Joint Dependency Parsing and Multiword Expression Tokenisation

Alexis Nasr, Carlos Ramisch, José Deulofeu, André Valli
Aix Marseille Université, CNRS, LIF UMR 7279
Marseille, France
FirstName.LastName@lif.univ-mrs.fr

Abstract

Complex conjunctions and determiners are often considered as pretokenized units in parsing. This is not always realistic, since they can be ambiguous. We propose a model for joint dependency parsing and multiword expressions identification, in which complex function words are represented as individual tokens linked with morphological dependencies. Our graph-based parser includes standard second-order features and verbal subcategorization features derived from a syntactic lexicon.We train it on a modified version of the French Treebank enriched with morphological dependencies. It recognizes 81.79% of ADV+*que* conjunctions with 91.57% precision, and 82.74% of *de*+DET determiners with 86.70% precision.

1 Introduction

Standard NLP tool suites for text analysis are often made of several processes that are organized as a pipeline, in which the input of a process is the output of the preceding one. Among these processes, one commonly finds a tokenizer, which segments a sentence into words, a part-of-speech (POS) tagger, which associates to every word a part-of-speech tag, and a syntactic parser, which builds a parse tree for the sentence[1]. These three processes correspond to three formal operations on the string: *segmentation* into linguistically relevant units (words), *tagging* the words with POS tags and *linking* the (word, POS) pairs by means of syntactic dependencies.

This setup is clearly not ideal, as some decisions are made too early in the pipeline (Branco and Silva, 2003). More specifically, some tokenization and tagging choices are difficult to make without taking syntax into account. To avoid the pitfall of premature decisions, probabilistic tokenizers and taggers can produce several solutions in the form of lattices (Green and Manning, 2010; Goldberg and Elhadad, 2011). Such approaches usually lead to severe computational overhead due to the huge search space in which the parser looks for the optimal parse tree. Besides, the parser might be biased towards short solutions, as it compares scores of trees associated to sequences of different lengths (De La Clergerie, 2013).

This problem is particularly hard when parsing *multiword expressions* (MWEs), that is, groups of tokens that must be treated as single units (Baldwin and Kim, 2010). The solution we present in this paper is different from the usual pipeline. We propose to jointly parse and tokenize MWEs, transforming segmentation decisions into linking decisions. Our experiments concentrate on two difficult tokenization cases. Hence, it is the parser that will choose, in such cases, whether to group or not several tokens.

Our first target phenomenon is the family of ADV+*que* constructions, a type of complex conjunction in French. They are formed by adverbs like *bien* (*well*) or *ainsi* (*likewise*) followed by the subordinative conjunction *que* (*that*). They function like English complex conjunctions *so that* and *now that*. Due to their structure, ADV+*que* constructions are generally ambiguous, like in the following examples:

1. *Je mange* **bien que** *je n'aie pas faim*
   I eat **although** *I am not hungry*
2. *Je pense* **bien que** *je n'ai pas faim*
   I think **indeed that** *I am not hungry*

In example 1, the sequence *bien que* forms a complex conjunction (*although*) whereas in example 2, the adverb *bien* (*indeed*) modifies the verb *pense* (*think*), and the conjunction *que* (*that*) introduces the sentential complement *je n'ai pas faim*

[1] This paper considers dependency syntactic structures.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1116–1126,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

(*I am not hungry*). In treebanks, the different readings are represented through the use of words-with-spaces in the case of complex conjunctions.

Our second target phenomenon is the family of partitive articles which are made of the preposition *de* (*of*) followed by the definite determiner *le*, *la*, *l'* or *les*[2] (*the*). These *de*+DET constructions are ambiguous, as shown in the following examples:

3. *Il boit* **de la** *bière*
He drinks **some** *beer*

4. *Il parle* **de la** *bière*
I talks **about the** *beer*

In example 3, the sequence *de la* forms a determiner (*some*) whereas in example 4, *de* is a preposition (*about*) and *la* is the determiner (*the*) of the noun *bière* (*bière*).

We focus on these constructions for two reasons. First, because they are extremely frequent. For instance, in the frWaC corpus, from a total of 54.8M sentences, 1.15M sentences (2.1%) contain one or more occurrences of our target ADV+*que* constructions and 26.7M sentences (48.6%) contain a *de*+DET construction (see Tables 1 and 2). Moreover, in a corpus of 370 M words in French,[3] *des* is the 7^{th} most frequent word. Second, because they are perfect examples of phenomena which are difficult to process by a tokenizer. In order to decide, in example 1, that *bien que* is a complex subordinate conjunction, non-trivial morphological, lexical and syntactic clues must be taken into account, such as the subcategorization frame of the verb of the principal clause and the mood of the subordinate clause. All these clues are difficult to take into account during tokenization, where the syntactic structure of the sentence is not yet explicit.

Ask the parser to perform tokenization will not always solve the problem. Even state-of-the-art parsers can fail to predict the right structure for the cases we are dealing with. The main reason is that they are trained on treebanks of limited size, and some lexico-syntactic phenomena cannot be well modeled. This brings us to the second topic of this paper, which is the integration of external linguistic resources in a treebank-trained probabilistic parser. We show that, in order to correctly solve the two problems at hand, the parser must have access to lexico-syntactic information that can be found in a syntactic lexicon. We propose a simple way to introduce such information in the parser by defining new linguistic features that blend smoothly with treebank features used by the parser when looking for the optimal parse tree.

The paper is organized as follows: Section 2 describes related work on MWE parsing. Section 3 proposes a way to represent multiword units by means of syntactic dependencies. In Section 4, we briefly describe the parser that has been used in this work, and in Section 5, we propose a way to integrate a syntactic lexicon into the parser. Section 6 describes the data sets used for the experiments, which results are presented and discussed in Section 7. Section 8 concludes the paper.

2 Related Work

The famous "pain-in-the-neck" article by Sag et al. (2002) discusses MWEs in parsers, contrasting two representation alternatives in the LinGO ERG HPSG grammar of English: compositional rules and words-with-spaces. The addition of compositional rules for flexible MWEs has been tested in a small-scale experiment which showed significant coverage improvements in HPSG parsing by the addition of 21 new MWEs to the grammar (Villavicencio et al., 2007).

It has been demonstrated that pre-grouping MWEs as words-with-spaces can improve the performance of shallow parsing for English (Korkontzelos and Manandhar, 2010). Nivre and Nilsson (2004) obtained similar results for dependency parsing of Swedish. They compare models trained on two representations: one where MWEs are linked by a special ID dependency, and another one based on gold pre-tokenization. Their results show that the former model can recognize MWEs with F1=71.1%, while the latter can significantly improve parsing accuracy and robustness in general. However, the authors admit that "it remains to be seen how much of theoretically possible improvement can be realized when using automatic methods for MWU recognition".

Several methods of increasing complexity have been proposed for fully automatic MWE tokenization: simple lexicon projection onto a corpus (Kulkarni and Finlayson, 2011), synchronous lexicon lookup and parsing (Wehrli et al., 2010; Seretan, 2011), token-based classifiers trained using

[2] Sequences *de le* and *de les* do not appear as such in French. They have undergone a morpho-phonetic process known as *amalgamation* and are represented as tokens *du* and *des*. In our pipeline, they are artificially *detokenized*.

[3] Newspaper *Le Monde* from 1986 to 2002.

association measures and other contextual features (Vincze et al., 2013a), or contextual sequence models like conditional random fields (Constant and Sigogne, 2011; Constant et al., 2013b; Vincze et al., 2013b) and structured perceptron (Schneider et al., 2014). In theory, compound function words like ADV+*que* and *de*+DET allow no internal variability, thus they should be represented as words-with-spaces. However, to date no satisfactory solution has been proposed for automatically tokenizing *ambiguous* MWEs.

Green et al. (2013) propose a constituency parsing model which, as a by-product, performs MWE identification. They propose a flat representation for contiguous expressions in which all elements are attached to a special node, and then they compare several parsing models, including an original factored-lexicon PCFG and a tree substitution grammar. These generic parsing models can be used for parsing in general, but they have interesting memorization properties which favor MWE identification. Their experiments on French and Arabic show that the proposed models beat the baseline in MWE identification while producing acceptable general parsing results.

Candito and Constant (2014) and Vincze et al. (2013c) present experiments on dependency parsing for MWE identification which are the closest to our settings. Vincze et al. (2013c) focus on light verb constructions in Hungarian. They propose distinguishing regular verbal dependencies from light verbs and their complements through four special labels prefixed by LCV-. Then, they train the Bohnet parser (Bohnet, 2010) using standard parameters and features, and evaluate on a gold test set. They report no significant changes in attachment scores, whereas F1 for light verb identification is 75.63%, significantly higher than the baseline methods of lexicon projection (21.25%) and classification (74.45%).

Candito and Constant (2014) compare several architectures for dependency parsing and MWE identification in French. For regular MWEs like noun compounds, they use regular expressions to automatically generate an internal syntactic structure, combining standard and MWE-dedicated dependency labels. Irregular expressions like complex conjunctions are represented as separate tokens, with a special DEP_CPD dependency that links all tokens to the first MWE word (Constant et al., 2013a). They compare different architectures for MWE identification before, during and after parsing, showing that the best architecture depends on whether the target MWEs are regular or irregular.

Similarly to these two papers, we use a special dependency to model MWEs and evaluate parsing and identification accuracy. Our work departs from theirs on three important aspects. First, we concentrate on syntactically irregular compounds, that we represent with a new kind of dependency. Second, we integrate into the parser a syntactic lexicon in order to help disambiguate ADV+*que* and *de*+DET constructions. Third, we built a specific evaluation corpus to get a better estimation of the performances of our model on ADV+*que* and *de*+DET constructions.

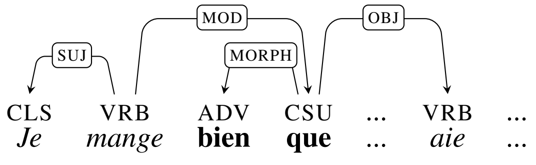
3 The MORPH Dependency

In order to let the parser take the tokenization decisions, we propose *not* to group sequences of tokens of the form ADV+*que* and *de*+DET at tokenization time. Instead, we transform the task of segmentation decision into a parsing decision task.

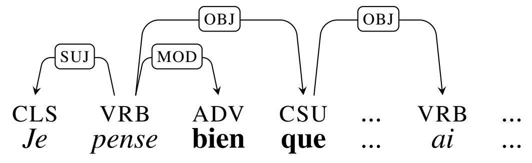
We associate a syntactic structure to ADV+*que* and *de*+DET constructions by introducing a new type of dependency that we call MORPH. It is not a standard syntactic dependency, but a reminiscent of the morphological dependencies of Mel'čuk (1988), similar to the DEP_CPD label proposed by Candito and Constant (2014) or the ID dependency of Nivre and Nilsson (2004), except that we focus on syntactically-motivated MWEs, proposing a regular structure for them.

The syntactic structures of examples 1 and 2, introduced in Section 1, are represented below[4].

Example 1.


MOD
OBJ
SUJ
MORPH
CLS VRB ADV CSU ... VRB ...
Je *mange* **bien** **que** ... *aie* ...

Example 2.

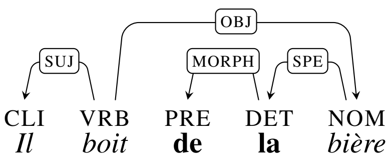

OBJ
OBJ
SUJ
MOD
CLS VRB ADV CSU ... VRB ...
Je *pense* **bien** **que** ... *ai* ...

[4] In the examples, parts of speech CLS, VRB, ADV and CSU respectively stand for subject clitic pronoun, verb, adverb and subordinating conjunction. Syntactic labels SUJ, MOD, OBJ, DE-OBJ and SPE stand for subject, modifier, object, indirect object introduced by the preposition *de* and specifier.

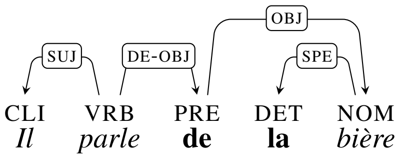
In example 1, the complex conjunction *bien que* is represented by the presence of the MORPH dependency, whereas, in example 2, the adverb *bien* modifies the verb *pense* and *que* introduces its object. From an NLP perspective, the two readings are treated the same way by the tokenizer and the tagger. It is only at parsing time that the presence of the complex conjunction is predicted.

The syntactic structures of examples 3 and 4 are represented below. In example 3, the partitive article *de la* is represented by means of the MORPH dependency. Example 4 exhibits a standard prepositional phrase structure.

Example 3.



Example 4.



4 Parsing

The parser used in this study is a second-order graph-based parser (Kübler et al., 2009). Given a sentence $W = w_1 \dots w_l$, the parser looks for the dependency tree $\hat{T}$ of W that maximizes the score s:

$$\hat{T} = \arg\max_{T \in \mathcal{T}(W)} \sum_{F \in \mathcal{F}(T)} s(F)$$

where $\mathcal{T}(W)$ is the set of all possible dependency trees for sentence W and $\mathcal{F}(T)$ is the set of all relevant subparts, called *factors*, of tree T and $s(F)$ is the score of factor F. The values of these scores are parameters estimated during training.

We can define different models of increasing complexity depending on the decomposition of the tree into factors. The most simple one is the *arc-factored* or *first-order model*, which simply decomposes a tree into single dependencies and assigns them a score, independently of their context. We used a second-order parser which decomposes a tree into factors of three types:

1. *first-order factors*, made of one dependency;
2. *sibling factors*, made of two dependencies sharing a common governor;
3. *grandchildren factors*, made of two dependencies where the dependent of one of them is the governor of the other one.

5 Integration with a Syntactic Lexicon

Although this kind of parsers achieve state-of-the-art performances (Bohnet, 2010), their predictions are limited to the phenomena that occur in the treebanks they are trained on. In particular, they often fail at correctly distinguishing elements that are subcategorized by a verb (henceforth *complements*) from others (*modifiers*). This is due to the fact that the nature and number of the complements is specific to each verb. If the verb did not occur, or did not occur often enough, in the treebank, the nature and number of its complements will not be correctly modeled by the parser.

A precise description of verb complements plays an important role in the task of predicting the MORPH dependency, as we illustrate in example 1. In this example, the verb *manger* (*eat*) does not accept an object introduced by the subordinate conjunction *que* (*that*) . This is a vital information in order to predict the correct syntactic structure of the sentence. If the parser cannot link the conjunction *que* to the verb *manger* with an OBJ dependency, then it has to link it with a MOD dependency (it has no other reasonable solution). But *que* by itself cannot be a MOD of the verb unless it is a complex conjunction. The parser has therefore no other choice than linking *que* with the adverb using a MORPH dependency.

In order to help the parser build the right solution in such cases, we have introduced information derived from a syntactic lexicon in the parser. The syntactic lexicon associates, each verb lemma, the features +/-QUE and +/-DE, that indicate respectively if the verb accepts an object introduced by the subordinating conjunction *que* and by the preposition *de*. The verbs of our examples would have the following values:

manger	-QUE	-DE
penser	+QUE	-DE
boire	-QUE	-DE
parler	-QUE	+DE

We will call such features *subcat features* (SFs). The semantics of positive feature values are quite different from the semantics of negative ones. The former indicates that a verb *may* (but does not need to) license a complement introduced by the conjunction *que* or the preposition *de*, whereas the latter indicates that the verb *cannot* license such a complement. Negative feature values have, therefore, a higher predictive power.

Every verbal lemma occurrence in the treebank is enriched with subcat features and three new factor templates have been defined in the parser in order to model the co-occurrence of subcat features and some syntactic configurations. These templates are represented in Figure 1. The first one is a first-order template and the others are grandchildren templates. In the template description, G, D and GD stand respectively for governor, dependent and grand-dependent. SF, POS, FCT and LEM respectively stand for subcat feature, part of speech, syntactic function and lemma.

1	G.SF	G.POS	D.FCT	D.POS	
2	G.SF	G.POS	D.FCT	D.POS	GD.POS
3	G.SF	G.POS	D.FCT	D.LEM	GD.POS

Figure 1: Factor templates modeling the co-occurrence of subcat features and syntactic configurations.

Two factors, of the types 1 and 3, have been represented in Figure 2. The first one models the co-occurrence of subcat feature -QUE and an object introduced by a subordinating conjunction. Such feature will receive a negative score at the end of training, since a verb having the -QUE feature should not license a direct object introduced by a subordinating conjunction. The second feature models the co-occurrence of the feature -QUE and a modifier introduced by the subordinating conjunction QUE and having an adverb as a dependent. Such a feature will receive a positive score.

1	-QUE	VRB	OBJ	CSU	
3	-QUE	VRB	MOD	QUE	ADV

Figure 2: Two factors modeling the co-occurrence of subcat features and syntactic configurations.

6 Experimental Setup

We test the proposed model to verify the linguistic plausibility and computational feasibility of using MORPH links to represent syntactically idiosyncratic MWEs in a dependency parser enriched with subcat features. Therefore, we train a probabilistic dependency parsing model on modified treebank, representing ADV+*que* and *de*+DET constructions using this special syntactic relation instead of pretokenization. Furthermore, in addition to regular features learned from the treebank, we also introduce and evaluate subcat features based on a lexicon of verbal valency, which helps identifying subordinative clauses and *de* prepositional phrases (see Section 5). We evaluate parsing precision and MWE identification on a test treebank and, more importantly, on a dataset built specifically to study the representation of our target constructions. All experiments used the NLP tool suite MACAON[5], which comprises a second-order graph-based parser.

6.1 Data Sets and Resources

French Treebank (FTB) The parser was trained on the French Treebank, a syntactically annotated corpus of news articles from *Le Monde* (Abeillé et al., 2003). We used the version which was transformed into dependency trees by Candito et al. (2009), and which was also used by Candito and Constant (2014) for experiments on MWE parsing. We used a standard split of 9,881 sentences (278K words) for training and 1,235 sentences for test (36K words). We applied simple rules to transform the flat representation of ADV+*que* and *de*+DET constructions into MORPH-linked individual tokens. All other MWEs are kept unchanged in training and test data. They are represented as single tokens, not decomposed into individual words.

MORPH **Dataset** The test portion of the FTB contains relatively few instances of our target constructions (see Tables 4 and 6). Thus, we have created two specific data sets to evaluate the prediction of MORPH links. As for ADV+*que* constructions, we manually selected the 7 most potentially ambiguous combinations from the top-20 most frequent combinations in the French Web as Corpus – frWaC (Baroni and Bernardini, 2006).[6] As for *de*+DET constructions, we selected all 4 possible combinations. For each target ADV+*que* and *de*+DET construction, we randomly selected 1,000 sentences from the frWaC based on two criteria: (1) sentences should contain only one occurrence of the target construction and (2) sentences should have between 10 and 20 words, to avoid distracting the annotators while still providing enough context. Additionally, for *de*+DET we selected only sentences in which a verb preceded the construction, in order to minimize the occur-

[5] http://macaon.lif.univ-mrs.fr

[6] http://wacky.sslmit.unibo.it/

ADV+*que*	#sent	conj.	other	#occur
ainsi	103	76.7	23.3	498,377
alors	110	88.2	11.8	291,235
autant	107	86.0	14.0	39,401
bien	99	37.4	62.6	156,798
encore	93	21.5	78.5	18,394
maintenant	120	55.8	44.2	16,567
tant	98	20.4	79.6	168,485
Total	730	56.4	43.6	1,189,257

Table 1: Annotations for ADV+*que* combinations in MORPH dataset: number of annotated sentences, proportion (%) of complex conjunction uses (MORPH) and other uses, number of occurrences in frWaC.

de+DET	#sent	det.	other	#occur
le (du)	136	33.1	66.9	16,609,049
la	138	21.0	79.0	10,849,384
les (des)	129	77.5	22.5	23,395,857
l'	136	16.9	83.1	8,204,687
Total	539	36.5	63.5	59,058,977

Table 2: Annotations for *de*+DET combinations MORPH dataset: number of annotated sentences, proportion (%) of complex determiner uses (MORPH) and other uses, number of occurrences in frWaC.

rence of nominal complements (*président de la république - president of the republic*) and focus on the determiner/preposition ambiguity. Two expert French native speakers annotated around 100 sentences per construction. Malformed or ambiguous sentences were discarded. Disagreements were either discussed and resolved or the sentence was discarded.[7]

We can see in Table 1 that ADV+*que* constructions are highly ambiguous, with 56.4% of the cases being complex conjunctions. However, they also present high variability: even though they share identical syntactic behavior, some of them tend to form complex conjunctions very often (*alors*) while others occur more often in other syntactic configurations (*tant* and *encore*). As one can see in Table 2, *de*+DET sequences tend to function as prepositions followed by a determiner with the notable exception of *de les*. The reason is that *de les* (actually the amalgame *des*) is actually the plural of the indefinite article (*un*), used with any plural noun, while the other determiners are partitives that tend to be used only with massive nouns. The last column of these tables shows the number of occurrences of each construction in the frWaC corpus. We can see that they are very recurrent combinations, specially *de*+DET constructions, which account for 3.7% of the total number of bigrams in the corpus. This underlines the importance of correctly predicting their syntactic structure in a parser.

DicoValence Lexicon DicoValence (van den Eynde and Mertens, 2003) is a lexical resource which lists the subcategorization frames of more than $3,700$ French verbs.[8] It describes more specifically the number and nature of the verbs' complements. Dicovalence gives a more fine-grained description of the complements than what is needed in our feature templates. We have only kept, as described in Section 5, the subcat features -QUE, +QUE, -DE and +DE of each verb. Table 3 below shows the number of verbal entries having each of our four subcat features. Although the number of verbs described in DicoValence is moderate, its coverage is high on our data sets. It is equal to 97.82% on the FTB test set and is equal to 95.48% on the MORPH dataset.

-QUE	+QUE	-DE	+DE
3,814	356	3,450	720

Table 3: Number of verbs in DicoValence per value of subcat feature.

6.2 Evaluation

We evaluate our models on two aspects: parsing quality and MWE identification (Nivre and Nilsson, 2004; Vincze et al., 2013c; Candito and Constant, 2014). First, we use standard *parsing attachment scores* to verify whether our models impact parsing performance in general. We compare the generated dependency trees with the reference in the test portion of the FTB, reporting the proportion of matched links, both in terms of structure – unlabeled attachment score (UAS) – and of labeled links – labeled attachment score (LAS).

Since our focus is on MWE parsing, we are also

[7] The dataset is available at `http://pageperso.lif.univ-mrs.fr/%7Ecarlos.ramisch/?page=downloads/morph`

[8] `http://bach.arts.kuleuven.be/dicovalence/`

interested in *MWE identification metrics*. We focus on words whose dependency label is MORPH and calculate the proportion of correctly predicted MORPH links among those in the parser output (precision), among those in the reference (recall) and the F1 average. Since some of the phenomena are quite rare in the FTB test portion, we focus on the MORPH dataset, which contains around 100 instances of each target construction.

We compare our approach with two simple baselines. The first one consists in pretokenizing ADV+*que* systematically as a single token, while *de*+DET is systematically left as two separate tokens. This baseline emulates the behavior of most parsing pipelines, which deal with functional complex words during tokenization. This corresponds to choosing the majority classes in the last row of Tables 1 and 2. For ADV+*que*, the precision of the baseline is 56.4%. If we assume recall is 100%, this yields an F1 score of 72.2%. For *de*+DET, however, recall is 0% since no MORPH link is predicted at all. Therefore, we only look at the baseline's precision of 63.5%. A second, slightly more sophisticated baseline, consists in choosing the majority class for each individual construction and average precisions over the constructions. In this case, the average precision is 75.3% for ADV+*que* and 76.6% for *de*+DET.

We compare our model to the one proposed by Green et al. (2013). We used the pretrained model available as part of the Stanford parser[9]. Their model outputs constituent trees, which were automatically converted to unlabeled dependency structures. We ignore the nature of the dependency link, only checking whether the target construction elements are linked in the correct order.

Our experiments use the MACAON tool suite. For the FTB, gold POS and gold lemmas are given as input to the parser. In the case of the MORPH dataset, for which we do not have gold POS and lemmas, they are predicted by MACAON. The first best prediction is given as input to the parser.

7 Evaluation Results

7.1 ADV+*que* Constructions

Table 4 reports the performances of the parser[10] on the test set of FTB. The rows of the table respectively display the results obtained without and with the use of subcat features (SF). The second and third columns represent standard attachment metrics, column four displays the number of ADV+*que* conjunctions present in the FTB test set FTB and the two last columns show the precision and recall of the MORPH dependency prediction. The table shows that the number of occurrences of ADV+*que* conjunctions is very small (27). It is therefore difficult draw clear conclusions concerning the task of predicting the MORPH dependency. The precision and recall have nevertheless been reported. The recall is perfect (all MORPH dependencies have been predicted) and the the precision is reasonable (the parser overpredicts a little). The table also shows that the use of subcat features is not beneficial, as attachment scores as well as precision decrease. The decrease of precision is misleading, though, due to the small number of occurrences it has been computed on.

SF	LAS	UAS	MORPH	Prec.	Rec.
no	88.98	90.63	27	87.10	100
yes	88.96	90.56	27	81.81	100

Table 4: Attachment scores, count, precision and recall of the MORPH dependency for ADV+*que* in FTB test, without and with subcat features (SF).

Table 5 displays the precision, recall and F1 of the prediction of the MORPH dependency on the 730 ADV+*que* sentences of the MORPH dataset, without and with the use of subcat features. The scores obtained are lower than the same experiments on the FTB.Precision is higher than recall, which indicates that the parser has a tendency to underpredict. We also present the precision of the two baselines described in Section 6.2. Only in two cases the per-construction majority baseline (indiv.) outperforms our parser without subcat features. These two constructions do not tend to form complex conjunctions, that is, the parser overgenerates MORPH dependencies. Here, subcat features help increasing precision, systematically outperforming the baselines.

The introduction of subcat features has a beneficial but limited impact on the results, increasing precision and lowering a bit recall, augmenting the tendency of the parser to under predict MORPH dependencies. Overall, our models are more precise than the Stanford parser at predicting MORPH links, specially for *bien que* and *en-*

[9] http://nlp.stanford.edu/software/lex-parser.shtml

[10] Trained on the modified train set of the FTB, where complex conjunctions and partitive determiners have been represented by means of the MORPH dependency

	Baseline prec.		Green et	Without SF			With SF		
ADV+*que*	**global**	**indiv.**	**al. (2013)**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
ainsi que	76.7	76.7	81.44	96.00	91.14	93.50	95.94	89.87	92.81
alors que	88.2	88.2	95.10	92.78	92.78	92.78	93.81	93.81	93.81
autant que	86.0	86.0	92.00	86.95	65.21	74.53	86.66	70.65	77.84
bien que	37.4	62.6	55.22	86.84	89.18	88.00	91.66	89.18	90.41
encore que	21.5	78.5	64.52	72.72	80.00	76.19	92.85	65.00	76.47
maintenant que	55.8	55.8	87.01	85.24	77.61	81.25	90.91	74.62	81.96
tant que	20.4	79.6	90.91	78.94	75.00	76.92	82.35	70.00	75.67
Total	56.4	75.3	83.06	88.71	**82.03**	85.24	**91.57**	81.79	**86.41**

Table 5: MORPH link prediction for ADV+*que* constructions: precision of global majority baseline, precision of individual per-construction baseline, precision of Green et al. (2013) constituent parser, precision, recall and F1 of our dependency parser without and with subcat features.

core que. However, this is not verified for all individual ADV+*que* constructions. The table also shows an important variety among the seven complex conjunctions studied. Some of them are very well predicted (F1 = 93.5) while others are poorly predicted (F1 = 75.67). This is partly due to the tendency of some ADV+*que* sequences to be part of larger frozen or semi-frozen constructions and to be used with a different semantico-syntactic behavior. An error analysis performed on the *tant que* sequence revealed that 40% of the errors were due to the occurrence of *tant que* as part of the larger *en tant que* expression, while 20% of the errors were due to the usage of *tant que* as a comparative expression.

7.2 *de*+DET Constructions

SF	LAS	UAS	MORPH	Prec.	Rec.
no	89.02	90.23	145	85.85	81.12
yes	88.37	89.67	145	86.52	83.92

Table 6: Attachment scores, count, precision and recall of the MORPH dependency for *de*+DET in FTB test, without and with subcat features (SF).

Table 6 reports the results of the same experiments on *de*+DET constructions. It shows that the frequency of *de*+DET constructions is higher than ADV+*que* constructions. It also shows that the introduction of subcat features has a positive impact on the prediction of the MORPH dependency, but a negative effect on the attachment scores.

Table 7 reveals that the prediction of the correct structure of *de*+DET constructions is more difficult than that of ADV+*que* constructions for the parser. Here, not only the majority class is the non-MWE analysis (63.5%), but also there is higher ambiguity because of nominal and adverbial complements that have the same structure. This impacts the performance of the Stanford parser, which overgenerates MORPH links, achieving the lowest precision for all constructions except for *des*. Results also show that the introduction of subcat features has an important impact on the quality of the prediction (F1 jumps from 75% to 84.67%). The use of subcat features slightly improves the identification of *de les*, which is a determiner most of the time. On the other hand, it greatly improves F1 for other constructions, which appear less often as determiners. We believe that the higher impact of subcat frames on *de*+DET is mainly due to the fact that the number of verbs licensing complements introduced by the preposition *de* is higher than the number of verbs licensing complements introduced by the conjunction *que* (see Table 3). Therefore, the parser trained without subcat features can only rely on the examples present in the FTB which are proportionally smaller in the first case than in the second.

8 Conclusions

This paper introduced and evaluated a joint parsing and MWE identification model that can effectively detect and represent ambiguous complex function words. The difficulty of processing such expressions is underestimated because of their limited variability. They often are pre-grouped as words-with-spaces in many parsing architectures (Sag et al., 2002). However, we did not use gold tokenization, unrealistic for ambiguous MWEs (Nivre and Nilsson, 2004; Korkontze-

	Baseline prec.		Green et	Without SF			With SF		
de+DET	**global**	**indiv.**	**al. (2013)**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
de le	66.9	79.0	56.96	72.50	64.44	68.23	85.41	91.11	88.17
de la	79.0	77.5	22.83	58.13	86.20	69.44	81.25	89.65	85.24
de les	22.5	66.9	87.72	97.36	74.00	84.09	98.70	76.00	85.87
de l'	83.1	83.1	18.55	57.14	69.56	62.74	64.51	86.95	74.07
Total	63.5	76.6	44.37	77.00	73.09	75.00	**86.70**	**82.74**	**84.67**

Table 7: MORPH link prediction for *de*+DET constructions: precision of global majority baseline, precision of individual per-construction baseline, precision of Green et al. (2013) constituent parser, precision, recall and F1 of our dependency parser without and with subcat features.

los and Manandhar, 2010).

We proposed to deal with these constructions during parsing, when the required syntactic information to disambiguate them is available. Thus, we trained a graph-based dependency parser on a modified treebank where complex function words were linked with a MORPH dependency. Our results demonstrate that a standard parsing model can correctly learn such special links and predict them for unseen constructions. Nonetheless, the model is more accurate when we integrate external information from a syntactic lexicon. This improved precision for ADV+*que* and specially *de*+DET constructions. For the latter, F1 improved in almost 10%, going from 75% to 84.61%.

This study raised several linguistic and computational questions. Some complex function words include more than two elements, like *si bien que* (*so much that*) and *d'autant (plus) que* (*especially as*). Moreover, they may contain nested expressions with different meanings and structures, e.g. *tant que* (*as long as*) is a conjunction but *en tant que* (*as*) is a preposition. The same applies for quantified partitive determiners, like *beaucoup de* (*much*) and *un (petit) peu de* (*a (little) bit of*). Their identification and representation is planned as a future extension to this work.

We also would like to compare our approach to sequence models (Schneider et al., 2014). Careful error analysis could help us understand in which cases syntactic features can help. Moreover, different variants of the syntactic features and more sophisticated representation for syntactic lexicons can help improve MWE parsing further. For instance, we represent the subcat features of pronominal verbs and their simple versions with the same features, but they should be distinguished, e.g. *se rappeler* (*remember*) is +DE but *rappeler* (*remind*) is -DE.

Acknowledgements

This work has been funded by the French Agence Nationale pour la Recherche, through the projects ORFEO (ANR-12-CORP-0005) and ASFALDA (ANR-12-CORD-023) and by FAPERGS and CNRS through the AIM-WEST project.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for french. In Anne Abeillé, editor, *Treebanks: building and using parsed corpora*, pages 165–168. Kluwer academic publishers, Dordrecht, The Netherlands.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.

Marco Baroni and Silvia Bernardini, editors. 2006. *Wacky! Working papers on the Web as Corpus*. GEDIT, Bologna, Italy. 224 p.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In Huang and Jurafsky (Huang and Jurafsky, 2010), pages 89–97.

António Horta Branco and João Ricardo Silva. 2003. Contractions: Breaking the tokenization-tagging circularity. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, and Maria das Graças Volpe Nunes, editors, *Proc. of the 6th PROPOR (PROPOR 2003)*, volume 2721 of *LNCS (LNAI)*, pages 195–195, Faro, Portugal, Jun. Springer.

Marie Candito and Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proc. of the 52nd ACL (Volume 1: Long Papers)*, pages 743–753, Baltimore, MD, USA, Jun. ACL.

Marie Candito, Benoît Crabbé, Pascal Denis, and François Guérin. 2009. Analyse syntaxique du français : des constituants aux dépendances. In

Proc. of Traitement Automatique des Langues Naturelles, Senlis, France, Jun.

Matthieu Constant and Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In Kordoni et al. (Kordoni et al., 2011), pages 49–56.

Matthieu Constant, Marie Candito, and Djamé Seddah. 2013a. The LIGM-Alpage architecture for the SPMRL 2013 shared task: Multiword expression analysis and dependency parsing. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 46–52, Seattle, Washington, USA, October. Association for Computational Linguistics.

Matthieu Constant, Joseph Le Roux, and Anthony Sigogne. 2013b. Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. *ACM Trans. Speech and Lang. Process. Special Issue on MWEs: from theory to practice and use, part 2 (TSLP)*, 10(3).

Eric De La Clergerie. 2013. Exploring beam-based shift-reduce dependency parsing with DyALog: Results from the SPMRL 2013 shared task. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 53–62, Seattle, Washington, USA, October. Association for Computational Linguistics.

Yoav Goldberg and Michael Elhadad. 2011. Joint hebrew segmentation and parsing using a PCFGLA lattice parser. In *Proc. of the 49th ACL: HLT (ACL HLT 2011)*, pages 704–709, Portland, OR, USA, Jun. ACL.

Spence Green and Christopher D. Manning. 2010. Better Arabic parsing: Baselines, evaluations, and analysis. In Huang and Jurafsky (Huang and Jurafsky, 2010), pages 394–402.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Comp. Ling.*, 39(1):195–227.

Chu-Ren Huang and Dan Jurafsky, editors. 2010. *Proc. of the 23rd COLING (COLING 2010)*, Beijing, China, Aug. The Coling 2010 Organizing Committee.

Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors. 2011. *Proc. of the ACL Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, Portland, OR, USA, Jun. ACL.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Proc. of HLT: The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 636–644, Los Angeles, California, Jun. ACL.

S. Kübler, R. McDonald, and J. Nivre. 2009. Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 1(1):1–127.

Nidhi Kulkarni and Mark Finlayson. 2011. jMWE: A Java toolkit for detecting multi-word expressions. In Kordoni et al. (Kordoni et al., 2011), pages 122–124.

Igor A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, New York.

Joakim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Proc. of the LREC Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*, Lisbon, Portugal.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd CICLing (CICLing-2002)*, volume 2276/2010 of *LNCS*, pages 1–15, Mexico City, Mexico, Feb. Springer.

Nathan Schneider, Emily Danchik, Chris Dyer, and A. Noah Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1*, pages 193–206.

Violeta Seretan. 2011. *Syntax-Based Collocation Extraction*, volume 44 of *Text, Speech and Language Technology*. Springer, Dordrecht, Netherlands, 1st edition. 212 p.

Karel van den Eynde and Piet Mertens. 2003. La valence: l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, (13):63–104.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Jason Eisner, editor, *Proc. of the 2007 Joint Conference on EMNLP and Computational NLL (EMNLP-CoNLL 2007)*, pages 1034–1043, Prague, Czech Republic, Jun. ACL.

Veronika Vincze, István Nagy T., and Richárd Farkas. 2013a. Identifying English and Hungarian light verb constructions: A contrastive approach. In *Proc. of the 51st ACL (Volume 2: Short Papers)*, pages 255–261, Sofia, Bulgaria, Aug. ACL.

Veronika Vincze, István Nagy T., and János Zsibrita. 2013b. Learning to detect english and hungarian light verb constructions. *ACM Trans. Speech and Lang. Process. Special Issue on MWEs: from theory to practice and use, part 1 (TSLP)*, 10(2).

Veronika Vincze, János Zsibrita, and István Nagy T. 2013c. Dependency parsing for identifying hungarian light verb constructions. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 207–215, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Eric Wehrli, Violeta Seretan, and Luka Nerima. 2010. Sentence analysis and collocation identification. In Éric Laporte, Preslav Nakov, Carlos Ramisch, and Aline Villavicencio, editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 27–35, Beijing, China, Aug. ACL.