Discriminative Strategies to Integrate Multiword Expression Recognition and Parsing

Matthieu Constant
Université Paris-Est
LIGM, CNRS
France
mconstan@univ-mlv.fr

Anthony Sigogne
Université Paris-Est
LIGM, CNRS
France
sigogne@univ-mlv.fr

Patrick Watrin
Université de Louvain
CENTAL
Belgium
patrick.watrin
@uclouvain.be

Abstract

The integration of multiword expressions 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 integrate multiword units in a real constituency parsing context and shows that the results are not as promising as has sometimes been suggested. Firstly, we show that pre-grouping multiword expressions before parsing with a state-of-the-art recognizer improves multiword recognition accuracy and unlabeled attachment score. However, it has no statistically significant impact in terms of F-score as incorrect multiword expression recognition has important side effects on parsing. Secondly, integrating multiword expressions in the parser grammar followed by a reranker specific to such expressions slightly improves all evaluation metrics.

1 Introduction

The integration of Multiword Expressions (MWE) in real-life applications is crucial because such expressions have the particularity of having a certain level of idiomaticity. They form complex lexical units which, if they are considered, should significantly help parsing.

From a theoretical point of view, the integration of multiword expressions in the parsing procedure has been studied for different formalisms: Head-Driven Phrase Structure Grammar (Copestake *et al.*, 2002), Tree Adjoining Grammars (Schuler and Joshi, 2011), etc. From an empirical point of view, their incorporation has also been considered such as in (Nivre and Nilsson, 2004) for dependency parsing and in (Arun and Keller, 2005) in constituency parsing. Although 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 integrating 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.

Our proposal is to evaluate two discriminative strategies in a real constituency parsing context: (a) pre-grouping MWE before parsing; this would be done with a state-of-the-art recognizer based on Conditional Random Fields; (b) parsing with a grammar including MWE identification and then reranking the output parses thanks to a Maximum Entropy model integrating MWE-dedicated features. (a) is the direct realistic implementation of the standard approach that was shown to reach the best results (Arun and Keller, 2005). We will evaluate if real MWE recognition (MWER) still positively impacts parsing, i.e., whether incorrect MWER does not negatively impact the overall parsing system. (b) is a more innovative approach to MWER (despite not being new in parsing): we select the final MWE segmentation after parsing in order to explore as many parses as possible (as opposed to method (a)). The experiments were carried out on the French Treebank (Abeillé *et al.*, 2003) where MWEs are annotated.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 204–212,
Jeju, Republic of Korea, 8-14 July 2012.

The paper is organized as follows: section 2 is an overview of the multiword expressions and their identification in texts; section 3 presents the two different strategies and their associated models; section 4 describes the resources used for our experiments (the corpus and the lexical resources); section 5 details the features that are incorporated in the models; section 6 reports on the results obtained.

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[1] 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 procedure entirely relies on a lexicon and is unable to discover unknown MWEs. The use of collocation statistics 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 learnt 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 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. A recent trend is to couple MWE recognition 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

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

clearly evaluated, but seemed to reach around 70-80% F-score. 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.

3 Two strategies, two discriminative models

3.1 Pre-grouping Multiword Expressions

MWER can be seen as a sequence labelling task (like chunking) by using an IOB-like annotation scheme (Ramshaw and Marcus, 1995). This implies a theoretical limitation: recognized MWEs 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 MWEs. Nevertheless, in practice, the compounds we are dealing with are very rarely discontinuous and if so, they solely contain a single word insert that can be easily integrated in the MWE sequence. 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*. In this paper, 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 Ramdom Fields (CRF) that are discriminative probabilistic models introduced by 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 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 used for MWER will be described in section 5.

3.2 Reranking

Discriminative reranking consists in reranking the *n-best* parses of a baseline parser with a discriminative model, hence integrating features associated with each node of the candidate parses. Charniak and Johnson (2005) introduced different features that showed significant improvement in general parsing accuracy (e.g. around +1 point in English). Formally, given a sentence s, the reranker selects the best candidate parse p among a set of candidates $P(s)$ with respect to a scoring function V_θ:

$$p^* = argmax_{p \in P(s)} V_\theta(p)$$

The set of candidates $P(s)$ corresponds to the *n-best* parses generated by the baseline parser. The scoring function V_θ is the scalar product of a parameter vector θ and a feature vector f:

$$V_\theta(p) = \theta.f(p) = \sum_{j=1}^{m} \theta_j.f_j(p)$$

where $f_j(p)$ corresponds to the number of occurrences of the feature f_j in the parse p. According to

Charniak and Johnson (2005), the first feature f_1 is the probability of p provided by the baseline parser. The vector θ is estimated during the training stage from a reference treebank and the baseline parser ouputs.

In this paper, we slightly deviate from the original reranker usage, by focusing on improving MWER in the context of parsing. Given the *n-best* parses, we want to select the one with the best MWE segmentation by keeping the overall parsing accuracy as high as possible. We therefore used MWE-dedicated features that we describe in section 5. The training stage was performed by using a Maximum entropy algorithm as in (Charniak and Johnson, 2005).

4 Resources

4.1 Corpus

The French Treebank[2] [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) that we preprocessed with the Stanford Parser preprocessing tools (Green *et al.*, 2011). It contains 473,904 tokens and 15,917 sentences. 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). 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.

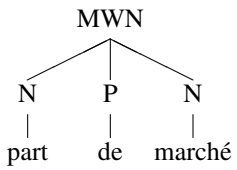

Figure 1: Subtree of MWE *part de marché* (market share): The MWN node indicates that it is a multiword noun; it has a flat internal structure N P N (noun – preposition – noun)

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. The non-terminal tagset is composed of 14 part-of-speech labels and 24 phrasal ones (including 11 MWE labels). The train/dev/test split is the same as in (Green *et al.*, 2011): 1,235 sentences for test, 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).

4.2 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, 1990; Courtois *et al.*, 1997) and Lefff (Sagot, 2010). The Dela was manually developed in the 90's by a team of linguists. We used the distribution freely available in the platform Unitex[3] (Paumier, 2011). It is composed 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[4] 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 a 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 compare compounds in these lexical resources with the ones in the French Treebank, we applied on the development corpus the dictionaries and the lexicon extracted from the training corpus. By a simple look-up, we obtained a preliminary lexicon-based MWE segmentation. The results are provided in table 1. They show that the use of external resources may improve recall, but they lead

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

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

[4] http://atoll.inria.fr/~sagot/lefff.html

to a decrease in precision as numerous MWEs in the dictionaries are not encoded as such in the reference corpus; in addition, the FTB suffers from some inconsistency in the MWE annotations.

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

Table 1: Simple context-free application of the lexical resources on the development corpus: `T` is the MWE lexicon of the training corpus, `L` is the lefff, `D` is the Dela. The given scores solely evaluate MWE segmentation and not tagging.

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.

5 MWE-dedicated Features

The two discriminative models described in section 3 require MWE-dedicated features. In order to make these models comparable, we use two comparable sets of feature templates: one adapted to sequence labelling (CRF-based MWER) and the other one adapted to reranking (MaxEnt-based reranker). The MWER templates are instantiated at each position of the input sequence. The reranker templates are instantiated only for the nodes of the candidate parse tree, which are leaves dominated by a MWE node (i.e. the node has a MWE ancestor). We define a template T as follows:

- MWER: for each position n in the input sequence x,
$$T = f(x, n)/y_n$$
- RERANKER: for each leaf (in position n) dominated by a MWE node m in the current parse tree p,
$$T = f(p, n)/label(m)/pos(p, n)$$

where f is a function to be defined; y_n is the output label at position n; $label(m)$ is the label of node m and $pos(p, n)$ indicates the position of the word corresponding to n in the MWE sequence: B (starting position), I (remaining positions).

5.1 Endogenous Features

Endogenous features are features directly extracted from properties of the words themselves or from a tool learnt 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.

Specific features. Due to their different use, each model integrates some specific features. In order to deal with unknown words and special tokens, we incorporate standard tagging features in the CRF: lowercase forms of the words, word prefixes of length 1 to 4, word suffice of length 1 to 4, whether the word is capitalized, whether the token has a digit, whether it is an hyphen. We also add label bigrams. The reranker models integrate features associated with each MWE node, the value of which is the compound itself.

5.2 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 resources are applied to the corpus through a lexical analysis that generates, for each sentence, a finite-state automaton TFSA which represents all the possible analyses. The features are computed from the automaton 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 TFSA. 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.

All feature templates are given in table 2.

Endogenous Features
$w(n+i), i \in \{-2,-1,0,1,2\}$
$w(n+i)/w(n+i+1), i \in \{-2,-1,0,1\}$
$t(n+i), i \in \{-2,-1,0,1,2\}$
$t(n+i)/t(n+i+1), i \in \{-2,-1,0,1\}$
$w(n+i)/t(n+j), (i,j) \in \{(1,0),(0,1),(-1,0),(0,-1)\}$
Exogenous Features
ac(n)
mwt(n)/mwpos(n)
mws(n)/mwpos(n)
c(n)/cs(n)/cpos(n)

Table 2: Feature templates (f) used both in the MWER and the reranker models: n is the current position in the sentence, $w(i)$ is the word at position i; $t(i)$ is the part-of-speech tag of $w(i)$; if the word at absolute position i is part of a compound in the Shortest Path Segmentation, $mwt(i)$ and $mws(i)$ are respectively the part-of-speech tag and the internal structure of the compound, $mwpos(i)$ indicates its relative position in the compound (B or I).

6 Evaluation

6.1 Experiment Setup

We carried out 3 different experiments. We first tested a standalone MWE recognizer based on CRF. We then combined MWE pregrouping based on this recognizer and the Berkeley parser[5] (Petrov *et al.*, 2006) trained on the FTB where the compounds were concatenated (BKYc). Finally, we combined the Berkeley parser trained on the FTB where the compounds are annotated with specific non-terminals (BKY), and the reranker. In all experiments, we varied the set of features: *endo* are all endogenous features; *coll* and *lex* include all endogenous features plus collocation-based features and lexicon-based ones, respectively; *all* is composed of both endogenous and exogenous features. The CRF recognizer relies on the software *Wapiti*[6] (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*[7] (Constant and Sigogne, 2011). To train the reranker, we used a MaxEnt algorithm[8] as in (Charniak and Johnson, 2005).

Results are reported using several standard measures, the $F_1 score$, *unlabeled attachment* and *Leaf Ancestor* scores. The labeled $F_1 score$ [F1][9], defined by the standard protocol called PARSEVAL (Black *et al.*, 1991), takes into account the bracketing and labeling of nodes. The *unlabeled attachement score* [UAS] evaluates the quality of unlabeled

[5]We used the version adapted to French in the software Bonsai (Candito and Crabbé, 2009): http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html. The original version is available at: http://code.google.com/p/berkeleyparser/. We trained the parser as follows: right binarization, no parent annotation, six split-merge cycles and default random seed initialisation (8).

[6]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%).

[7]Available at http://igm.univ-mlv.fr/~mconstan/research/software/.

[8]We used the following mathematical libraries PETSc et TAO, freely available at http://www.mcs.anl.gov/petsc/ and http://www.mcs.anl.gov/research/projects/tao/

[9]*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.

dependencies between words of the sentence[10]. And finally, the *Leaf-Ancestor* score [LA][11] (Sampson, 2003) computes the similarity between all paths (sequence of nodes) from each terminal node to the root node of the tree. The global score of a generated parse is equal to the average score of all terminal nodes. Punctuation tokens are ignored in all metrics. The quality of MWE identification was evaluated by computing the F_1 score on MWE nodes. We also evaluated the MWE segmentation by using the unlabeled F_1 score (U). In order to compare both approaches, parse trees generated by BKYc were automatically transformed in trees with the same MWE annotation scheme as the trees generated by BKY.

In order to establish the statistical significance of results between two parsing experiments in terms of F_1 and UAS, we used a unidirectional t-test for two independent samples[12]. The statistical significance between two MWE identification experiments was established by using the McNemar-s test (Gillick and Cox, 1989). The results of the two experiments are considered statistically significant with the computed value $p < 0.01$.

6.2 Standalone Multiword recognition

The results of the standalone MWE recognizer are given in table 3. They show that the lexicon-based system (*lex*) reaches the best score. Accuracy is improved by an absolute gain of +6.7 points as compared with BKY parser. The strictly endogenous system has a +4.9 point absolute gain, +5.4 points when collocations are added. That shows that most of the work is done by fully automatically acquired features (as opposed to features coming from a manually constructed lexicon). As expected, lexicon-based features lead to a 5.3 point recall improvement (with respect to non-lexicon based features) whereas precision is stable. The more precise system is the base one because it almost solely detects compounds present in the training corpus; nevertheless, it is unable to capture new MWEs (it has the lowest recall). BKY parser has 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).

	P	R	F_1	$F_1 \leq 40$	U
base	**78.0**	68.3	72.8	71.2	74.3
endo	75.5	74.5	75.0	74.0	76.3
coll	76.6	74.4	75.5	74.9	77.0
lex	76.0	79.8	**77.8**	**77.8**	**79.0**
all	76.2	79.2	77.7	77.3	78.8
BKY	67.6	75.1	71.1	70.7	72.5
Stanford*	-	-	-	70.1	-
DP-TSG*	-	-	-	71.1	-

Table 3: MWE identification with CRF: `base` are the features corresponding to token properties and word n-grams. The differences between all systems are statistically significant with respect to McNemar's test (Gillick and Cox, 1989), except `lex/all` and `all/coll`; `lex/coll` is "border-line". The results of the systems based on the Stanford Parser and the Tree Substitution Parser (DP-TSG) are reported from (Green *et al.*, 2011).

6.3 Combination of Multiword Expression Recognition and Parsing

We tested and compared the two proposed discriminative strategies by varying the sets of MWE-dedicated features. The results are reported in table 4. Table 5 compares the parsing systems, by showing the score differences between each of the tested system and the BKY parser.

Strat.	Feat.	Parser	F_1	LA	UAS	F_1(MWE)
-	-	BKY	80.61	92.91	82.99	71.1
pre	-	BKYc	75.47	91.10	76.74	0.0
pre	endo	BKYc	80.23	92.69	83.62	74.9
pre	coll	BKYc	80.32	92.73	83.77	75.5
pre	lex	BKYc	80.66	92.81	**84.16**	**77.4**
pre	all	BKYc	80.51	92.77	84.05	77.2
post	endo	BKY	80.87	92.94	83.49	72.9
post	coll	BKY	80.71	92.85	83.16	71.2
post	lex	BKY	**81.08**	**92.98**	83.98	74.5
post	all	BKY	81.03	92.96	83.97	74.3
pre	gold	BKYc	83.73	93.77	90.08	95.8

Table 4: Parsing evaluation: *pre* indicates a MWE pregrouping strategy, whereas *post* is a reranking strategy with $n = 50$. The feature *gold* means that we have applied the parser on a gold MWE segmentation.

[10]This score is computed by using the tool available at http://ilk.uvt.nl/conll/software.html. The constituent trees are automatically converted into dependency trees with the tool *Bonsai*.

[11]*Leaf-ancestor assessment* tool available at http://www.grsampson.net/Resources.html

[12]Dan Bikel's tool available at http://www.cis.upenn.edu/~dbikel/software.html.

	ΔF_1		ΔUAS		ΔF_1(MWE)	
	pre	post	pre	post	pre	post
endo	-0.38	+0.26	+0.63	+0.50	+3.8	+1.8
coll	-0.29	+0.10	+0.78	+0.17	+4.4	+0.1
lex	+0.05	+0.47	+1.17	+0.99	+6.3	+3.4

Table 5: Comparison of the strategies with respect to BKY parser.

Firstly, we note that the accuracy of the best realistic parsers is much lower than that of a parser with a golden MWE segmentation[13] (-2.65 and -5.92 respectively in terms of F-score and UAS), which shows the importance of not neglecting MWE recognition in the framework of parsing. Furthermore, pre-grouping has no statistically significant impact on the F-score[14], whereas reranking leads to a statistically significant improvement (except for collocations). Both strategies also lead to a statistically significant UAS increase. Whereas both strategies improve the MWE recognition, pre-grouping is much more accurate (+2-4%); this might be due to the fact that an unlexicalized parser is limited in terms of compound identification, even within n-best analyses (cf. Oracle in table 6). The benefits of lexicon-based features are confirmed in this experiment, whereas the use of collocations in the reranking strategy seems to be rejected.

	endo	coll	lex	all	oracle
n=1			80.61 (71.1)		
n=5	80.74 (71.5)	**80.88** (**71.7**)	81.03 (73.4)	**81.05** (73.3)	83.17 (74.6)
n=20	**80.98** (**72.9**)	80.72 (70.6)	81.09 (73.6)	81.01 (73.0)	84.76 (75.5)
n=50	80.87 (**72.9**)	80.71 (71.2)	81.08 (**74.5**)	81.03 (**74.3**)	85.21 (**76.4**)
n=100	80.69 (72.0)	80.53 (70.0)	**81.12** (74.4)	80.93 (73.7)	**85.54** (**76.4**)

Table 6: Reranker F_1 evaluation with respect to n and the types of features. The F_1(MWE) is given in parenthesis.

Table 7 shows the results by category. It indicates that both discriminative strategies are of interest in locating multiword adjectives, determiners and prepositions; the pre-grouping method appears to be particularly relevant for multiword nouns and adverbs. However, it performs very poorly in multiword verb recognition. In terms of standard parsing accuracy, the pre-grouping approach has a very heterogeneous impact: Adverbial and Adjective Modifier phrases tend to be more accurate; verbal kernels and higher level constituents such as relative and subordinate clauses see their accuracy level drop, which shows that pre-recognition of MWE can have a negative impact on general parsing accuracy as MWE errors propagate to higher level constituents.

cat	#gold	BKY	endo (pre)	lex (pre)	endo (post)	lex (post)
MWET	4	0.0	N/A	N/A	N/A	N/A
MWA	22	37.2	+15.2	+21.3	+0.9	+4.7
MWV	47	62.1	-9.7	-13.2	+1.7	+2.5
MWD	24	62.1	+7.3	+10.2	0.0	+1.2
MWN	860	68.2	+4.0	+7.0	+1.7	+4.2
MWADV	357	72.1	+3.8	+6.4	+3.4	+4.1
MWPRO	31	84.2	-3.5	-0.9	0.0	0.0
MWP	294	79.1	+4.3	+5.8	+0.4	+1.1
MWC	86	85.7	+0.9	+3.7	+0.2	+1.0
Sint	209	47.2	-7.7	-8.7	+0.1	-0.2
AdP	86	48.8	+1.2	+3.0	+3.4	+5.1
Ssub	406	60.8	-1.1	-1.1	-0.3	-0.5
VPpart	541	63.2	-2.8	-2.1	-0.5	-1.6
Srel	408	74.8	-3.4	-3.5	-0.3	-0.6
VPinf	781	75.2	0.0	-0.1	-0.3	-0.3
COORD	904	75.2	+0.2	+0.4	-0.3	-0.4
PP	4906	76.7	-0.8	-0.3	+0.5	+0.7
AP	1482	74.5	+3.2	+3.9	+0.7	+1.6
NP	9023	79.8	-1.1	-0.8	+0.1	+0.2
VN	3089	94.0	-2.0	-1.0	0.0	0.0

Table 7: Evaluation by category with respect to BKY parser. The BKY column indicates the F_1 of BKY parser.

7 Conclusions and Future Work

In this paper, we evaluated two discriminative strategies to integrate Multiword Expression Recognition in probabilistic parsing: (a) pre-grouping MWEs with a state-of-the-art recognizer and (b) MWE identification with a reranker after parsing. We showed that MWE pre-grouping significantly improves compound recognition and unlabeled dependency annotation, which implies that this strategy could be useful for dependency parsing. The reranking procedure evenly improves all evaluation scores. Future work could consist in combining both strategies: pre-grouping could suggest a set of potential MWE segmentations in order to make it more flexible for a parser; final decisions would then be made by the reranker.

[13] The F_1(MWE) is not 100% with a golden segmentation because of tagging errors by the parser.

[14] Note that we observe an increase of +0.5 in F_1 on the development corpus with lexicon-based features.

Acknowlegments

The authors are very grateful to Spence Green for his useful help on the treebank, and to Jennifer Thewissen for her careful proof-reading.

References

A. Abeillé and L. Clément and F. Toussenel. 2003. Building a treebank for French. *Treebanks*. In A. Abeillé (Ed.). Kluwer. Dordrecht.

A. Arun and F. Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *ACL*.

E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop*.

T. Baldwin and K.S. Nam. 2010. Multiword Expressions. *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group.

M. -H. Candito and B. Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. *Proceedings of IWPT 2009*.

E. Charniak and M. Johnson. 2005. Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

M. Constant and A. Sigogne. 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)*.

A. Copestake, F. Lambeau, A. Villavicencio, F. Bond, T. Baldwin, I. Sag, D. Flickinger. 2002. Multiword Expressions: Linguistic Precision and Reusability. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*.

B. Courtois. 1990. Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*. Vol. 87.

B. Courtois, M. Garrigues, G. Gross, M. Gross, R. Jung, M. Mathieu-Colas, A. Monceaux, A. Poncet-Montange, M. Silberztein and R. Vivés. 1997. *Dictionnaire électronique DELAC : les mots composés binaires*. Technical Report. n. 56. LADL, University Paris 7.

L. Gillick and S. Cox. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of ICASSP'89*.

S. Green, M.-C. de Marneffe, J. Bauer and C. D. Manning. 2011. Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. In *Empirical Method for Natural Language Processing (EMNLP'11)*.

M. Gross. 1986. Lexicon Grammar. The Representation of Compound Words. In *Proceedings of Computational Linguistics (COLING'86)*.

J. Lafferty and A. McCallum and F. Pereira. 2001. Conditional random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*.

T. Lavergne, O. Cappé and F. Yvon. 2010. Practical Very Large Scale CRFs. In *ACL*.

J. Nivre and J. Nilsson. 2004. Multiword units in syntactic parsing. In *Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

S. Paumier. 2011. Unitex 3.9 documentation. *http://igm.univ-mlv.fr/˜unitex*.

S. Petrov, L. Barrett, R. Thibaux and D. Klein. 2006. Learning accurate, compact and interpretable tree annotation. In *ACL*.

C. Ramisch, A. Villavicencio and C. Boitet. 2010. mwetoolkit: a framework for multiword expression identification. In *LREC*.

L. A. Ramshaw and M. P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the 3rd Workshop on Very Large Corpora*.

I. A. Sag, T. Baldwin, F. Bond, A. Copestake and D. Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *CICLING 2002*. Springer.

B. Sagot. 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)*.

G. Sampson and A. Babarczy. 2003. A test of the leaf-ancestor metric for parsing accuracy. *Natural Language Engineering*. Vol. 9 (4).

Seddah D., Candito M.-H. and Crabb B. 2009. Cross-parser evaluation and tagset variation: a French treebank study. *Proceedings of International Workshop on Parsing Technologies (IWPT'09)*.

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

M. Silberztein. 2000. INTEX: an FST toolbox. *Theoretical Computer Science*, vol. 231(1).

P. Watrin and T. François. 2011. N-gram frequency database reference to handle MWE extraction in NLP applications. In *Proceedings of the 2011 Workshop on MultiWord Expressions*.