-pregrouping followed by parsing, and especially to develop a state-of-the-art MWE recognizer based on Conditional Random Fields (CRF) to be plugged to a parser, such as in [Constant et al. 2012]. We combine these two strategies by using word lattices: the MWE recognizer generates a word lattice representing the n-best lexical segmentations; the word lattice is then used as input of the parser. The main

[1]Submitted to the Special Issue on Multiword Expressions.

benefit of this approach is to limit the search space of the parser by taking advantage of an ambiguous MWE recognition.

In our experiments, we focus on contiguous MWEs that we thereafter call *compounds*. We use state-of-the-art parsers based on Probabilistic Context-Free Grammars with Latent Annotations (PCFG-LAs) in order to keep the best general parsing accuracy as possible. All our strategies are evaluated on French. Although the proposed techniques are very well known and widely used in the NLP community, they are rarely applied all together to MWEs and evaluated in a realistic context.

This paper is organized as follows: in section 2, we briefly define multiword expressions and describe related works on MWE recognition (MWER) and on the integration of MWEs in parsing; section 3 presents all resources that will be used: annotated corpus and lexical resources; section 4 details and evaluates the first strategy consisting of applying a PCFG-LA parser that have been trained on a treebank with a specific compound annotation scheme; in section 5, we describe and evaluate a CRF-based compound recognizer integrating endogeneous and exogeneous features; in section 6, we show how to combine the two strategies with word lattices and discuss their benefits; and, finally, we discuss the final results.

2. MULTIWORD EXPRESSIONS

2.1. Overview

Multiword expressions are lexical items made up of multiple lexemes that undergo idiosyncratic constraints and therefore offer a certain degree of idiomaticity. They cover a wide range of linguistic phenomena: fixed and semi-fixed expressions, light verb constructions, phrasal verbs, named entities, etc. They may be contiguous (e.g. *traffic light*) or discontinuous (e.g. *John took your argument into account*). They are often divided into two main classes: multiword expressions defined through linguistic idiomaticity criteria (*lexicalized phrases* in the terminology of [Sag et al. 2002]) and those defined by statistical ones (i.e. simple collocations). Most linguistic criteria used to determine whether a combination of words is a MWE are based on syntactic and semantic tests such as the ones described in [Gross 1986]. For instance, the utterance *at night* is a MWE because it does display a strict lexical restriction (**at day*, **at afternoon*) and it does not accept any inserting material (**at cold night*, **at present night*). Such linguistically defined expressions may overlap with collocations which are the combinations of two or more words that cooccur more often than by chance. Collocations are usually identified through statistical association measures. A detailed description of MWEs can be found in [Baldwin and Nam 2010].

In this paper, we focus on contiguous MWEs that form a lexical unit which can be marked by a part-of-speech tag (e.g. *at night* is an adverb, *because of* is a preposition). They can undergo limited morphological and lexical variations – e.g. *traffic (light+lights)*, *(apple+orange+...) juice* – and usually do not allow syntactic variations[2] such as inserts (e.g. **at cold night*). Such expressions can be analyzed at the lexical level. In what follows, we use the term *compounds* to denote such expressions.

2.2. Identification

The idiomaticity property of MWEs makes them both crucial for Natural Language Processing applications and difficult to predict. Their actual identification in texts is therefore fundamental. There are different ways for achieving this objective. The simpler approach is lexicon-driven and consists in looking the MWEs up in an existing lexicon, such as in [Silberztein 2000]. The main drawback is that this proce-

[2]Such MWEs may very rarely accept inserts, often limited to single modifiers: e.g. *in the short term*, *in the very short term*.

dure entirely relies on a lexicon and is unable to discover unknown MWEs. The use of collocation statistics (e.g. [Pecina 2010]) is therefore useful. For instance, for each candidate in the text, Watrin and François [2011] compute on the fly its association score from an external ngram base learned from a large raw corpus, and tag it as MWE if the association score is greater than a threshold. They reach excellent scores in the framework of a keyword extraction task. Within a validation framework (i.e. with the use of a reference corpus annotated in MWEs), Ramisch et al. [2010] developped a Support Vector Machine classifier integrating features corresponding to different collocation association measures. The results were rather low on the GENIA corpus [Kim et al. 2003] and Green et al. [2011] confirmed these bad results on the French Treebank. This can be explained by the fact that such a method does not make any distinctions between the different types of MWEs and the reference corpora are usually limited to certain types of MWEs. Furthermore, the lexicon-driven and collocation-driven approaches do not take the context into account, and therefore cannot discard some of the incorrect candidates. Vincze et al. [2011] proposed to detect noun compounds by combining a CRF model and external MWE resources automatically extracted from wikipedia. They used two kinds of training corpus: (a) a manually validated one made up of 49 English wiki pages; (b) a larger one including 5,000 randomly selected wikipages that were automatically annotated thanks to external MWE resources. They displayed scores up to 68.7% for (a) and 56% for (b) on wikipedia articles. A recent trend is to couple MWE recognition (MWER) with a linguistic analyzer: a POS tagger [Constant and Sigogne 2011] or a parser [Green et al. 2011]. Constant and Sigogne [2011] trained a unified Conditional Random Fields model integrating different standard tagging features and features based on external lexical resources. They show a general tagging accuracy of 94% on the French Treebank. In terms of Multiword expression recognition, the accuracy was not clearly evaluated, but Constant and Tellier [2012] in a cross-validation framework evaluated it between 73 and 78% depending on the kinds of compound being annotated and whether the model incorporates lexicon-based features. Green et al. [2011] proposed to include the MWER in the grammar of the parser. To do so, the MWEs in the training treebank were annotated with specific non-terminal nodes. They used a Tree Substitution Grammar instead of a Probabilistic Context-free Grammar (PCFG) with latent annotations in order to capture lexicalized rules as well as general rules. They showed that this formalism was more relevant to MWER than PCFG (71% F-score vs. 69.5%). Both methods have the advantage of being able to discover new MWEs on the basis of lexical and syntactic contexts. In this paper, we will take advantage of the methods described in this section by integrating them as features of a MWER model.

2.3. Integration of Multiword Expression Recognition in Parsing

From a theoretical point of view, the integration of multiword expressions in syntactic parsing has been studied for several formalisms: Head-Driven Phrase Structure Grammar [Copestake et al. 2002], Tree Adjoining Grammars [Schuler and Joshi 2011], *inter alia*. From an empirical point of view, their incorporation has also been considered such as in [Nivre and Nilsson 2004; Eryigit et al. 2011] for dependency parsing and in [Arun and Keller 2005; Hogan et al. 2011] in constituency parsing. Although their experiments always relied on a corpus where the MWEs were perfectly preidentified, they showed that pre-grouping such expressions could significantly improve parsing accuracy. For consituency parsing, we can note the experiments in [Cafferkey et al. 2007] that combined real-world MWE recognizers and different probabilistic parsers for English. They worked on a reference corpus where MWEs were not annotated. MWEs were automatically pre-grouped by projecting external resources and applying a Named Entity Recognizer. Then, they applied a parser and finally

reinserted the subtrees corresponding to MWEs in order to perform the evaluation. They showed small but significant gains. Recently, some studies proposed to integrate the two tasks in the same model [Finkel and Manning 2009; Green et al. 2011]. Finkel and Manning [2009] coupled parsing and Named Entity Recognition in a CRF-based parsing model. Green et al. [2011] integrated the recognition of compounds in the grammar. They specifically showed, for French, that the best parser was driven by a non-lexicalized strategy (Berkeley parser), although the compound recognition was worse than a parser driven by a lexicalized strategy.

There also exists the study in [Wehrli et al. 2010] that ranks the candidates generated by a symbolic parser on the basis of the occurrence (or not) of collocations. We can note that [Constant et al. 2012] combined a n-best PCFG-LA parser with a reranker based on a Maximum Entropy model integrating MWE-dedicated features, and showed statistically significant improvements in all evaluation metrics including MWE recognition and general parsing accuracy.

3. RESOURCES

3.1. Corpus

The French Treebank[3] (FTB) [Abeillé et al. 2003] is a syntactically annotated corpus made up of journalistic articles from `Le Monde` newspaper. We used the latest edition of the corpus (June 2010). It contains 473,904 tokens and 15,917 sentences. Phrasal elements are annotated with 13 labels. One benefit of this corpus is that its compounds are marked. Their annotation was driven by linguistic criteria such as the ones in [Gross 1986]. We exploited two different instances of this corpus: one instance (FTB-STF) resulting from the preprocessing procedure described in [Green et al. 2011] and one instance (FTB-P7) resulting from the preprocessing tools of the Alpage Team at University Paris 7. FTB-STF contains 14 part-of-speech tags and was used to get comparable results in terms of MWE recognition accuracy with the ones reported in [Green et al. 2011]. Compounds are identified with a specific non-terminal symbol "MWX" where X is the part-of-speech of the expression. They have a flat structure made of the part-of-speech of their components as shown in figure 1. In this example, the MWN node indicates that the utterance *part de marché* (market share) is a multiword noun and that it has a flat internal structure N P N (noun – preposition – noun). There exist 11 MWE labels in this instance. FTB-P7 uses a part-of-speech tagset of 28 labels optimized for parsing [Candito and Crabbé 2009] and therefore very relevant to our experiments. Each compound is grouped into a single concatenated token. In order to carry out our experiments, we had to undo all compounds and represent them like in the instance FTB-STF. We assigned their part-of-speech to all simple elements of the compounds with the tagger *lgtagger* [Constant and Sigogne 2011]. There exist 18 MWE labels in this instance.

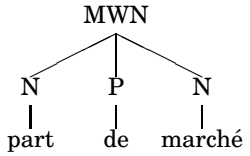


Fig. 1. Subtree of MWE *part de marché* (market share)

[3] `http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php`

The French Treebank is composed of 435,860 lexical units (34,178 types). Among them, 5.3% are compounds (20.8% for types). In addition, 12.9% of the tokens belong to a MWE, which, on average, has 2.7 tokens. Compounds are of different types: nominals such as *acquis sociaux* (social benefits), verbs such as *faire face à* (to face), adverbials like *dans l' immédiat* (right now), prepositions such as *en dehors de* (beside). Note that multiword verbs[4] are limited to verbs that tend to be contiguous, i.e. expressions that do not incorporate free arguments like in *to take something into account*. At worse, they can accept small adjuncts such as adverbials, e.g. ***fait*** *plus que jamais* ***partie*** 'be more than ever part'. Some Named Entities are also encoded: organization names like *Société suisse de microélectronique et d' horlogerie*, family names like *Strauss-Kahn*, location names like *Afrique du Sud* (South Africa) or *New York*.

The train/dev/test split is the same as in [Green et al. 2011]: 1,235 sentences for test[5], 1,235 for development and 13,347 for training. The development and test sections are the same as those generally used for experiments in French, e.g. [Candito and Crabbé 2009].

3.2. Lexical resources

French is a resource-rich language as attested by the existing morphological dictionaries which include compounds. In this paper, we use two large-coverage general-purpose dictionaries: Dela [Courtois 2009; Courtois et al. 1997] and Lefff [Sagot 2010]. The Dela was manually developed in the 90's. We used the distribution freely available in the platform Unitex[6] [Paumier 2011]. It consists of 840,813 lexical entries including 104,350 multiword ones (91,030 multiword nouns). The compounds present in the resources respect the linguistic criteria defined in [Gross 1986]. The lefff is a freely available dictionary[7] that has been automatically compiled by drawing from different sources and that has been manually validated. We used a version with 553,138 lexical entries including 26,311 multiword ones (22,673 multiword nouns). Their different modes of acquisition makes those two resources complementary. In both, lexical entries are composed of an inflected form, a lemma, a part-of-speech and morphological features. The Dela has an additional feature for most of the multiword entries: their syntactic surface form. For instance, *eau de vie* (brandy) has the feature `NDN` because it has the internal flat structure noun – preposition *de* – noun. In order to be compatible with the tagset of the FTB, we automatically converted the dictionary tags in their equivalents for each FTB instance. The dictionaries have a good coverage: we observed that 95.1% of the words[8] in the development section were present in our lexical resources. They also have a recall of 93.7% in the FTB-P7 instance and 94.4% in the FTB-STF instance.

In order to compare compounds in our lexical resources with the ones in the French Treebank, we performed a preliminary lexicon-based MWE segmentation. To do this, we applied, on the development corpus, the pexternal dictionaries and the internal lexicon (i.e. extracted from the training corpus). This lexical analysis generated a finite-state automaton representing all possible analyses including MWE ones. We then applied a shortest path algorithm to select a segmentation that favors MWE analyses. The results are provided in table I. The given scores solely evaluate MWE segmentation and not tagging. They show that the use of external resources may improve recall,

[4]The corpus contains only 14 discontinuous verbs, all occuring in the training section.

[5]Note that, in practice, we removed one sentence *(..)* consisting of punctuation marks that we considered incorrectly tokenized.

[6]`http://igm.univ-mlv.fr/~unitex`

[7]`http://atoll.inria.fr/~sagot/lefff-en.html`

[8]We did not include punctuation marks and digits.

but they lead to a decrease in precision as numerous MWEs in the dictionaries are not encoded as such in the reference corpus; some MWEs found are not actual MWEs in their occuring context. For instance, the utterance *sur ce* may be either a frozen adverbial meaning "thereupon" or a compositional sequence (meaning *on this*) part of a prepositional phrase e.g. *sur ce point* (on this point). A very frequent ambiguous sequence is *de la* which litteraly means *of the* and which is also a partitive determiner. In addition, the FTB suffers from some inconsistencies in the MWE annotations.

Table I. Simple context-free application of the lexical resources on the development corpus

	T	L	D	L+D	T+L	T+D	T+L+D
recall	75.9	31.7	59.0	70.1	77.3	83.4	84.0
precision	61.2	52.0	55.6	50.5	58.7	51.2	49.9
f-score	67.8	39.4	57.2	58.7	66.8	63.4	62.6

Notations: We note `D` the Dela lexicon, `L` the lefff lexicon and `T` the MWE lexicon of the training corpus. The scores indicate the MWE segmentation accuracy.

In terms of statistical collocations, Watrin and François [2011] described a system that lists all the potential nominal collocations of a given sentence along with their association measure. The authors provided us with a list of 17,315 candidate nominal collocations occurring in the French treebank with their log-likelihood and their internal flat structure. By applying them with the shortest path method described above, we observed a very low recall and precision, as shown in table II. When combined with all other resources, the recall is slightly lower than with the other resources alone. This can be explained by lexical segment overlapping. For instance, *régime d'assurance-chômage*[9] (unemployment insurance scheme) is a potential collocation but not an encoded MWE in the FTB, while *assurance-chômage* (unemployment insurance) is considered an MWE and is present in the lexical resources. With the shortest path segmentation, the longest segment is preferred, therefore selects *régime d'assurance-chômage* when all resources are applied, which make recall decrease as compared with when collocation resources are not applied. We also observe a drop in the precision. Collocation resources might be useful for detecting MWE counter-examples.

Table II. Simple context-free application of the collocation resources on the development corpus

	C	C+L+D	C+L+D+T	T+L+D
recall	15.8	71.4	82.5	84.0
precision	29.4	43.6	44.1	49.9
f-score	20.6	54.1	57.5	62.6

Notations: We note `C` the collocation lexicon, `D` the Dela lexicon, `L` the lefff lexicon and `T` the MWE lexicon of the training corpus. The scores indicate the MWE segmentation accuracy.

[9]Actually, *régime d'assurance-chômage* might be considered as an accurate compound depending on the compound definition used. This shows the difficulty to evaluate MWE recognition. This issue is out of the scope of this paper. But we believe that it is a fundamental one that should be deeply discussed in the MWE community.

4. PARSING COMPOUNDS

A first simple strategy to integrate compound recognition and parsing is to incorporate the multiword recognition in the grammar, as in [Green et al. 2011]: compounds are annotated with a specific non-terminal symbol. In this paper, we experiment a PCFG-LA strategy because we want to reach the best general parsing accuracy, although it has been shown in [Green et al. 2011] not to be the best strategy for compound recognition. We focus on improving MWE recognition accuracy by modifying the labelling of the MWEs in the treebank while keeping the best parsing accuracy as possible.

4.1. Products of PCFG-LA

Probabilistic Context-Free Grammars with Latent Annotations were first introduced in [Matsuzaki et al. 2005]. The key idea underlying this grammatical formalism is the notion of latent annotations that refine, or specialize, observable grammar symbols. These annotated symbols are in turn used to create specialized rules. Annotations for each symbol, and the probabilities of the corresponding annotated rules, are learned automatically on the training corpus, usually relying on an iterative procedure like the Expectation-Maximization (EM) algorithm[10].

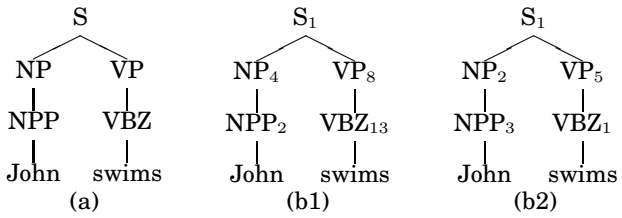


Fig. 2. An observable tree (a) and two possible annotated trees (b1,b2) for the sentence *John swims*

For this evaluation, we are only interested in the observable trees, in other words the trees with basic (not specialized) categories as appearing in the treebank (for example, tree (a) on Fig. 2). The score of such trees is the sum of the scores of their annotated counterparts (for example, trees (b1) and (b2) on Fig. 2). Formally, for a PCFG-LA G and an observed tree t, we write $D(t)$ the set of annotated trees corresponding to t and $R(d)$ the set of annotated rules r appearing in an annotated tree d. Then the probability associated with t is:

$$P_G(t) = \sum_{d \in D(t)} P_G(d) = \sum_{d \in D(t)} \prod_{r \in R(d)} P_G(r)$$

There is no efficient way to calculate such an alternation of sum and product given a sequence of words – a sentence – in order to determine what is the most probable tree for this sentence. Hence implementations perform approximations based on Bayesian variational inference, as described in [Matsuzaki et al. 2005] and [Petrov and Klein 2007], with a polynomial time complexity.

Finally, [Petrov 2010] presented a parsing algorithm that can use a sequence of PCFG-LAs, named grammar-product parsing, where the probabilities are combined at the consistuent level when scoring trees, in order to overcome the problem of inference

[10]Although recent communications like [Cohen et al. 2012] show that an analytic inference is also possible.

relying on EM: the resulting grammars are not guaranteed to reach the maximum likelihood over the training data[11]. For this algorithm, we create several PCFG-LAs from the same training corpus, only changing the EM initialization parameters – as these parameters are initialized randomly, in practice this amounts to changing the random seed used by the number generator – to obtain grammars with the same rules, but where annotations and weights associated with rules may differ. These weights are combined during parsing, at the constituent level. A symbol (in the parse chart) is scored with the product of the scores given by each grammar. The actual details of this algorithm are beyond the scope of this paper and we redirect the readers to the original paper [Petrov 2010] for a complete description.

With this setting, we expect to get a competitive baseline against which we can evaluate our method thoroughly.

4.2. Varying Compound Annotation Scheme

A simple way of improving MWE parsing with PCFG-LA is to modify the annotation scheme of the MWEs, and especially the way to annotate their components. In the annotation scheme defined in [Green et al. 2011], the POS tagset of the MWE components is the same as the one used to tag simple words as shown in Fig. 3. We consider it our *baseline* annotation scheme.

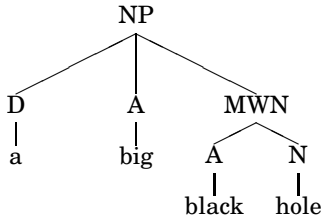


Fig. 3. Baseline annotation scheme for compounds

We propose to use specific POS tags inside compounds in order to guide their parsing. A first annotation scheme is to assign each MWE component its POS augmented with a symbol ($^+$). In the example in Fig. 4, *black* is assigned the POS A^+ because it is an adjective (A) and is part of the multiword noun *black hole*. Note that the word *big* is considered as a simple adjective and is simply assigned the POS *A*. There are two drawbacks when using this strategy: (1) the POS tagset size increases from 14 to 28 for FTB-STF and from 28 to 53 for FTB-P7. Hence learning these grammars is more difficult; (2) two consecutive multiword units cannot be delimited on the POS level. One way to resolve the delimitation issue is to augment the MWE component POS with the symbols *B* when it is at the beginning of a compound and *I* for the remaining positions in the compound. In the example in Fig. 5, the adjective *black* is assigned the part-of-speech *A+B* because it is at the beginning of the multiword *black hole*; the noun *hole* is then given the tag *N+I* because it does not start the compound. The main drawback is that the POS tagset size still rises: 39 tags for FTB-STF and 75 for FTB-P7.

[11]Another drawback of the EM algorithm is that variations of the initial conditions can have a dramatic impact over the final grammar. Hence it may be difficult to analyze the effects of changing tagsets or, more generally, any parameter of the system. Using a product of grammars helps reducing this noise factor.

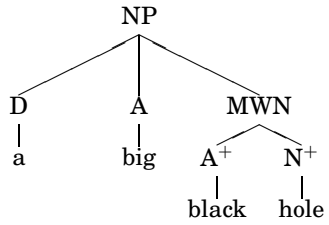


Fig. 4. *specialized* annotation scheme for compounds

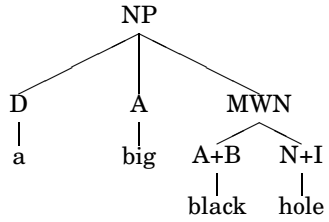


Fig. 5. *specialized* annotation scheme with BI for compounds

4.3. Experimental Setup

4.3.1. Parsers. In our experiments, we use the LORG parser [Attia et al. 2010], known to perform well on out-of-domain texts (also used in [Le Roux et al. 2012] and [Seddah et al. 2012])[12]. The input of this software can be either mere sequences of words or word lattices. This feature provides a way to see what happens when the multiword tokenization is performed by a preprocessing module only, by the parser only, or by the parser according to a set of hypotheses selected by the preprocessing module and presented as a word lattice. As we use the product-of-grammars algorithm (see section 4.1), for each experiment we trained 8 different grammars (with random seeds in $[1; 8]$, cf. section 4.1).

In order to get comparable results with [Green et al. 2011], we also used the Berkeley Parser [Petrov and Klein 2007]. In particular, we used the version adapted for French and available in the Bonsai toolkit[13] [Candito and Crabbé 2009]. This version uses the Berkeley Parser and does not implement the product-of-grammars parsing algorithm. Hence, to mitigate the impact of initial parameters over learning, for each experiment with Bonsai, we trained 3 grammars[14]. Each score provided is the average of the applications of these grammars.

4.3.2. Evaluation metrics. Results are reported using several standard measures, the F_1score and *unlabeled attachement* scores. The labeled F_1score (also noted F thereafter)[15], defined by the standard protocol called PARSEVAL [Black et al. 1991], takes into account the bracketing and labelling of the constituent nodes. Two nodes of different parses of the same sentence are considered equivalent if they have the same label and the same span. The *unlabeled attachement score* [UAS] evaluates the quality of unlabeled dependencies between the words of the sentence. This score is computed by using the CoNLL 2006 evaluation tool[16]. The evaluation requires to automatically

[12] available at `https://github.com/CNGLdlab/LORG-Release`
[13] available at `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`
[14] We used the following random seeds: 2, 5 and 8.
[15] We used *Evalb* tool available at `http://nlp.cs.nyu.edu/evalb/`. We also used the evaluation by category implemented in the class `EvalbByCat` in the Stanford Parser.
[16] Available at `http://ilk.uvt.nl/conll/software.html`.

convert constituent trees into dependency trees. We used the conversion procedure described in [Candito et al. 2010] implemented in the tool *Bonsai*. We had to integrate new rules detecting the lexical heads in the compound constituents, whenever it was possible. For instance, we considered that the lexical head of a multiword noun is the first noun from the left (when it is available): e.g. *part* is the head of the compound *part de marché*; the lexical head of a multiword adverb starting with a preposition is this preposition; the head of a multiword verb is the first verb from the left. When none of the rules apply, the lexical head of a constituent is its last element from the left. The quality of the compound identification was evaluated by computing the F_1 score on MWE nodes[17]. We also evaluated the MWE segmentation by using the unlabeled F_1 score (U).

In order to establish the statistical significance of results between two parsing experiments in terms of F_1 and UAS, we used an unidirectional t-test for two independent samples[18]. The statistical significance between two MWE identification experiments was established by using the McNemar-s test [Gillick and Cox 1989]. The results of two experiments are considered statistically significant with the computed value $p < 0.01$.

4.3.3. First results. The parsing results on the development sections of FTB-STF and FTB-P7 are provided in table III. The *gold* experiment corresponds to parsing the gold compound segmentation with a grammar learned on the treebank where the MWEs are pre-grouped in single tokens. In that case, the pre-grouped MWEs in the resulting parse trees are undone and represented in the baseline scheme, in order to get comparable results.

First, we can see that there is a 2 to 3.6 point difference in terms of parsing accuracy between the *gold*[19] and *baseline* scores. This shows that compound recognition cannot be neglected within a parsing framework. The parser LORG shows the best parsing accuracy by 2.5 points in F-score as compared with the Bonsai parser. This confirms the strong efficiency of the grammar product strategy. We also note that using a specialized POS tagset for compounds does not display statistically significant effects in terms of general parsing accuracy (F and UAS). Though, this strategy improves compound recognition in a statistically significant way: between +0.6 and +3.3 points. The use of the BI annotation scheme shows the best MWER scores, although it makes the POS tagset wider. Note that the LORG parser is so competitive that the different strategies to improve accuracy show lower effects than with the Bonsai parser. For instance, on the FTB-STF corpus, modifying the POS tagset for compounds improves MWE recognition by +2 points and +0.6 points with respect to the baseline by using the Bonsai parser and the LORG parser, respectively. On the FTB-P7, we have a +3.3 point gain for Bonsai and a +1.4-point gain for LORG.

5. A CRF-BASED COMPOUND RECOGNIZER

The parsing strategy with compound pre-grouping requires a MWE recognizer. This section is devoted to the description and the evaluation of such a tool. We use CRF models in which we integrate features that correspond to different state-of-the-art techniques. The pre-grouping strategy can also be combined with POS tagging to be used as input of a parser. By using the approach of Constant and Sigogne [2011], we adapted our MWE recognizer to a joint multiword tokenizer and POS tagger.

[17]We used the class EvalByCat available in the Stanford Parser.
[18]We used Dan Bikel's tool that is available at `http://www.cis.upenn.edu/~dbikel/software.html`.
[19]Note that MWER is not 100% accurate in the gold multiword tokenisation configuration: while MWE segmentation is 100% accurate, this is not the case for tagging.

Table III. Parsing on development corpus

		FTB-STF			FTB-P7		
Parser	Annotation	F	F(MWE)	UAS	F	F(MWE)	UAS
BON	baseline	80.03^a	72.7	85.19^b	79.72^c	68.6	84.78
	specialized	79.88^a	73.3	85.34^b	79.68^c	71.2	85.24^d
	specialized + BI	**80.05**^{a}	**74.7**	**85.48**^{b}	**79.81**^{c}	**71.9**	**85.25**^{d}
	gold	83.04	96.0	90.32	83.34	93.1	90.77
LORG	baseline	82.36^e	74.5^f	**86.82**^{g}	**82.25**^{h}	71.7	86.35^j
	specialized	82.20^e	74.3^f	86.72^g	82.07^h	73.0^i	**86.53**^{j}
	specialized + BI	**82.45**^{e}	**75.1**	86.75^g	81.93^h	**73.1**^{i}	86.50^j
	gold	84.45	95.6	90.59	85.09	92.8	91.03

Note: Scores with the same letters are not statistically significant compared with each others.

5.1. A labelling Task

MWER can be seen as a sequence labelling task (like chunking) by using an BIO-like annotation scheme [Ramshaw and Marcus 1995]. This implies a theoretical limitation: recognized compounds must be contiguous. The proposed annotation scheme is therefore theoretically weaker than the one proposed by [Green et al. 2011] that integrates the MWER in the grammar and allows for discontinuous compounds. Nevertheless, in practice, the compounds we are dealing with are very rarely discontinuous and if so, they contain small inserts. Constant and Sigogne [2011] proposed to combine MWE segmentation and part-of-speech tagging into a single sequence labelling task by assigning to each token a tag of the form TAG+X where TAG is the part-of-speech (POS) of the lexical unit the token belongs to and X is either B (i.e. the token is at the beginning of the lexical unit) or I (i.e. for the remaining positions): *John*/*N+B hates*/*V+B traffic*/*N+B jams*/*N+I*. For MWER, as our task consists in jointly locating and tagging MWEs, we limited the POS tagging to MWEs only (TAG+B/TAG+I), simple words being tagged by O (outside): *John*/*O hates*/*O traffic*/*N+B jams*/*N+I*.

For such a task, we used Linear chain Conditional Random Fields (CRF) that are discriminative probabilistic models introduced in [Lafferty et al. 2001] for sequential labelling. Given an input sequence of tokens $x = (x_1, x_2, ..., x_N)$ and an output sequence of labels $y = (y_1, y_2, ..., y_N)$, the model is defined as follows:

$$P_\lambda(y|x) = \frac{1}{Z(x)}.\sum_{t}^{N}\sum_{k}^{K} log\lambda_k.f_k(t, y_t, y_{t-1}, x)$$

where $Z(x)$ is a normalization factor depending on x. It is based on K features each of them being defined by a binary function f_k depending on the current position t in x, the current label y_t, the preceding one y_{t-1} and the whole input sequence x. The tokens x_i of x integrate the lexical value of this token but can also integrate basic properties which are computable from this value (for example: whether it begins with an upper case, it contains a number, its tags are in an external lexicon, etc.). The feature is activated if a given configuration between t, y_t, y_{t-1} and x is satisfied (i.e. $f_k(t, y_t, y_{t-1}, x) = 1$). Each feature f_k is associated with a weight λ_k. The weights are the parameters of the model to be estimated. The features are defined by users. Usually, they are generated from templates that are instantiated at each position of the input sequence. The features used in our system are described in the following subsection.

5.2. Features

5.2.1. Endogenous Features. Endogenous features are features directly extracted from properties of the words themselves or from a tool learned from the training corpus (e.g. a tagger).

Word n-grams. We use word unigrams and bigrams in order to capture multiwords present in the training section and to extract lexical cues to discover new MWEs. For instance, the bigram *coup de* is often the prefix of compounds such as *coup de pied* (kick), *coup de foudre* (love at first sight), *coup de main* (help).

POS n-grams. We use part-of-speech unigrams and bigrams in order to capture MWEs with irregular syntactic structures that might indicate the idiomacity of a word sequence. For instance, the POS sequence *preposition – adverb* associated with the compound *depuis peu* (recently) is very unusual in French. We also integrated mixed bigrams made up of a word and a part-of-speech.

Basic features. In order to deal with unknown words and special tokens, we incorporate standard tagging features in the CRF: lowercase form of the words, word prefixes of length 1 to 4, word suffixes of length 1 to 4, whether the word is capitalized, whether the token has a digit, whether it is a hyphen. We also add label bigrams.

5.2.2. Exogenous Features. Exogenous features are features that are not entirely derived from the (reference) corpus itself. They are computed from external data (in our case, our lexical resources). The lexical resources might be useful to discover new expressions: usually, expressions that have standard syntax like nominal compounds and are difficult to predict from the endogenous features. The generation of exogeneous features is based on two steps. The first step consists of a lexical analysis of each sentence using the external resources. This analysis is non-deterministic and the result is a finite-state automaton (TFSA). The second step is the extraction of relevant discriminative features from the generated TFSA.

Lexicon-based features. We associate each word with its part-of-speech tags found in our external morphological lexicon. All tags of a word constitute an ambiguity class ac. If the word belongs to a compound, the compound tag is also incorporated in the ambiguity class. For instance, the word *night* (either a simple noun or a simple adjective) in the context *at night*, is associated with the class *adj_noun_adv+I* as it is located inside a compound adverb. This feature is directly computed from the automaton. The lexical analysis can lead to a preliminary MWE segmentation by using a shortest path algorithm that gives priority to compound analyses. This segmentation is also a source of features: a word belonging to a compound segment is assigned different properties such as the segment part-of-speech mwt and its syntactic structure mws encoded in the lexical resource, its relative position $mwpos$ in the segment ('B' or 'I').

Collocation-based features. In our collocation resource, each candidate collocation of the French treebank is associated with its internal syntactic structure and its association score (log-likelihood). We divided these candidates into two classes: those whose score is greater than a threshold and the other ones. Therefore, a given word in the corpus can be associated with different properties whether it belongs to a potential collocation: the class c and the internal structure cs of the collocation it belongs to, its position $cpos$ in the collocation (B: beginning; I: remaining positions; O: outside). We manually set the threshold to 150 after some tuning on the development corpus.

5.2.3. Feature templates. All feature templates are given in table IV: n is the current position in the sentence; $w(n)$ is the token at position n; $lowercase(n)$ is the lowercase form of $w(n)$; $prefix_i(n)$ is the prefix of size i of $w(n)$; $suffix_i(n)$ is the suffix of size i of $w(n)$; $hasHyphen(n)$ indicates whether $w(n)$ contains an hyphen; $hasDigit(n)$ indicates whether $w(n)$ includes a digit; $allUppercase(n)$ indicates whether $w(n)$ is capitalized;

$t(n)$ is the predicted part-of-speech of $w(n)$; ac is the ambiguity class of $w(n)$; if $w(n)$ is part of a compound in the Shortest Path Segmentation, $mwt(n)$ and $mws(n)$ are respectively the part-of-speech and the internal structure of the compound, $mwpos(n)$ indicates its relative position in the compound (B or I). Each template is instantiated at each position of the text. Each instance of a template corresponds to a feature that is activated each time the instance is activated. For example, in the tagged sequence *the/D big/A black/N+B hole/N+I*, at position 2 (*black*), the template $w(n+0)\&y(n)$ is instantiated as $w(n+0)\&y(n) = black\&N+B$, and this instance corresponds to the binary feature function f_{2020}:

$$f_{2020}(x, y_t, t, y_{t-1}) = \begin{array}{l} 1 \text{ if } x_t = \text{"black" and } y_t = \text{"N+B"} \\ 0 \text{ otherwise} \end{array}$$

Table IV. Feature templates

Basic Feature Patterns	
$lowercase(n)$	$\&y(n)$
$prefix_i(n), i \in \{1,2,3,4\}$	$\&y(n)$
$suffix_i(n), i \in \{1,2,3,4\}$	$\&y(n)$
$hasHyphen(n)$	$\&y(n)$
$hasDigit(n)$	$\&y(n)$
$allUppercase(n)$	$\&y(n)$
$isCapitalized(n)$	$\&y(n)$
Endogenous Feature Patterns	
$w(n+i), i \in [-2,2]$	$\&y(n)$
$w(n+i)/w(n+j), (i,j) \in \{(-1,0),(-1,1),(0,1)\}$	$\&y(n)$
$t(n+i), i \in \{-2,-1,0,1,2\}$	$\&y(n)$
$t(n+i)/t(n+j), (i,j) \in \{(0,1),(1,2),(-1,1)\}$	$\&y(n)$
$t(n-i)/t(n-j), (i,j) \in \{(0,1),(1,2)\}$	$\&y(n)$
$w(n+i)/t(n+j), (i,j) \in \{(-1,0),(0,-1),(0,1),(1,0)\}$	$\&y(n)$
$y(n-1)$	$\&y(n)$
Exogenous Feature Patterns	
ac($n+i$),$i \in [-2,2]$	$\&y(n)$
mwt(n)	$\&y(n)$
mwt(n)/mwpos(n)	$\&y(n)$
mws(n)	$\&y(n)$
mws(n)/mwpos(n)	$\&y(n)$
c(n)/cs(n)/cpos(n)	$\&y(n)$
mwpos(n)	$\&y(n)$

5.3. First Results

5.3.1. Experimental setup. The MWE recognizer relies on the software *Wapiti*[20] [Lavergne et al. 2010] to train and apply the model, and on the software *Unitex* [Paumier 2011] to apply lexical resources. The part-of-speech tagger used to extract POS features was *lgtagger*[21] [Constant and Sigogne 2011]. In all experiments, we varied the set of features: *base* are the basic features; *endo* corresponds to all endogeneous features; *coll* and *lex* include all endogenous features plus respectively collocation-based features and lexicon-based ones; *all* is composed of both endogenous and exogenous features.

[20]Wapiti can be found at `http://wapiti.limsi.fr`. It was configured as follows: rprop algorithm, default L1-penalty value (0.5), default L2-penalty value (0.00001), default stopping criterion value (0.02).
[21]Available at `http://igm.univ-mlv.fr/~mconstan/research/software/`.

The quality of MWE identification was evaluated by computing the F_1 score on MWE nodes. Two nodes of different parses of the same sentence are considered equivalent if they have the same label and the same span. For each experiment, we detailed the precision (P) and the recall (R). We also evaluated the MWE segmentation by using the unlabeled F_1 score (U).

The statistical significance between two MWE identification experiments was established by using the McNemar-s test [Gillick and Cox 1989]. The results of two experiments are considered statistically significant with the computed value $p < 0.01$.

5.3.2. Compound recognition. The results of the standalone compound recognizer on the development sections of the two FTB instances are reported in the table V. They show that the system with all features reaches the best score for both FTB instances. The recognition is better by +2.6 points on the FTB-STF because of its smaller POS tagset. Accuracy is improved by an absolute gain of +4.7 points and +4.9 points as compared with the Bonsai parser on FTB-STF and FTB-P7 respectively. As compared with the LORG parser, the absolute gain is similar for FTB-STF (+4.3). The LORG parser seems to perform better on the POS tagset of FTB-P7: the absolute gain of our system is reduced to +3.7. The strictly endogenous system has +0.6/0.7 point absolute gain with respect to the LORG parser. The addition of collocation features does not have any statistically significant effects on the system. As expected, exogeneous features lead to a 3.4 point recall improvement on the FTB-STF with respect to endogeneous features (+2.8 points on the FTB-P7). The more precise system is the base one because it mainly detects compounds present in the training corpus; nevertheless, it hardly captures new MWEs (it has the lowest recall). The two parsers have the best recall among the non lexicon-based systems, i.e. it is the best one to discover new compounds as it is able to precisely detect irregular syntactic structures that are likely to be MWEs. Nevertheless, as it does not have a lexicalized strategy, it is not able to filter out incorrect candidates; the precision is therefore very low (the worst).

Table V. Compound recognition on the development corpus of FTB-STF

	FTB-STF (11 tags)				FTB-P7 (19 tags)			
	P	R	F1	U	P	R	F1	U
base	**83.5**	66.9	74.3^e	75.2	**81.6**	64.2	71.9^f	74.4
endo	81.1	71.2	75.8^a	75.9	79.5	68.7	73.7^b	76.5
coll	81.5	71.1	75.9^a	76.0	79.8	69.1	74.1^b	76.6
lex	82.2	**76.5**	79.2^c	80.2	80.4	73.0	76.5^d	79.3
all	82.7	76.3	**79.4**^{c}	**80.4**	80.9	**73.1**	**76.8**^{d}	**79.5**
Bonsai	74.6	74.7	74.7^e	76.1	71.2	72.6	71.9^f	75.5
LORG	75.6	74.5	75.1	76.2	73.4	72.9	73.1	76.0

Note: The Bonsai and LORG parsers were trained on a corpus using the specialized POS tagset and the BI annotation scheme for compounds. Scores with the same letters are not statistically significant compared with each others.

Around 25% of the compounds in the development corpus are unknown, i.e. not present in the training corpus. We observed that 19% and 28% of them are correctly segmented with the recognizer based on endogenous features (*CRF-tag*) and on all features (*CRF-all*), respectively. They are rather good at discovering multiword expressions composed of words separated by hyphens (*pare-brise* – windshield –), numerical determiners (*dix-huit mille* – eighteen thousand –), named entities including capitalized words, foreign words or words occuring solely in compounds in the training section such as *Bank* (e.g. *Bank of credit and commerce international*). Some unknown expressions whose structure is *noun + adjective* [NA] (*conseiller municipal* – city councillor

–, *parti conservateur* – conservative party –) or *noun* + *preposition* + *noun* (*femme de ménage* – cleaning maid –) are also detected. For instance, the unknown compound *conseiller municipal* is recognized because (a) *conseiller* is often part of a compound noun with the NA structure in the training section: *conseiller régional* (regional counsellor), *conseiller social* (social counsellor), *conseiller technique* (technical adviser), etc; (b) *municipal* very often co-occur with the NA compound *conseil municipal* (town council). The unknown *femme de ménage* (cleaning maid) is identified because the word bigram *femme de* (litt. woman of) is always part of the compound *femme de chambre* (chambermaid) in the training section. The *CRF-all* has better accuracy than *CRF-endo* on such regular syntax compound nouns because it also uses information from external MWE lexicons. Despite this, the rather low accuracy for unknown compounds shows that there is room for improvements.

Tables VI and VII show the results by category. First, we can see that all systems are good at detecting multiword prepositions (MWP), conjunctions (MWC), determiners (MWD) and pronouns (MWPRO) on FTB-STF, and multiword prepositions, subordinating conjunctions (MWCS) and determiners (MWDET) on FTB-P7. Nonetheless, they perform very badly to recognize multiword verbs on FTB-STF, and multiword adjectives (MWADJ) as well as coordinating conjunctions (MWCC) on FTB-P7. As compared with LORG, our CRF-based system is much better at recognizing multiword prepositions, verbs and adjectives. On FTB-STF, LORG is also worse at identifying multiword conjunctions. Moreover, we can see that integrating lexicon-based features improves the detection of multiword nouns (MWN) on FTB-STF and common nouns (MWNC) on FTB-P7. Multiword adverbs reach higher accuracy for both corpus. We can note that the stricly endogeneous CRF-based recognizer perform slightly better on proper nouns (NPP) on FTB-P7.

Table VI. Compound recognition by category on the development corpus of FTB-STF

Cat.	#gold	LORG	CRF-endo	CRF-all
MWN	974	71.6	71.4	76.5
MWADV	360	75.2	76.7	79.0
MWP	346	81.6	83.6	85.7
MWC	93	79.6	89.4	90.1
MWD	50	80.9	83.2	83.2
MWV	31	56.5	71.4	69.7
MWA	25	48.8	63.2	61.5
MWPRO	17	81.3	90.3	93.8
MWET	3	N/A	N/A	N/A
MWCL	1	66.7	100	100
		74.3	75.8	79.4

5.3.3. Joint lexical segmentation and POS Tagging. In this subsection, we focus on the evaluation of the joint task of lexical segmentation and POS tagging. To do so, we used the annotation scheme described in [Constant and Sigogne 2011] and in section 5.1. The task consists in jointly recognizing compounds and tag all lexical units (including compounds). The output will also be used as input of the parser in section 6. The results on the development corpus of FTB-P7[22] are given in table VIII. The best system includes all features without tag-based features[23]. It shows an absolute gain of 1 point for overall tagging and 4.6 points for compound recognition with respect to the base

[22]We do not display the results on FTB-STF because they show equivalent behaviours.
[23]Tag-based features are features using POS of the tokens predicted by the POS tagger *lgtagger*.

Table VII. Compound recognition by category on the development corpus of FTB-P7

Cat.	#gold	LORG	CRF-endo	CRF-all
MWNC	710	69.1	68.0	74.3
MWADV	360	76.1	77.7	80.9
MWP	346	81.6	84.0	85.1
MWNPP	263	66.8	66.0	64.6
MWCS	82	85.4	86.1	87.5
MWDET	48	83.5	83.7	83.7
MWADJ	25	46.5	61.9	57.8
MWPRO	17	82.4	90.3	93.8
MWVPP	13	66.7	54.6	58.3
MWVINF	13	77.4	96.3	96.3
MWCC	11	53.9	45.5	64.3
MWV	5	25.0	44.4	36.4
MWET	3	N/A	N/A	N/A
MWCLS	1	100.0	100.0	100.0
MWVPR	0	N/A	N/A	N/A
		73.0	73.7	76.8

system. We can note that there is no significant effect of the tag-based and collocation-based features as compared with the base system. Moreover, tag-based features tend to significantly decrease the overall tagging accuracy, when they are combined with exogeneous features.

Table VIII. Joint MWER and POS tagging on development corpus

	FTB-P7 (28 tags)	
	F(POS)	F(MWE)
base	93.64^{a}	71.9^{c}
endo	93.76^{a}	72.8^{d}
coll	93.73^{a}	72.8^{d}
coll^{-}	93.65^{a}	71.9^{c}
lex	94.37	76.9^{e}
lex^{-}	94.52^{b}	75.8
all	94.23	76.7^{e}
all^{-}	94.64^{b}	76.5^{e}

Note: The $^{-}$ symbol indicates that all tag-based features are removed. Scores with the same letters are not statistically significant compared with each others. F(POS) corresponds to the overall accuracy of the POS tagger including lexical segmentation.

6. LIMITING SEARCH SPACE

In this section, we describe a method to reduce the search space of the parser by limiting the lexical segmentations and POS tags to the ones found by the compound recognizer/POS tagger. Given the n best outputs of the recognizer, we construct a word lattice including all analyses belonging to these outputs. This word lattice is then parsed by the parser.

6.1. Parsing Word Lattices

Theoretically, parsing word lattices is grounded in the *parsing as intersection* paradigm resulting from the seminal work of [Bar-Hillel et al. 1961], extended to probabilistic context-free grammars in [Nederhof and Satta 2003], where the intersection of a (weighted) regular language, represented as a finite-state automaton, with a (weighted) context-free language, represented as a context-free parser, is shown to be a new context-free language recognizing solely the intersection of the two languages. This new grammar can also be interpreted as a compact way to represent the set of possible derivations of the regular input with the original grammar.

Practically, word lattices are used to represent ambiguous input for a parser and efficient standard parsing algorithms working on strings can be modified accordingly to work on word lattices [Chappelier et al. 1999]. This ambiguity naturally arises in pipeline NLP architectures when the input of the parser comes from for example a speech recognition system or, as it is the case here, an ambiguous tokenization.

In a word lattice, each node represents a position in the sentence and each arc represents a possible token between two positions. A complete path between the initial state and the final state is a possible tokenization for the input sentence. An example of such lattice is shown on Fig. 6, taken from the development set of the FTB, for which there are 16 distinct paths between position 0 and position 15. Each arc is labelled by a token and optionally by a POS. If no POS is provided, we consider all POS of the tagset for the token.

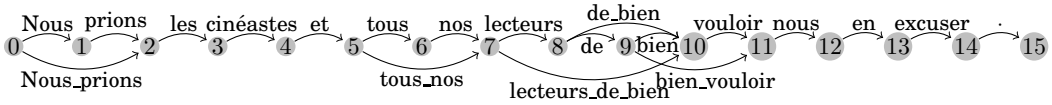


Fig. 6. Example of word lattice for the sentence *Nous prions les cinéastes et tous nos lecteurs de bien vouloir nous en excuser.*

6.2. Results

We now evaluate the use of word lattices for parsing. In our experiments, we considered two types of word lattices:

— word lattices where arcs can be compound candidates with their components merged in a single token as in Fig. 6. We used two different options: either we associate all arcs with their predicted POS (option *merged + POS*) or we solely provide POS for compound arcs (option *merged*).
— word lattices where arcs correspond to simple words (no compounds); all the words are given their part-of-speech in one of the specialized annotation scheme described in section 4: *decomposed* with no BI annotation; *decomposed + BI* with BI annotation.

The word lattices were obtained by using the n best results generated by our compound recognizer for lattices *merged* or by our joint compound recognizer and POS tagger for the other kinds of lattices. For lattices *decomposed* and *decomposed + BI*, each MWE component was tagged with *lgtagger*. For each type of word lattice, we trained a particular kind of parser. For lattices *merged + POS* and *merged*, LORG was trained on a corpus where compounds were merged into single tokens. For lattices *decomposed* and *decomposed + BI*, it was trained on a corpus where compounds and their components were respectively annotated with a specific non-terminal symbol and a specialized POS tagset. In our experiments, we varied n from 1 to 10. Figures 7 and 8 show

the results obtained by our parsers on the development section of FTB-P7[24]. We only showed results for the lattices obtained by the best compound recognizer (i.e. including all features) and by the best joint compound recognizer and POS tagger (i.e. including all features minus tag-based ones).

We first see that the accuracies of the parsers *merged* and *merged + POS* drastically drop with respect to n because the parsers tend to favour the shortest paths (i.e. paths with the longest lexical units)[25]. We can note that assigning a POS for each arc causes a smoother decrease (by 5 points from n = 1 to 10 instead of 10 points for *merged*). The parser *decomposed* reaches its top accuracy for $n = 2$ and then its accuracy slightly decrease (by 1 point from n = 2 to 10). This shows that the use of word lattices is not as interesting as expected. Especially, when compounds are pre-grouped in a single token, using the best segmentation is enough to get the best parser.

On the other hand, parsing filtered input is dramatically faster. On average we observed that using the best CRF tagger output divides parsing time by two. Even though parsing (and MWER) performance is not improved, there is still a positive aspect into running a pipeline architecture.

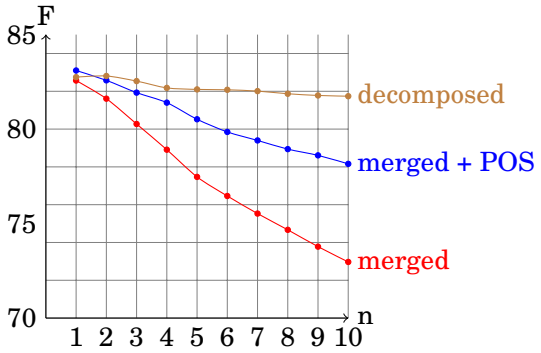


Fig. 7. Evolution of parsing accuracy

In tables IX and X, we provide the details of the best results for each parser on the development section of FTB-P7 and FTB-STF respectively. The parser *specialized* corresponds to the parser with the same name in section 4. Firstly, the systems including exogeneous features reach much better scores than systems integrating endogeneous features only (e.g. around 1 point difference in parsing accuracy and 3 points in compound recognition). Moreover, simple parsers described in section 4 get similar accuracies as the systems based on endogeneous features only. Generally, the different systems using the same kinds of features have no statistically significant differences between each others. There is an important exception on the FTB-STF where the systems parsing pre-grouped compounds have very low scores with respect to systems parsing word sequences that have been pre-tagged in a specialized annotation scheme.

[24]The scores obtained on FTB-STF are very similar in terms of score evolution with respect to n.

[25]We tried to use the Iterative Viterbi decoding method for language models [Silaghi 2005] also used for PCFG parsing [Rozenknop and Silaghi 2001] in order to penalize short paths, but the variational probability approximation used in PCFG-LA does not perform well with this technique. On the other hand, we did not try to combine CRF scores with the PCFG-LA scores. Although it could be useful in practice, multiplying scores from discriminative and generative devices is not consistent.

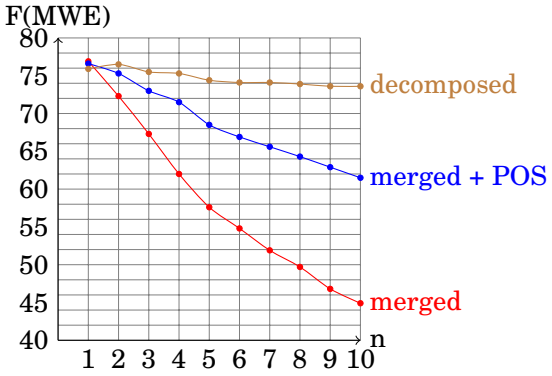


Fig. 8. Evolution of compound recognition accuracy

This confirms that the FTB-P7 tagset is optimized for parsing texts where compounds have been pre-grouped [Candito and Crabbé 2009].

Table IX. Parsing word lattices with LORG on FTB-P7 development corpus

Parser	Feats	n	F	F(MWE)	UAS
specialized	-	-	82.07^{a}	73.0^{e}	86.53^{c}
merged	endo	1	82.04^{a}	**73.7**	**86.65**^{c}
merged + POS	base	1	81.93^{a}	71.9	86.37^{c}
decomposed	base	2	82.02^{a}	72.8^{e}	86.48^{c}
decomposed + BI	base	2	**82.11**^{a}	73.1^{e}	86.50^{c}
merged	all	1	82.57	**76.9**^{g}	87.40^{d}
merged + POS	all^{-}	1	**83.10**^{b}	76.6^{g}	**87.58**^{d}
decomposed	all^{-}	2	82.91^{b}	76.0^{f}	87.47^{d}
decomposed + BI	all^{-}	2	82.82^{b}	$76.3^{f,g}$	87.46^{d}

Note: Scores with the same letters are not statistically significant compared with each others.

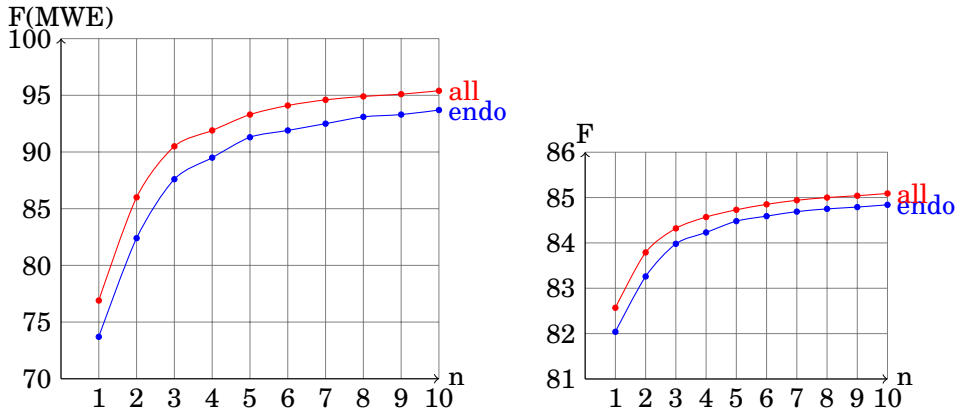
Table X. Parsing word lattices with LORG on FTB-STF development corpus

Parser	Feats	n	F	MWE	UAS
specialized + BI	-	-	82.45^{k}	75.1^{b}	86.75^{d}
merged	endo	1	81.26^{x}	**75.9**^{c}	**86.61**^{d}
merged + POS	base	1	80.95^{x}	73.7	85.97
decomposed	base	2	82.05^{a}	75.1^{b}	86.43^{d}
decomposed + BI	base	2	**82.26**^{a,k}	$75.4^{b,c}$	86.54^{d}
merged	all	1	81.83^{f}	**79.4**	87.30^{i}
merged + POS	all^{-}	1	82.08^{f}	78.5	87.15^{i}
decomposed	all^{-}	2	82.81^{g}	77.5^{h}	**87.47**^{i}
decomposed + BI	all^{-}	2	**82.92**^{g}	77.7^{h}	87.46^{i}

Note: Scores with the same letters are not statistically significant compared with each others.

6.3. Parsing Oracle Segmentation

The use of word lattices may, at first sight, look disappointing as, at most, the 2 best analyses generated by our CRF-based analyzers are sufficient to get the best scores. In this section, we want to show that this word lattice approach is promising by examining the oracle scores of our parsing systems. We define the oracle score of a sentence as the score of the parse of the best path in the input lattice (for a given n). The best path is the one that has the lexical segmentation and tagging the most similar to the golden one (i.e. the one in the reference corpus). To compute the best path, we assign a weight to each arc in the lattice. If the analysis associated with the transition corresponds to the one in the reference, then we assign it a negative weight; the default value is 0. The best path is then the shortest path according to the sum of their weights. Fig. 9 represents the oracle evolution of the CRF-based compound recognition accuracy with respect to n on the FTB-P7 development section. The *endo* curve displays the scores of the system integrating endogeneous features only and the *all* curve displays the scores of the system incorporating all features. Fig. 9 shows the parsing accuracy evolution of the parser when its input is the oracle lexical segmentation. These curves show asymptotic evolutions and almost reach their asymptote with $n = 10$. The asymptote corresponds to the score obtained with the gold parser (cf. section 4). For $n = 4$, the score differences between $n = 1$ and 10 is reduced by 80 % both in terms of general parsing accuracy and compound reconition accuracy. This observation opens very interesting perspectives. For instance, we could use a reranking strategy such as in [Constant et al. 2012] to improve performances. In addition, by putting together all oracle results, we can show that there is a linear correlation between parsing accuracy and compound recognition accuracy (cf. Fig. 10).



(a) Oracle evolution for compound recognition

(b) Parsing oracle lexical segmentation

Fig. 9. Oracle scores

7. DISCUSSIONS

7.1. Parsing Compounds or CRF-based Recognition ?

The compound recognition results on the test section of FTB-STF are provided in table XI. *Baseline 1* and *Baseline 2* corresponds to the baselines of the Bonsai and LORG

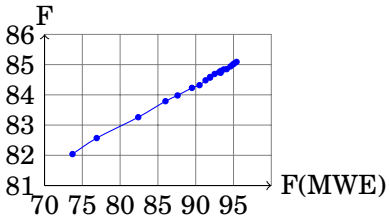


Fig. 10. Correlation between compound recognition and parsing accuracy

parsers respectively (cf. section 4). BON-BI and LORG-BI respectively correspond to Bonsai and LORG parsers trained on a corpus using the best annotation scheme for compounds (specialized + BI). We directly compare the results with the ones in [Green et al. 2011] as we used the same evaluation corpus. F1 $\leq$ 40 corresponds to the compound recognition F-score on sentences whose length is less or equal than 40 words.

The results confirm that the best recognizer is based on CRF and includes all features described in this paper. It outperforms by at least 3 points the second-best parser which is the LORG-BI parser. The LORG-BI parser achieves similar results as the CRF-based recognizer including endogeneous features (no statistically significant differences). This shows that, by choosing a relevant annotation scheme for compounds, PCFG-LA is as competitive as CRF with endogenous features only. The grammar product optimization is also a strong factor of improvement of the recognition accuracy (around 2-point difference between Bonsai and LORG) when using the best annotation scheme. We can also note that our two strategies – i.e. (1) CRF and (2) compound annotation scheme optimization – outperform the ones described and evaluated in [Green et al. 2011] by at least 1.5 points and at most 7.0 points. An interesting future work would consist in extending the work by Green et al. [2011] by using a specialized annotation scheme for compounds and adapting grammar product algorithm to Tree substitution grammars.

Table XI. Comparison on test section of FTB-STF corpus with other MWE Recognizers

	F1 $\leq$ 40	F1
CRF-all	78.1	78.0
LORG-BI	74.2	74.9^a
CRF-endo	74.1	74.5^a
BON-BI	72.6	72.8
DP-STG*	71.1	-
Baseline 1	70.2	70.7
Stanford*	70.1	-
Baseline 2	69.3	70.0

*The results of the Stanford Parser (Stanford) and the parser based on a Tree-substitution grammar (DP-STG) are directly reported from [Green et al. 2011]. Scores with the same letters are not statistically significant compared with each others.

7.2. Pre-grouping compounds ?

In order to create a realistic evaluation context, we selected the best parsing configurations from the results on the development sections. Final parsing results are provided in table XII. Firstly, the test section appears easier to parse than the development section: simple parsers with specialized POS tagsets described in section 4 achieve similar parsing accuracy as parsing systems using the best CRF-based analyzers. Morever, the effect of compound recognition improvement on general parsing does not clearly appear, contrary to the evaluations on the development sections.

Table XII. Final parsing results on test sections

FTB	Parser	Feats	n	F1	MWE	UAS
STF	baseline	-	-	83.09^{a}	72.6	$87.37^{d,e}$
	specialized + BI	-	-	83.38^{b}	74.9^{c}	$87.31^{d,e}$
	decomposed + BI	base	2	83.08^{a}	74.8^{c}	$87.17^{d,e}$
	decomposed + BI	all^{-}	2	83.39^{b}	76.4	87.69^{e}
	gold	gold	-	85.40	95.6	90.88
P7	baseline	-	-	83.18^{f}	70.1	86.89^{g}
	specialized	-	-	83.42^{f}	71.5	86.99^{g}
	merged	endo	1	83.16^{f}	73.2	86.94^{g}
	merged + POS	all	1	83.30^{f}	75.8	87.45
	gold	gold	-	86.36	93.0	91.23

Note: Scores with the same letters are not statistically significant compared with each others.

8. CONCLUSIONS AND FUTURE WORK

In this paper, we described and evaluated two different strategies to integrate compound recognition and PCFG-LA parsing. The first strategy consists in using grammars incorporating compound identification thanks to specialized annotation schemes for compounds. The second strategy consists in pre-grouping compounds with a state-of-the-art discriminative CRF-based compound pre-recognizer integrating endogenous and exogenous features. We showed how to combine these two strategies by means of word lattices. The proposed strategies displayed significant gains in terms of multiword recognition and often in terms of standard parsing accuracy. An oracle analysis showed that the combined strategy offers new promising research directions like the use of a discriminative reranker such as in [Constant et al. 2012]. The combination of the weights coming from the CRF-based compound pre-recognizer and the ones coming from the PCFG-LA-based parser is also an interesting point to discuss and experiment in the future as the direct combination via the weighting of the word lattice is not trivial because the weights are calculated from two different types of models (discriminative vs. generative). Furthermore, the scope of this paper can be extended to other types of multiword expressions.

REFERENCES

ABEILLÉ, A., CLÉMENT, L., AND TOUSSENEL, F. 2003. Building a treebank for french. In *Treebanks*, A. Abeillé, Ed. Kluwer, Dordrecht.

ARUN, A. AND KELLER, F. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of french. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

ATTIA, M., FOSTER, J., HOGAN, D., LE ROUX, J., TOUNSI, L., AND VAN GENABITH, J. 2010. Handling unknown words in statistical latent-variable parsing models for Arabic, English and French. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL'10)*. Los Angeles, CA.

BALDWIN, T. AND NAM, K. 2010. Multiword expressions. In *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group.

BAR-HILLEL, Y., PERLES, M., AND SHAMIR, E. 1961. On formal properties of simple phrase structure grammars. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 14,* 2, 143–172.

BLACK, E., ABNEY, S., FLICKINGER, D., GDANIEC, C., GRISHMAN, R., HARRISON, P., HINDLE, D., INGRIA, R., JELINEK, F., KLAVANS, J., LIBERMAN, M., MARCUS, M., ROUKOS, S., SANTORINI, B., AND STRZALKOWSKI, T. 1991. A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop*.

CAFFERKEY, C., HOGAN, D., AND VAN GENABITH, J. 2007. Multi-word units in treebank-based probabilistic parsing and generation. In *Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing (RANLP'07)*.

CANDITO, M. H. AND CRABBÉ, B. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*.

CANDITO, M.-H., CRABBÉ, B., AND DENIS, P. 2010. Statistical french dependency parsing: treebank conversion and first results. In *Proceedings of the Language and Resource Evaluation Conference (LREC'10)*.

CHAPPELIER, J., RAJMAN, M., ARAGÜÉS, R., AND ROZENKNOP, A. 1999. Lattice parsing for speech recognition. In *6ème conférence sur le Traitement Automatique du Langage Naturel (TALN'99)*. Cargse, France.

COHEN, S. B., STRATOS, K., COLLINS, M., FOSTER, D. P., AND UNGAR, L. 2012. Spectral learning of latent-variable PCFGs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*.

CONSTANT, M. AND SIGOGNE, A. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE'11)*.

CONSTANT, M., SIGOGNE, A., AND WATRIN, P. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*.

CONSTANT, M. AND TELLIER, I. 2012. Evaluating the impact of external lexical resources into a CRF-based multiword segmenter and part-of-speech tagger. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

COPESTAKE, A., LAMBEAU, F., VILLAVICENCIO, A., BOND, F., BALDWIN, T., SAG, I., AND FLICKINGER, D. 2002. Multiword expressions: Linguistic precision and reusability. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*.

COURTOIS, B. 2009. Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française 87*, 1941 – 1947.

COURTOIS, B., GARRIGUES, M., GROSS, G., GROSS, M., JUNG, R., MATHIEU-COLAS, M., MONCEAUX, A., PONCET-MONTANGE, A., SILBERZTEIN, M., AND VIVÉS, R. 1997. Dictionnaire électronique DELAC : les mots composés binaires. Tech. Rep. 56, University Paris 7, LADL.

ERYIGIT, G., ILBAY, T., AND ARKAN CAN, O. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the IWPT Workshop on Statistical Parsing of Morphologically-Rich Languages (SPRML'11)*.

FINKEL, J. R. AND MANNING, C. D. 2009. Joint parsing and named entity recognition. In *Proceedings of the conference of the North American Chapter of the Association for Computational Linguistics (NAACL'09)*.

GILLICK, L. AND COX, S. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of ICASSP'89*.

GREEN, S., DE MARNEFFE, M.-C., BAUER, J., AND MANNING, C. D. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with french. In *Proceedings of the conference on Empirical Method for Natural Language Processing (EMNLP'11)*.

GROSS, M. 1986. Lexicon grammar. the representation of compound words. In *Proceedings of Computational Linguistics (COLING'86)*.

HOGAN, D., FOSTER, J., AND VAN GENABITH, J. 2011. Decreasing lexical data sparsity in statistical syntactic parsing - experiments with named entities. In *Proceedings of ACL Workshop Multiword Expressions: from Parsing and Generation to the Real World (MWE'11)*.

KIM, J.-D., OHTA, T., TATEISI, Y., AND ICHI TSUJII, J. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. In *ISMB (Supplement of Bioinformatics)*. 180–182.

LAFFERTY, J., MCCALLUM, A., AND PEREIRA, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML'01)*. 282–289.

LAVERGNE, T., CAPPÉ, O., AND YVON, F. 2010. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*. Association for Computational Linguistics, 504–513.

LE ROUX, J., FOSTER, J., WAGNER, J., SAMAD ZADEH KALJAHI, R., AND BRYL, A. 2012. DCU-Paris13 Systems for the SANCL 2012 Shared Task. Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL).

MATSUZAKI, T., MIYAO, Y., AND TSUJII, J. 2005. Probabilistic CFG with latent annotations. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 75–82.

NEDERHOF, M. AND SATTA, G. 2003. Probabilistic parsing as intersection. In *Proceedings of the 8th International Conference on Parsing Technologies (IWPT'03)*. 137–148.

NIVRE, J. AND NILSSON, J. 2004. Multiword units in syntactic parsing. In *Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

PAUMIER, S. 2011. Unitex 2.1 - user manual. http://igm.univ-mlv.fr/~unitex.

PECINA, P. 2010. Lexical association measures and collocation extractione. *Language Resources and Evaluation 44*, 137–158.

PETROV, S. 2010. Products of random latent variable grammars. In *Proceedings of the conference on Human Language Technologies and the conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'10)*. Association for Computational Linguistics, Los Angeles, California, 19–27.

PETROV, S. AND KLEIN, D. 2007. Improved inference for unlexicalized parsing. In *Proceedings of the conference on Human Language Technologies and the conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'07)*.

RAMISCH, C., VILLAVICENCIO, A., AND BOITET, C. 2010. mwe-toolkit: a framework for multiword expression identification. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

RAMSHAW, L. A. AND MARCUS, M. P. 1995. Text chunking using transformation-based learning. In *Proceedings of the 3rd Workshop on Very Large Corpora*. 88 – 94.

ROZENKNOP, A. AND SILAGHI, M. 2001. Algorithme de décodage de treillis selon le critère du coût moyen pour la reconnaissance de la parole. In *Actes de la 8ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'01)*. Number 1. 391–396.

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A. A., AND FLICKINGER, D. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'02)*. Springer-Verlag, London, UK, 1–15.

SAGOT, B. 2010. The lefff, a freely available, accurate and large-coverage lexicon for french. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

SCHULER, W. AND JOSHI, A. 2011. Tree-rewriting models of multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE'11)*.

SEDDAH, D., SAGOT, B., AND CANDITO, M. 2012. Robust Pre-Processing and Semi-Supervised Lexical Bridging for User-Generated Content Parsing. Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL).

SILAGHI, M. 2005. Spotting subsequences matching an HMM using the average observation probability criteria with application to keyword spotting. In *Proceedings Of The National Conference On Artificial Intelligence*. Vol. 20. 1118.

SILBERZTEIN, M. 2000. Intex: an fst toolbox. *Theoretical Computer Science 231,* 1, 33–46.

VINCZE, V., NAGY, I., AND BEREND, G. 2011. Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the conference on Recent Advances in Natural Language Processing (RANLP'11)*. 289–295.

WATRIN, P. AND FRANÇOIS, T. 2011. N-gram frequency database reference to handle mwe extraction in nlp applications. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE'11)*.

WEHRLI, E., SERETAN, V., AND NERIMA, L. 2010. Sentence analysis and collocation identification. In *Proceedings of the Workshop on Multiword Expression: From Theory to Applications (MWE'10)*.