Chapter 10

Sequence models and lexical resources for MWE identification in French

Manon Scholivet
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Carlos Ramisch
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Silvio Cordeiro
Institute of Informatics, Federal University of Rio Grande do Sul, Brazil
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

We present a simple and efficient sequence tagger capable of identifying continuous multiword expressions (MWEs) of several categories in French texts. It is based on conditional random fields (CRF), using as features local context information such as previous and next word lemmas and parts of speech. We show that this approach can obtain results that, in some cases, approach more sophisticated parser-based MWE identification methods without requiring syntactic trees from a treebank. Moreover, we study how well the CRF can take into account external information coming from both high-quality hand-crafted lexicons and MWE lists automatically obtained from large monolingual corpora. Results indicate that external information systematically helps improving the tagger's performance, compensating for the limited amount of training data.

1 Introduction

Identifying multiword expressions (MWEs) in running texts with the help of lexicons could be considered as a trivial search-and-replace operation. In theory, one could simply scan the text once and mark (e.g. join with an underscore)

Manon Scholivet, Carlos Ramisch & Silvio Cordeiro. 2018. Sequence models and lexical resources for MWE identification in French. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 263–297. Berlin: Language Science Press. DOI:10.5281/zenodo.1469567

all sequences of tokens that appear as headwords in the MWE lexicons. Direct matching and projection of lexical entries onto the corpus can be employed as a simple yet effective preprocessing step prior to dependency parsing (Nivre & Nilsson 2004) and machine translation (Carpuat & Diab 2010). Upon recognition, the identified member words of an MWE can be concatenated and treated as single token, that is, a "word with spaces", as suggested by Sag et al. (2002).

However, this simple pipeline will fail when dealing with frequent categories of MWEs that present some challenging characteristics such as **variability** and **ambiguity**. For many MWE categories, **variability** due to morphological inflection may pose problems. For instance, if a lexicon contains the idiom *to make a face*, string matching will fail to identify it in *children are always* ***making faces*** because the verb and the noun are inflected.[1] Since lexicons usually contain canonical (lemmatised) forms, matching must take inflection into account. This can be carried out by (a) pre-analysing the text and matching lemmas instead of surface-level word forms (Finlayson & Kulkarni 2011), or by (b) looking up lexicons containing inflected MWEs (Silberztein et al. 2012).

Things get more complicated when the target MWEs are ambiguous, though. An MWE is **ambiguous** when its member words can co-occur without forming an expression. For instance, *to make a face* is an idiom meaning 'to show a funny facial expression', but it can also be used literally when someone is making a snowman (Fazly et al. 2009). Additionally, the words in this expression can co-occur by chance, not forming a phrase (Boukobza & Rappoport 2009; Shigeto et al. 2013). This is particularly common for multiword function words such as prepositions (e.g. *up to*), conjunctions (e.g. *now that*) and adverbials (e.g. *at all*). For example, *up to* is an MWE in *they accepted* ***up to*** *100 candidates* but not in *you should look it up to avoid making a typo*. Similarly, *at all* is an adverbial in *they accepted no candidates* ***at all***, but not in *this train does not stop at all stations*. Context-dependent statistical methods (Fazly et al. 2009; Boukobza & Rappoport 2009) and syntax-based methods (Candito & Constant 2014; Nasr et al. 2015) are usually employed to deal with semantic ambiguity and accidental co-occurrence, respectively.

In addition to variability and ambiguity, an additional challenge stems from the absence or limited coverage of high-quality hand-crafted lexical resources containing MWEs for many languages. Therefore, it is not always possible to em-

[1] In addition, the determiner *a* is not mandatory. However, discontinuous expressions containing optional intervening words are out of the scope of this work because our method is based on sequence models and our corpora only contain continuous MWEs. An adaptation of sequence models to discontinuous expressions has been proposed by Schneider, Danchik, et al. (2014).

ploy purely symbolic look-up methods for MWE identification. Statistical methods are an interesting alternative, since one can learn generic models for MWE identification based on corpora where MWEs have been manually annotated. If enough evidence is provided and represented at the appropriate level of granularity, the model can make generalizations based on commonly observed patterns. It may then be able to identify MWE instances that have never occurred in annotated training data. However, annotated corpora often do not contain enough training material for robust MWE identification. Complementary evidence can be obtained with the help of unsupervised MWE discovery methods that create MWE lists from raw corpora, which are then considered as if they were hand-crafted lexicons. In short, the heterogeneous landscape in terms of available resources (annotated corpora, hand-crafted lexicons) motivates the development of statistical MWE identification models that can exploit external hand-crafted and automatically constructed lexicons as a complementary information source (Constant & Sigogne 2011; Schneider, Danchik, et al. 2014; Riedl & Biemann 2016).

We propose a simple, fast and generic sequence model for tagging continuous MWEs based on conditional random fields (CRF). It cannot deal with discontinuous expressions, but is capable of modelling variable and highly ambiguous expressions. Moreover, we propose a simple adaptation to integrate information coming from external lexicons. Another advantage of our CRF is that we do not need syntactic trees to train our model, unlike methods based on parsers (Le Roux et al. 2014; Nasr et al. 2015; Constant & Nivre 2016). Moreover, for expressions that are syntactically fixed, it is natural to ask ourself if we really need a parser for this task. Parsers are good for non-continuous MWEs, but we hypothesise that continuous expressions can be modelled by sequence models that take ambiguity into account, such as CRFs. Regardless of the syntactic nature of these ambiguities, we expect that the highly lexicalised model of the CRF compensates for its lack of structure.

The present chapter contains three significant extensions with respect to our previous publication at the MWE 2017 workshop (Scholivet & Ramisch 2017). First, we train and test our models on two complementary datasets containing nominal expressions and general MWEs in French. Second, we study the integration of automatically constructed MWE lexicons obtained with the help of MWE discovery techniques. Third, we study the performance of our system on particularly hard MWE instances such as those including variants and those that do not occur in the training corpora.

In short, we demonstrate that, in addition to being well suited to identifying highly ambiguous MWEs in French (Scholivet & Ramisch 2017), the proposed

model and its corresponding free implementation[2] can also be applied to identify other MWE categories and use other types of external lexicons. We believe that this approach can be useful (a) when no treebank is available to perform parsing-based MWE identification, (b) when large monolingual corpora are available instead of hand-crafted lexical resources, and (c) as a preprocessing step to parsing, which can improve parsing quality by reducing attachment ambiguities (Candito & Constant 2014; Nivre & Nilsson 2004).

2 Related work

Token identification of MWEs in running text can be modelled as a machine learning problem, building an identification model from MWE-annotated corpora and treebanks. To date, it has been carried out using mainly two types of models: sequence taggers and parsers. Sequence taggers such as CRFs, structured support vector machines and structured perceptron allow disambiguating MWEs using local feature sets such as word affixes and surrounding word and POS *n*-grams. Parsers, on the other hand, can take into account longer-distance relations and features when building a parse tree, at the expense of using more complex models.

Sequence taggers have been proven useful in identifying MWEs. MWE identification is sometimes integrated with POS tagging in the form of special tags. Experiments have shown the feasibility of sequence tagging for general expressions and named entities in English (Vincze et al. 2011), verb-noun idioms in English (Diab & Bhutada 2009) and general expressions in French (Constant & Sigogne 2011) and in English (Schneider, Danchik, et al. 2014; Riedl & Biemann 2016). Shigeto et al. (2013) tackle specifically English function words and build a CRF from the Penn Treebank, additionally correcting incoherent annotations. We develop a similar system for French, using the MWE annotation of the French Treebank as training data and evaluating the model on a dedicated dataset.

Parsing-based MWE identification requires a treebank annotated with MWEs. Lexicalised constituency parsers model MWEs as special non-terminal nodes included in regular rules (Green et al. 2013). In dependency parsers, it is possible to employ a similar approach, using special dependency labels to identify relations between words that make up an expression (Candito & Constant 2014).

The work of Nasr et al. (2015) is a parsing-based approach evaluated on highly ambiguous grammatical MWEs in French (§5.1). In their work, they link word

[2]The CRF-MWE tagger described in this chapter is included in the mwetoolkit in the form of 2 scripts: `train_crf.py` and `annotate_crf.py`, freely available at http://mwetoolkit.sf.net/

sequences belonging to complex conjunctions such as *bien que* 'well that' ⇒ 'although' and partitive determiner such as *de la* 'of the' ⇒ 'some', using a special dependency link called `morph`, similar to Universal Dependencies' `compound` relation (Nivre et al. 2016). On the other hand, these word sequences can occur by chance, such as in *Je pense bien que je suis malade.* 'I think well that I am sick.' ⇒ 'I really think that I am sick'. Then, the adverb *well* modifies the verb *think*, which in turn has a complement introduced by *that*. Nasr et al. (2015) train a second-order graph-based dependency parser to distinguish `morph` from other syntactic relations, implicitly identifying MWEs. In addition to standard features, they extract features from a valence dictionary specifying whether a given verb licences complements introduced by *que* 'that' or *de* 'of'.

Our hypothesis is that parsing-based techniques like this are not required to obtain good performances on continuous expressions. Our paper adapts a standard CRF model inspired on the ones proposed by Constant & Sigogne (2011), Riedl & Biemann (2016) and Shigeto et al. (2013) to deal with continuous MWEs.

Concerning external lexical resources, Nasr et al. (2015) have shown that their features extracted from a valence dictionary can significantly improve identification. Moreover, most systems based on sequence taggers also integrate additional hand-crafted lexicons to obtain good results (Constant & Sigogne 2011; Schneider, Danchik, et al. 2014). Nonetheless, the integration of automatically discovered lexicons of MWEs has not been explored by many authors, with the notable exception of Riedl & Biemann (2016). We show that our CRF can naturally handle automatically and manually constructed lexicons and that, in both cases, the system benefits from the extra information present in the lexicons.

3 A CRF-based MWE tagger

Linear-chain conditional random fields (CRFs) are an instance of stochastic models that can be employed for sequence tagging (Lafferty et al. 2001). Each input sequence T is composed of $t_1 \dots t_n$ tokens considered as an observation. Each observation is tagged with a sequence $Y = y_1 \dots y_n$ of tags corresponding to the values of the hidden states that generated them. CRFs can be seen as a discriminant version of hidden Markov models, since they model the conditional probability $P(Y|T)$. This makes them particularly appealing since it is straightforward to add customised features to the model. In first-order linear-chain CRFs, the probability of a given output tag y_i for an input word x_i depends on the tag of the neighbour token y_{i-1}, and on a rich set of features of the input $\phi(T)$, that can range over any position of the input sequence, including but not limited to

the current token t_i. CRF training consists in estimating individual parameters proportional to $p(y_i, y_{i-1}, \phi(T))$.

The identification of continuous MWEs is a segmentation problem. We use a tagger to perform this segmentation, employing the well-known Begin-Inside-Outside (BIO) encoding (Ramshaw & Marcus 1995). In BIO, every token t_i in the training corpus is annotated with a corresponding tag y_i with values `B`, `I` or `O`. If the tag is `B`, it means the token is the beginning of an MWE. If it is `I`, this means the token is inside an MWE. `I` tags can only be preceded by another `I` tag or by a `B`. Finally, if the token's tag is `O`, this means the token is outside the expression, and does not belong to any MWE. An example of such encoding for the 2-word expression *de la* 'some' in French is shown in Figure 1.

i	-2	-1	0	1	2	3
w_i	*Il*	*jette*	*de*	*la*	*nourriture*	*périmée*
y_i	O	O	B	I	O	O
	He	*discards*	*some*		*food*	*expired*

Figure 1: Example of BIO tagging of a French sentence containing a *De*+determiner expression, assuming that the current word (w_0) is *de*.

For our experiments, we have trained a CRF tagger with the CRFSuite toolkit[3] (Okazaki 2007). We used a modified version of the French treebank (Abeillé et al. 2003) as training, development, and test data, and the MORPH dataset[4] (Nasr et al. 2015) as development and test data. We additionally include features from an external valence lexicon, DicoValence[5] (van den Eynde & Mertens 2003), and from an automatically constructed lexicon of nominal MWEs obtained automatically from the frWaC corpus (Baroni & Bernardini 2006) with the help of the mwetoolkit (Ramisch 2014).

3.1 CRF features

Our set of features $\phi(T)$ contains 37 different combinations of values (henceforth referred to as the ALL feature set). Our features are inspired on those proposed by Constant & Sigogne (2011), and are similar to those used by Schneider, Danchik, et al. (2014) and Riedl & Biemann (2016). The feature templates described below

[3]http://www.chokkan.org/software/crfsuite/

[4]http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/morph

[5]http://bach.arts.kuleuven.be/dicovalence/

consider that the current token T_0 has surface form W_0, lemma L_0 and POS P_0. In addition to output tag bigrams (CRF's first-order assumption), we consider the following feature templates in our model, to be regarded in conjunction with the current tag to predict:

- Single-token features (T_i):[6]
  - W_0 : wordform of the current token
  - L_0 : lemma of the current token
  - P_0 : POS tag of the current token
  - W_i, L_i and P_i: wordform, lemma or POS of previous ($i \in \{-1, -2\}$) or next ($i \in \{+1, +2\}$) tokens
- N-gram features (bigrams $T_{i-1}T_i$ and trigrams $T_{i-1}T_iT_{i+1}$):
  - $W_{i-1}W_i$, $L_{i-1}L_i$, $P_{i-1}P_i$: wordform, lemma and POS bigrams of previous-current ($i = 0$) and current-next ($i = 1$) tokens
  - $W_{i-1}W_iW_{i+1}$,$L_{i-1}L_iL_{i+1}$, $P_{i-1}P_iP_{i+1}$: wordform, lemma and POS trigrams of previous-previous-current ($i = -1$), previous-current-next ($i = 0$) and current-next-next ($i = +1$) tokens
- Orthographic features (ORTH):
  - HYPHEN and DIGITS: the current wordform W_i contains a hyphen or digits
  - F-CAPITAL: the first letter of the current wordform W_i is uppercase
  - A-CAPITAL: all letters of the current wordform W_i are uppercase
  - B-CAPITAL: the first letter of the current word W_i is uppercase, and it is at the beginning of a sentence.
- Lexicon features (LF), described in more detail in §4.3:
  - QUEV: the current wordform W_i is *que*, and the closest verb to the left licences a complement introduced by *que* according to the valence dictionary DicoValence.[7]
  - DEV: the current wordform W_i is *de*, and the closest verb to the left licences a complement introduced by *de* according to the valence dictionary DicoValence.

[6] T_i is a shortcut denoting the group of features W_i, L_i and P_i for a token T_i. In other words, each token T_i is a tuple (W_i,L_i,P_i). The same applies to n-grams.

[7] QUEV and DEV are sequential versions of the *subcat features* proposed by Nasr et al. (2015).

- Association measures (AM) between the current token's lemma L_i and the previous tokens' lemmas:
  - MLE: probability of the lemma sequence estimated using maximum likelihood estimation
  - PMI: pointwise mutual information of the lemma sequence.
  - DICE: Dice's coefficient of the lemma sequence
  - T-MEAS: Student's t-score of the lemma sequence
  - LL: log-likelihood ratio between the current lemma and the previous lemma

Our proposed feature set is similar to previous work, with only minor differences (Constant & Sigogne 2011; Schneider, Onuffer, et al. 2014; Riedl & Biemann 2016). Like all previous models, we encode output tags with BIO, and we consider as features the surface form of the current token, of surrounding tokens, and bigrams of those. Our orthographic features are practically identical to related work, but all previously proposed models include 4- to 5-character prefixes and suffixes, which we do not. The features proposed by Constant & Sigogne (2011) are only based on surface forms of words, given that their task is to jointly predict POS and MWE tags. On the other hand, the features of Schneider, Onuffer, et al. (2014) and Riedl & Biemann (2016) are based on current and surrounding lemmas and POS tags, and so are ours. Differently from these two articles, we rely on token trigram features and we do not use mixed lemma+POS features. The lemma-based features of Schneider, Onuffer, et al. (2014) are quite different from ours, because they are conditioned on particular POS tags. The main differences between previous models and ours are in the lexicon features: Constant & Sigogne (2011) and Schneider, Onuffer, et al. (2014) employ hand-crafted lexicons and extract more detailed information from them than we do. Riedl & Biemann (2016) cover both hand-crafted and automatically built lexicons. Their feature set has one feature in common with ours: Student's t-score. In short, the features are similar in nature, but present some arbitrary variation in their implementations, in addition to some variation due to the nature of the available lexicons and corpora.

Our training corpora contain syntactic dependency information. However, we decided not to include it as CRF features for two main reasons. First, we wanted to evaluate the hypothesis that sequence-based methods can perform MWE identification without resorting to treebanks, as opposed to parser-based identification. Second, representing syntactic structure in a CRF is tricky as the linear-chain model in our experiments is not adequate for representing general graphs.

Nonetheless, adding features based on simplified syntactic information (e.g. the dependency label of each word with respect to its parent) is feasible and represents an interesting idea for future work.

4 Experimental setup

In order to evaluate our systems, we test them on four MWE categories in French:

- Adverb+*que* expressions (AQ): in French, adverbs (such as *bien* 'well') and the subordinating conjunction *que* 'that' are frequently combined to build complex conjunctions such as *bien que* 'well that' ⇒ 'although'. This category was chosen because (a) these expressions present little variability,[8] and (b) they are highly ambiguous, since their components can co-occur by chance, as in *il sait bien que tu mens.* 'he knows well that you lie.' ⇒ 'he really knows that you are lying'. Thus, we can focus on ambiguity as a challenging problem to model.

- *De*+determiner expressions (DD): in French, partitive and plural determiners are formed by the word *de* 'of' followed by a definite article, for instance, *il mange de la salade, du pain et des fruits.* 'he eats of the.SG.FEM salad, of-the.SG.MASC bread and of-the.PL fruit.' ⇒ 'he is eating some salad, bread and fruit'. Similarly to AQ, these constructions present little variation[9] and are ambiguous with preposition+article combinations such as *il parle de la salade* (lit. *he talks of the salad*) 'he talks about the salad'. Their disambiguation is challenging because it relies on the argumental structure of the verb governing the noun. Moreover, these are among the most frequent tokens in a corpus of French (Nasr et al. 2015).

- Nominal expressions: at a first moment, we focus on the identification of nominal expressions for two reasons. First, they present morphological variability but are syntactically fixed, making CRFs particularly suitable to model them. Second, we test the inclusion of automatically calculated association measures as features in the identification model, and our lexicon of pre-calculated association measures contains only nominal MWEs.

- General MWEs: finally, we evaluate our model on a corpus containing several categories of continuous MWEs. These include nominal expressions,

[8] The only variability that must be taken into account is that *que* is sometimes written as *qu'* when the next word starts with a vowel.

[9] Except for contractions *de+le=du* and *de+les=des*

complex conjunctions and determiners such as AQ and DD combinations, fixed prepositional phrases, multiword named entities, some limited continuous verbal expressions such as *avoir lieu* (lit. *have place*) 'take place', and so on. Our training and test corpora do not contain any labels distinguishing these MWE categories. Therefore the only category-based analysis we perform relies on the POS tags of the component words (for nominal MWEs).

In our experiments, we used two annotated corpora: the French treebank and the MORPH dataset. Other corpora annotated with MWEs in French do exist (Laporte et al. 2008; Savary et al. 2017). However, we chose to evaluate our model on a dataset for which, at the time of writing this chapter, many studies on MWE identification methods have been reported (the French treebank) and on an in-house dataset focusing on ambiguous MWEs (MORPH). Hence, we can compare our sequence model with state-of-the art results and verify whether they are adequate to recognise ambiguous MWEs. Evaluation on other corpora is left for future work.

4.1 The French treebank

We train and test our models on the MWE-annotated French treebank (FTB), available in CONLL format and automatically transformed into the CRFsuite format. The FTB is traditionally split into three parts: train, dev and test. We train our systems systematically on the training part of the FTB, that we adapted to keep only the MWEs we are interested in. For the experiments where we considered general MWEs and nominal MWEs, we used the FTB version of the SPMRL shared task (Seddah et al. 2013). The FTB dev and test corpora were employed respectively for feature engineering and evaluation. For each word, the corpus contains its wordform, lemma, POS (15 different coarse POS tags), and syntactic dependencies (that were ignored).

In the original corpus, MWE information is represented as words with spaces. For instance, *bien_que* appears as a single token with underscores when it is a complex conjunction, whereas accidental co-occurrence is represented as two separate tokens *bien* and *que*. We argue that using such gold tokenisation is unrealistic, especially in the case of ambiguous MWEs. We thus systematically split single-token MWEs and added an extra column containing MWE tags using BIO encoding (§3). Even though this preprocessing might sound artificial, we believe that it provides a more uniform treatment to ambiguous constructions, closer to their raw-text form. This assumption is in line with the latest developments in

the dependency parsing community, which has evolved from parser evaluation on gold tokenisation (Buchholz & Marsi 2006) to evaluation on raw text (Zeman et al. 2017).

The MWE-BIO tags were generated using the following transformation heuristics in the case of ambiguous AQ and DD MWEs:

- For AQ expressions:
  1. We scan the corpus looking for the lemmas *ainsi_que*, *alors_que*, *autant_que*, *bien_que*, *encore_que*, *maintenant_que* and *tant_que*.
  2. We split them into two tokens and tag the adverb as B and *que* as I.
- For DD expressions:
  1. We scan the corpus looking for the wordforms *des*, *du*, *de_la* and *de_l'*. Due to French morphology, *de* is sometimes contracted with the articles *les* (determinate plural) and *le* (determinate singular masculine). Contractions are mandatory for both partitive and preposition+determiner uses. Therefore, we systematically separate these pairs into two tokens.[10]
  2. If a sequence was tagged as a determiner (D), we split the tokens and tag *de* as B and the determiner as I.
  3. Contractions (*des*, *du*) tagged as P+D (preposition+determiner) were split in two tokens, both tagged as O.
- All other tokens are tagged as O, including all other categories of MWEs.

For the newly created tokens, we assign individual lemmas and POS tags. The word *de* is systematically tagged as P (preposition), not distinguishing partitives from prepositions at the POS level. The input to the CRF is a file containing one word per line, BIO tags as targets, and FeatureName=value pairs including *n*-grams of wordforms, lemmas and POS tags, as described in §3.1.

In the case of nominal MWEs, we applied the same procedure as for AQ pairs to the MWEs matching certain sequences of POS tags[11]. We accept that tokens can be separated by punctuation marks, as in the proper noun *Bouches-du-Rhône*.

[10] An alternative to this preprocessing would be to keep contractions untokenised, and to assign a single B tag to those representing determiners. However, this would actually move the task of MWE identification to the POS tagger, which would need to choose whether the token is a determiner or a contracted preposition before MWE identification.

[11] The exact regular-expression pattern is: `(A.N) | (N.(PONCT.)?(A |(P+D.(PONCT.)?N) | (P.(PONCT.)?(D.)?(PONCT.)?N) | N)+)`.

When the MWE starts with a noun (N), it can be followed by one or more adjectives (A), nouns (N), or nouns preceded by prepositions (P) optionally including determiners (D) between the preposition and the noun. The matched nominal MWEs include combinations composed of:

- adjective noun: *premier ministre* 'prime minister';
- noun adjective: *corps médical* 'medical community';
- noun-noun: *maison mère* 'parent company';
- noun preposition noun: *motion de censure* 'motion of censure';
- noun preposition determiner noun: *impôt sur le revenu* 'income tax';
- noun preposition+determiner noun: *ironie du sort* 'twist of fate'.

4.2 MORPH dataset

We used the MORPH dataset introduced by Nasr et al. (2015) as test and development corpora for ambiguous AQ and DD expressions. It contains a set of 1,269 example sentences, each containing one of 7 ambiguous AQ constructions and 4 ambiguous DD constructions. To build this corpus, around 100 sentences for each of the 11 target constructions were extracted from the frWaC corpus and manually annotated as to whether they contain a multiword function word (MORPH) or accidental cooccurrence (OTHER). We have preprocessed the raw sentences as follows:

1. We have automatically POS tagged and lemmatized all sentences using an off-the-shelf POS tagger and lemmatizer independently trained on the FTB.[12] This information is given to the CRF as part of its features.
2. We have located the target construction in the sentence and added BIO tags according to the annotation provided: target pairs annotated as MORPH were tagged B + I, target pairs annotated as OTHER were tagged O.
3. For each target construction, we have taken the first 25 sentences as development corpus (dev, 275 sentences) and the remaining sentences for testing (test, 994 sentences).

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

4. We created four targeted datasets: DEV$_{AQ}$, DEV$_{DD}$, FULL$_{AQ}$ and FULL$_{DD}$, where the different construction classes are separated, in order to perform feature selection.

Table 1 summarises the corpora covered by our experiments in terms of number of tokens and MWEs. We trained all systems on the training portion of the FTB with different tokenisation choices, depending on the target MWE.[13] The density of AQ and DD being too low in FTB-dev and FTB-test, we tune and evaluate our model for AQ and DD constructions on the MORPH dataset. For nominal and general MWEs, however, we use the FTB-dev and FTB-test portions.

Table 1: Number of tokens and MWEs in each corpus of our experiments.

Corpus	Portion	Target MWEs	#tokens	#MWEs
FTB	train	AQ	285,909	216
FTB	train	DD	285,909	1,356
FTB	train	Nominal	443,115	6,413
FTB	train	General	443,115	23,522
FTB	dev	Nominal	38,820	686
FTB	dev	General	38,820	2,117
FTB	test	Nominal	75,217	1,019
FTB	test	General	75,217	4,041
MORPH	FULL$_{AQ}$	AQ	11,839	730
MORPH	FULL$_{DD}$	DD	8,319	539

4.3 External lexicons

The verbal valence dictionary DicoValence specifies the allowed types of complements per verb sense in French. For each verb, we extract two binary flags:

- QUEV: one of the senses of the verb has one object that can be introduced by *que*.[14]

[13] FTB-train for AQ/DD and for nominal/general MWEs is the same corpus, but the number of tokens differs because all MWEs other than AQ and DD were represented using words-with-spaces in FTB-train for AQ/DD.

[14] In DicoValence, an object P1, P2 or P3 licenses a complementizer QPIND

- DEV: one of the senses of the verb has a locative, temporal or prepositional paradigm that can be introduced by *de*.[15]

We also use a lexicon containing nominal MWEs that were automatically extracted from the frWaC. They were obtained with the help of the mwetoolkit by first extracting all lemma sequences that match the nominal MWE pattern described above. Then, for each sequence, we calculate its number of occurrences as well as the number of occurrences of its member words, which are then used to calculate the association measures listed in §3.1.

When integrating this lexicon in the CRF as features, special treatment was required for overlapping expressions. If a given token belonged to more than one overlapping MWE, we considered the maximum value of the association measures. Moreover, since CRFs cannot deal with real-valued features, we have quantized each association score through a uniform distribution that assigned an equal number of expressions to each bin.

4.4 Evaluation measures

For general and nominal MWEs, we analyse the performance on the FTB reported by the evaluation script of PARSEME shared task (see Savary et al. 2018 [this volume]).[16] The script provides us with two different scores: one based on MWEs, and one based on MWE tokens. The MWE-based measure requires that all tokens in the MWE be predicted by the system, whereas the token-based measure is calculated based on each token individually, so that partially correct predictions are taken into account. Each variant (MWE-based and token-based) is reported in terms of precision, recall and F-measure. In this work, we will particularly focus on the F-measure.

For AQ and DD combinations, we evaluated on the MORPH dataset. We report precision (P), recall (R) and F-measure (F_1) of MWE tags. In other words, instead of calculating micro-averaged scores over all BIO tags, we only look at the proportion of correctly guessed `B` tags. Since all of our target expressions are composed of exactly 2 contiguous words, we can use this simplified score because all `B` tags are necessarily followed by exactly one `I` tag. As a consequence, the measured precision, recall and F-measure scores on `B` and `I` tags are identical.

[15] In DicoValence, the paradigm is PDL, PT or PP.

[16] http://multiword.sf.net/sharedtask2017

5 Results

We present our results for different categories of MWEs, performing feature selection on the *dev* datasets. §5.1 presents an evaluation of our approach on ambiguous AQ and DD expressions. §5.2 evaluates the broader category of nominal MWEs. §5.3 then extends the latter results to an evaluation of all MWEs. §5.4 compares the best results we obtained against the state of the art. Finally, §5.5 presents the results of a detailed analysis focusing on variable and unseen MWEs.

5.1 Experiments on AQ and DD expressions

Our first evaluation was performed on the *dev* part of the MORPH dataset. We consider a subset of around 1/4 sentences containing AQ constructions (DEV_{AQ}, 175 sentences) and DD constructions (DEV_{DD}, 100 sentences). We evaluate the results under different levels of feature selection, regarding both coarse groups and individual features.

In these experiments, the CRF is trained to predict BIO labels on the training part of the FTB, where only the target AQ and DD constructions have been annotated as MWEs, as described in §4.1. Feature selection is performed on development set of the MORPH dataset, in which each sentence contains exactly one occurrence to disambiguate (MWE or accidental co-occurrence).

5.1.1 First feature selection: coarse

As shown in the first row of Table 2, when we include all features described in §3 (ALL), we obtain an F_1 score of 75.47 for AQ and 69.70 for DD constructions. The following rows of the table show the results of a first ablation study, conducted to identify coarse groups of features that are not discriminant and may hurt performance.

When we ignore orthographic features (ALL - ORTH), all scores increase for DEV_{AQ} and DEV_{DD}, suggesting that MWE occurrences are not correlated with orthographic characteristics. F_1 also increases when we remove all surface-level wordform features, including single words and *n*-grams (represented by W). We hypothesize that lemmas and POS are more adequate, as they can reduce sparsity by conflating variations of the same lexeme, while wordforms only seem to introduce noise.

We then evaluate the removal of lexicon features (ALL - LF). At a first intuition, one would say that this information is important to our system, as it allows assigning O tags to conjunctions and prepositions that introduce verbal

Table 2: Ablation study results on the dev portion of the MORPH dataset focusing on AQ and DD expressions - impact of the removal of coarse-grained feature sets.

Feature set	DEV$_{AQ}$			DEV$_{DD}$		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**
ALL	89.55	65.22	75.47	92.00	56.10	69.70
ALL - ORTH	90.28	70.65	79.27	95.83	56.10	70.77
ALL - W	90.79	75.00	82.14	87.10	65.85	75.00
ALL - LF	91.18	67.39	77.50	88.89	58.54	70.59
ALL - $t_{\pm 2}$	87.67	69.57	77.58	88.00	53.66	66.67
ALL - $T_{i-1}T_iT_{i+1}$	87.84	70.65	78.31	91.67	53.66	67.69
ALL - $T_{i-1}T_i$	**93.55**	63.04	75.32	95.83	56.10	70.77
ALL - $T_{i-1}T_i$ - $T_{i-1}T_iT_{i+1}$	88.57	67.39	76.54	**96.00**	58.54	72.73
ALL - ORTH - W	90.24	**80.43**	**85.06**	87.10	65.85	75.00
ALL - ORTH - W - $t_{\pm 2}$ (REF$_1$)	89.74	76.09	82.35	85.29	**70.73**	**77.33**

complements. Surprisingly, though, the system performs better without them. We presume that this is a consequence of the sparsity of these features: since there are many features overall in the system, the CRF will naturally forgo LF features when they are present, rendering them superfluous to the system. These features will be analyzed individually later (see Table 4).

One might expect that single tokens located 2 words apart from the target token do not provide much useful information, so we evaluate the removal of the corresponding features (ALL - $t_{\pm 2}$). While this intuition may be true for DEV$_{AQ}$, it does not hold for DEV$_{DD}$. Next, we try to remove all trigrams, and then all trigam & bigram features at once. When we remove trigrams, F$_1$ decreases by 2.01 absolute points in DEV$_{DD}$ and increases by 2.84 absolute points in DEV$_{AQ}$. Bigrams are somehow included in trigrams, and their removal has little impact on the tagger's performance. When we remove bigram and trigram features altogether, scores are slightly better, even though a large amount of information is ignored. Since these results are inconclusive, we perform a more fine-grained selection considering specific *n*-grams in §5.1.2.

Finally, we try to remove several groups of features at the same time. When we remove both orthographic and wordform features, F$_1$ increases to 85.06 for DEV$_{AQ}$ and 75.00 for DEV$_{DD}$. When we also remove tokens located far away from the current one, performance increases for DEV$_{DD}$, but not for DEV$_{AQ}$. Un-

reported experiments have shown, however, that further feature selection also yields better results for DEV_{AQ} when we ignore $t_{\pm 2}$ features. Therefore, our reference (REF_1) for the fine-grained feature selections experiments will be this set of features, corresponding to the last row of Table 2.

5.1.2 Second feature selection: fine

Table 3 presents the results from fine-grained feature selection. In the first row of the table, we replicate the reference (REF_1) feature set defined above. In the second row, we try to remove the lexicon features (LF) once again. When they were removed in previous experiments, shown in Table 2, we had a gain in performance, suggesting that these features were superfluous. When we remove them from REF_1, however, the precision and recall observed for DEV_{DD} decrease by about 10 points. That is, the removal of LF yields a performance drop with respect to a relatively good model (REF_1), suggesting that these features are valuable after all. We hypothesise that LF can be better taken into account now that there are less noisy features overall in the whole system.

Table 3: Ablation study results on the dev portion of the MORPH dataset focusing on AQ and DD expressions - impact of the removal of fine-grained feature sets.

Feature set	DEV_{AQ}			DEV_{DD}		
	P	**R**	**F_1**	**P**	**R**	**F_1**
REF_1	89.74	76.09	82.35	**85.29**	70.73	77.33
REF_1 - LF	90.00	78.26	83.72	75.76	60.98	67.57
REF_1 - $w_{-1}w_0$	**90.54**	72.83	80.72	**85.29**	70.73	77.33
REF_1 - w_0w_{+1}	89.87	77.17	83.04	84.85	68.29	75.68
REF_1 - $w_0w_{+1}w_{+2}$ (BEST_1)	87.36	**82.61**	**84.92**	83.78	**75.61**	**79.49**

The last three rows of the table presents the results from attempts at removing individual *n*-gram features that we expected to be redundant or not highly informative. First, we consider the removal of two types of bigram features independently (towards the left and towards the right of the target word). We remove their wordforms, POS and lemmas. The results suggest that bigrams can be mildly useful, as their removal causes the most scores to drop.

In the last row of the table, we present the results from removing all trigram features of the form $w_0w_{+1}w_{+2}$. As a result, we can see that performance increases for both datasets. While trigram features could be potentially useful to recognise

longer expressions, we assume that the number of all possible trigrams is actually too large, making the feature values too sparse. In other words, a much larger annotated corpus would be required for trigram features to be effective. This is the best configuration obtained on the development datasets, and we will refer to it as BEST_1 in the next experiments.

Our last feature selection experiments consider the influence of lexicon features (LF) individually, as shown in Table 4. We observe that DEV is an important feature, because when we remove it, F_1 decreases by almost 7 absolute points on the DEV_{DD} set. The feature QUEV, however, seems less important, and its absence only slightly decreases the F_1 score on the DEV_{AQ} set. This is in line with what was observed by Nasr et al. (2015) for the whole dataset. In sum, these features seem to help, but we would expect the system to benefit more from them with a more sophisticated representation.

Table 4: Ablation study results on the dev portion of the MORPH dataset focusing on AQ and DD expressions - impact of the removal of lexicon features (LF).

Dataset	Feature set	P	R	F_1
DEV_{AQ}	BEST_1	87.36	82.61	84.92
	BEST_1-QUEV	91.25	79.35	84.88
DEV_{DD}	BEST_1	83.78	75.61	79.49
	BEST_1-DEV	77.78	68.29	72.73

The results obtained in this section focus on a limited number of very frequent expressions. Since our evaluation focuses on a small sample of 11 such MWEs only, it would be tempting to train one CRF model per target expression. However, there are a few more expressions with the same characteristics in French, and many of them share similar syntactic behaviour (e.g. conjunctions formed by an adverb and a relative conjunction). An approach with a dedicated model per expression would miss such regular syntactic behaviour (e.g. the fact that the surrounding POS are similar).

The experiments reported up to here show how it is possible to identify highly ambiguous (and frequent) expressions with a CRF, but they are hard to generalise to other MWE categories. Therefore, in the next sections, we evaluate our model on broader MWE categories such as nominal MWEs and general continuous MWEs (as defined in the FTB).

5.2 Experiments on nominal MWEs

We now focus on the identification of nominal MWEs in the FTB. As above, we separate our experiments in coarse-grained and fine-grained feature selection. In these experiments, the CRF was trained on the training part of the FTB where only nominal MWEs were tagged as B-I and all other words and MWEs were tagged as O. The feature selection experiments are performed on the development set of the FTB, also transformed in the same way. For the comparison with the state of the art, we report results for the test portion of FTB.

5.2.1 First feature selection: coarse

Table 5 presents the results obtained on FTB for different levels of feature selection. In the first row (ALL), we present the evaluation of all the features described in §3, except DEV and QUEV (only relevant to the previous experiments). We obtain a baseline with MWE-based F_1 score of 71.57%, and token-based score F_1 score of 73.85%.

Table 5: Ablation study results on FTB-dev focusing on nominal MWEs - impact of the removal of coarse-grained feature sets.

Feature set	MWE-based			Token-based		
	P	R	F_1	P	R	F_1
ALL	80.86	64.19	71.57	81.23	67.70	73.85
ALL - ORTH	81.85	64.78	72.32	82.16	68.02	74.43
ALL - W	80.41	64.78	71.75	80.95	68.44	74.17
ALL - AM	81.37	61.72	70.19	81.48	65.16	72.41
ALL - $t_{\pm 2}$	81.49	**65.84**	72.83	81.80	**69.50**	75.15
ALL - T_{+2}	80.96	65.51	72.48	81.18	69.08	74.64
ALL - $T_{i-1}T_i$	80.41	64.31	71.47	80.99	67.84	73.83
ALL - $T_{i-1}T_iT_{i+1}$ (REF_2)	81.61	**65.84**	**72.88**	82.05	69.40	**75.20**
ALL - $T_{i-1}T_iT_{i+1}$ - AM	81.69	63.60	71.52	82.09	67.28	73.95
ALL - ORTH - W - $t_{\pm 2}$ - $T_{i-1}T_iT_{i+1}$	79.59	63.37	70.56	81.00	67.88	73.86
ALL - ORTH - $T_{i-1}T_iT_{i+1}$	**82.51**	65.05	72.73	**82.74**	67.93	74.61

We consider the removal of the same groups of features that we removed on the AQ and DD experiments. We evaluate the independent removal of orthographic features, wordforms, association measures, $t_{\pm 2}$, t_{i+2}, bigrams and trigrams. We notice that all of these columns have better results than ALL, except

for the column where we removed the bigrams and the one in which we removed association measures. In particular, we notice that the absence of the AMs significantly hurts recall, which in turn has an impact on the F_1 score (−1.38% for the MWE-based measure and −1.41% for the token-based measure). This is the first clue that indicates the importance of these features.

We then evaluate the removal of different groups of features at the same time. We begin by deleting all of the previous groups, except for AMs and bigrams, which seemed to provide useful information above. Nevertheless, we did not obtain better results. We then tried to remove only the trigrams and the orthographic features. Results were slightly higher than ALL, but still remain worse than the results with only the trigrams removed. Finally, we decided to verify if the AM features are still relevant to obtain this performance. This was confirmed, as without the AM, the MWE-based F_1 score decreased by 1.36%, and the token-based F_1 score decreased by 1.25%. Overall, the highest results were obtained by removing only trigrams from ALL, and so we take this feature set as our new reference (REF_2).

5.2.2 Second feature selection: fine

Experiments above have shown that association measures (AM) are a vital component of our system. We proceed now to evaluate the importance of individual association measures towards the identification of nominal MWEs. The results are shown in Table 6. We consider the impact of the different AMs in two baseline configurations: all features (ALL), and the features of the reference only (REF_2). We then remove individual measures and evaluate the new feature set on FTB-dev.

We consider the removal of multiple combinations of features. In most cases, we notice a slight improvement in the results against ALL, but not when compared to the reference group . The removal of the DICE measure did improve the results in both cases, ALL and REF_2. Therefore, this configuration was chosen as the $BEST_2$ set of features. We then evaluated these $BEST_2$ features on the FTB-test dataset, obtaining a MWE-based F_1 score of 71.38%, and a token-based score of 73.43%. As a sanity check, we have also evaluated the system without AMs on FTB-test (ALL - AM). The $BEST_2$ system is significantly different from both ALL and ALL - AM on the test set. Moreover, the large margin between ALL - AM and the two other systems indicates that association measures do provide useful features for this task.

Table 6: Ablation study results on FTB-dev focusing on nominal MWEs - impact of the removal of fine-grained feature sets.

Feature set	MWE-based			Token-based		
	P	R	F_1	P	R	F_1
ALL	80.86	64.19	71.57	81.23	67.70	73.85
ALL - DICE	81.07	64.55	71.87	81.39	68.02	74.11
ALL - T-MEAS	81.07	64.55	71.87	81.40	68.07	74.14
ALL - PMI	81.26	63.84	71.50	81.51	67.33	73.74
ALL - MLE - LL	81.13	64.31	71.75	81.40	67.65	73.89
ALL - T-MEAS - DICE	80.76	64.78	71.90	81.23	68.30	74.20
ALL - MLE - LL - T-MEAS - DICE	81.72	63.72	71.61	81.58	67.05	73.61
REF_2	81.61	65.84	72.88	82.05	69.40	75.20
REF_2 - DICE ($BEST_2$)	**81.84**	65.84	72.98	82.33	**69.45**	**75.34**
REF_2 - T-MEAS	81.61	65.84	72.88	82.01	69.22	75.08
REF_2 - PMI	81.80	65.14	72.52	**82.36**	68.71	74.92
REF_2 - MLE - LL	81.67	65.61	72.76	82.03	69.08	75.00
REF_2 - T-MEAS - DICE	81.75	**65.96**	**73.01**	82.18	69.36	75.23
REF_2 - MLE - LL - T-MEAS - DICE	81.41	65.49	72.58	81.51	68.94	74.70
ALL (on FTB-test)	77.06	65.66	70.90	79.10	68.23	73.27
ALL - AM (on FTB-test)	76.96	61.81	68.56	78.94	64.91	71.24
$BEST_2$ (on FTB-test)	76.00	67.28	71.38	77.74	69.58	73.43

5.3 Experiments on general MWEs

We extend the experiments above to evaluate the feature sets against the whole FTB corpus, keeping all annotated MWEs in the training, development and test parts of the FTB. We would like to verify if our system is able to take into account the different MWE categories at the same time. This time, we only present coarse-grained feature selection (Table 7), since unreported fine-grained feature selection resulted in similar findings as in experiments focusing on nominal MWEs.

The first row in the table (ALL) presents the evaluation of all features described in §3. The prediction of general MWEs with ALL features yields a MWE-based F_1 score of 78.89% and a token-based F_1 score of 81.61%. We then consider what happens when one removes the same groups of features as in the previous sections. This time the results are quite different: all of these tests have worse results than ALL, except when we remove t_{+2} features. In some unreported experiments,

Table 7: Ablation study results on FTB-dev focusing on general MWEs - impact of the removal of feature sets.

Feature set	MWE-based			Token-based		
	P	R	F_1	P	R	F_1
ALL	**85.60**	73.16	78.89	**87.32**	76.60	81.61
ALL - ORTH	85.09	72.97	78.57	86.97	76.56	81.44
ALL - W	83.96	72.59	77.86	86.13	76.37	80.96
ALL - AM	85.11	72.78	78.46	86.89	76.33	81.27
ALL - $t_{\pm 2}$	84.03	72.45	77.81	86.57	76.94	81.47
ALL - t_{+2}	85.50	73.68	79.15	87.19	77.21	81.90
ALL - $T_{i-1}T_i$	84.36	71.75	77.54	86.61	75.47	80.66
ALL - $T_{i-1}T_iT_{i+1}$	84.78	73.07	78.49	86.39	76.31	81.04
ALL - T_{+2} - ORTH (REF_3)	85.52	**73.82**	**79.24**	87.30	**77.35**	**82.03**
REF_3 - AM	85.37	72.69	78.52	87.08	76.33	81.35
REF_3 - T-MEAS ($BEST_3$)	85.62	**73.87**	**79.31**	87.40	**77.43**	**82.11**
ALL (on FTB-test)	83.80	74.51	78.88	86.58	78.23	82.19
ALL - AM (on FTB-test)	84.19	73.52	78.49	86.90	77.30	81.82
$BEST_3$ (on FTB-test)	84.03	74.71	79.10	86.72	78.47	82.39

we have tried to remove other groups of features along with t_{+2}. We found that removing orthographic features along with t_{+2} increased the results more than only removing t_{+2} features. This group of features will be our new reference from now on (REF_3). Once again, we tried to remove AMs from the reference to verify their impact. Here again, we notice that the removal of these features decreases the overall performance scores, even if their impact is weaker than it was in the case of nominal MWEs. Unreported experiments have shown that we obtain better results when we ignore the T-MEAS feature ($BEST_3$).

Then, we applied the feature group $BEST_3$ on the FTB-test dataset, and we obtained a MWE-based F_1 score of 79.10%, and a token-based score of 82.39%. For the feature selection experiments on the test part of the FTB (both nominal and general MWEs), we calculated the p-value of the difference between the configuration called BEST and the one called ALL, using approximate randomisation with stratified shuffling. None of the observed differences was considered statistically significant with $\alpha = 0.05$.

5.4 Comparison with state of the art

We now compare the highest-scoring reference results with the state of the art. We begin by evaluating the identification of *DD* and *AQ* constructions, and then proceed to evaluate more generally the quality of our reference system for general MWE identification. The comparisons presented here focus on MWE identification only, and our model takes gold POS and lemma information as input (except on the MORPH dataset). On the other hand, some of the works mentioned in our comparisons also predict POS and/or syntactic structure, which makes the task considerably harder. Therefore, results presented here should be taken as an indication of our position within the current landscape of MWE identification, rather than as a demonstration of our model's superiority.

5.4.1 AQ and DD constructions

We report the performance of MWE identification on the full MORPH dataset, split in two parts: sentences containing AQ constructions (FULL_{AQ}) and sentences containing DD constructions (FULL_{DD}). The use of the full datasets is not ideal, given that we performed feature selection on part of these sentences, but it allows a direct comparison with related work.

Table 8 presents a comparison between the best system score obtained after feature selection (BEST_1) and the results reported by Nasr et al. (2015). We include two versions of the latter system, since they also distinguish their results based on the presence of lexicon features (LF) coming from DicoValence.

Table 8: Comparison with baseline and state of the art of AQ and DD identification on the full MORPH dataset.

System	FULL_{AQ}			FULL_{DD}		
	P	**R**	**F_1**	**P**	**R**	**F_1**
Baseline	56.08	100.00	71.86	34.55	100.00	51.35
Nasr et al. (2015)-LF	88.71	82.03	85.24	77.00	73.09	75.00
Nasr et al. (2015)+LF	91.57	81.79	86.41	86.70	82.74	84.67
BEST_1	91.08	78.31	84.21	79.14	74.37	76.68

We additionally report results for a simple baseline:

1. We extract a list of all pairs of contiguous AQ and DD from the FTB-train.
2. We calculate the proportion of cases in which they were annotated as MWEs (B-I tags) with respect to all of their occurrences.
3. We keep in the list only those constructions which were annotated as MWE at least 50% of the time.
4. We systematically annotate these constructions as MWEs (B-I) in all sentences of the MORPH dataset, regardless of their context.

Table 8 shows that this baseline reaches 100% recall, covering all target constructions, but precision is very low, as contextual information is not taken into account during identification. Our BEST_1 system can identify the target ambiguous MWEs much better than the baseline for both FULL_{AQ} and FULL_{DD}.

For some constructions, we do obtain results that are close to those obtained by the parsers (see Table 9 for more details). For FULL_{AQ}, our BEST_1 system obtains an F_1 score that is 1.2 absolute points lower than the best parser. For FULL_{DD}, however, our best system, which includes lexicon features (LF), is comparable with a parser without lexicon features. When the parser has access to the lexicon, it beats our system by a significant margin of 7.99 points, indicating that the accurate disambiguation of DD constructions indeed requires syntax-based methods rather than sequence taggers. These results contradict our hypothesis that sequence models can deal with continuous constructions with a performance equivalent to parsing-based approaches. While this may be true for non-ambiguous expressions, parsing-based methods are superior for AQ and DD constructions, given that they were trained on a full treebank, have access to more sophisticated models of a sentence's syntax, and handle long-distance relations and grammatical information.

Despite the different results obtained depending on the nature of the target constructions, the results are encouraging, as they prove the feasibility of applying sequence taggers for the identification of highly ambiguous MWEs. Our method has mainly two advantages over parsing-based MWE identification: (a) it is fast and only requires a couple of minutes on a desktop computer to be trained; and (b) it does not require the existence of a treebank annotated with MWEs.

Table 9 shows the detailed scores for each expression in the MORPH dataset. We notice that some expressions seem to be particularly difficult to identify, especially if we look at precision, whereas for others we obtain scores well above 90%. When we compare our results to those reported by Nasr et al. (2015), we can see that they are similar to ours: *ainsi* 'likewise', *alors* 'then' and *bien* 'well'

have F_1 higher than 90%, while *autant* 'as much' and *tant* 'while' have a score lower than 80%. The AQ constructions with *encore* 'still' and *maintenant* 'now' are the only ones which behave differently: our system is better for *encore* 'still', but worse for *maintenant* 'now'. Likewise, for DD expressions, our system obtains a performance that is close to their system without lexicon features (LF), but considerably worse than their system including LFs for three out of 4 expressions. Both approaches are more efficient in identifying the plural article *de les* 'of the.PL' than the partitive constructions.

Table 9: Performance of the $BEST_1$ configuration broken down by expression, along with the results for the best model of Nasr et al. (2015) (with LF).

	$BEST_1$ system			Nasr et al. (2015)		
Expression	**P**	**R**	**F_1**	**P**	**R**	**F_1**
ainsi que	94.44	93.15	93.79	95.94	89.87	92.81
alors que	84.00	97.67	90.32	93.81	93.81	93.81
autant que	93.48	51.81	66.67	86.66	70.65	77.84
bien que	100.00	91.43	95.52	91.66	99.18	90.41
encore que	76.19	94.12	84.21	92.85	65.00	76.47
maintenant que	97.62	64.06	77.36	90.91	74.62	81.96
tant que	100.00	60.00	75.00	82.35	70.00	75.67
de le	78.05	71.11	74.42	85.41	91.11	88.17
de la	67.74	72.41	70.00	81.25	89.65	85.24
de les	92.41	71.57	80.66	98.70	76.00	85.87
de l'	61.11	95.65	74.58	64.51	86.95	74.07

5.4.2 General MWEs

We now compare our system with two baselines and with the system proposed in Le Roux et al. (2014).[17] $Baseline_1$ consists in identifying as MWE every continuous occurrence of tokens that has been seen as an MWE in the training corpus. For example, the MWE *bien sûr* (lit. *well sure*) 'of course' can be seen in the training corpus, and so every occurrence of this expression was predicted as an MWE

[17] The comparison with Le Roux et al. (2014) is not ideal, since we predict MWEs with the help of gold POS and lemmas, whereas they try to predict both POS and MWEs. However, we could not find a fully comparable evaluation in the literature.

for the test corpus. Baseline_2 filters the list of MWEs seen in the training corpus, so that only the expressions which had been annotated more than 40% of the time are predicted as MWEs. For example, the expression *d'un côté* (lit. *of a side*) 'on the one hand' is not predicted as MWE, as it was only annotated in 38% of its occurrences in the training corpus. The baselines were directly inspired by a predictive model applied to the English language in a similar task, where a threshold of 40% was found to yield the best results (Cordeiro et al. 2016). The applied threshold in Baseline_2 only eliminates 6.46% of the MWEs from the list, but it contributes to an increase of 20–30 points in precision without impacting the recall.

Our system (BEST_3 configuration) is more accurate than the baselines, both in terms of precision and recall. It also has a higher precision than the approach proposed by Le Roux et al. (2014), but the recall is considerably worse (9.48% less than their system). This means that our system misses more expressions, even if its predictions have higher precision. This could be partly explained by the fact that they employed dictionaries, and have access to more expressions that our system has never seen and could not predict. Nonetheless, our results are sufficiently close and represent a decent alternative if high-quality external resources are not available.

Table 10: Comparison with baseline and state of the art of general MWE identification on FTB-test.

System	MWE-based			Token-based		
	P	R	F_1	P	R	F_1
Baseline_1	52.93	66.20	58.82	62.70	69.73	66.03
Baseline_2	82.76	69.36	75.47	84.80	69.62	76.46
BEST_3 configuration	84.03	74.71	79.10	86.72	78.47	82.39
Le Roux et al. (2014)	80.76	84.19	82.44	—	—	—

5.5 Analysis of results

The performance of our CRF identification model depends on the characteristics of the identified MWEs and of the training and test corpora. Therefore, we have performed a detailed analysis of its performance focusing on a subset of the test corpus. We focus on two phenomena: variants and unseen MWEs. We define a **variant** as an MWE whose lemmatised form occurs both in the training and in

the test corpus, but whose surface form in the test corpus is different from all of its surface forms in the training corpus. We define an **unseen** MWE as one whose lemmatised form occurs in the test corpus but never (under any surface form) in the training corpus. MWEs which have identical occurrences in the training and test corpora will be referred to as **seen** MWEs.

Both variants and unseen MWEs are harder to identify than seen MWEs. Nonetheless, we hypothesise that our model is able to recognise variants correctly, since its features are based on lemmas. On the other hand, we expect that unseen MWEs cannot be easily predicted given that our system is absed on categorical features and does not have access to much information about an expression that has never been seen in the training corpus, except for its association measures in a large unannotated corpus. To verify these hypotheses, we create sub-corpora of FTB-test, where the density of variants and unseen MWEs is higher than in the full FTB-test corpus. In these experiments, the model is not newly trained, but the $BEST_2$ and $BEST_3$ models are applied to different sub-corpora with a high density of variant/unseen MWEs.

The evaluation measures reported in our experiments (§4.4) consider the best bijection between predicted and gold MWEs. Therefore, we cannot simply remove seen MWEs from the test set, since they will be predicted anyway, artificially penalising precision. Therefore, instead of completely removing seen MWEs, we remove sentences that contain only seen MWEs and keep sentences that contain (a) at least one variant MWE or (b) at least one unseen MWE.

Table 11: Results of $BEST_2$ (nominal MWEs) and $BEST_3$ (general MWEs) on FTB-test, on sub-corpus containing unseen variants of a seen MWEs, and on sub-corpus containing unseen MWEs. Columns %var and %unseen show the proportion of variants/unseen MWEs in each sub-corpus.

Feature set	%var	%unseen	MWE-based			Token-based		
			P	R	F_1	P	R	F_1
Nominal full	5%	28%	76.00	67.28	71.38	77.74	69.58	73.43
Nom. variants	65%	N/A	86.42	63.64	73.30	85.84	66.20	74.75
Nom. unseen	N/A	72%	82.01	42.70	56.16	87.78	46.75	61.01
General full	5%	23%	84.03	74.71	79.10	86.72	78.47	82.39
Gen. variants	32%	N/A	88.91	69.44	77.98	92.77	74.05	82.36
Gen. unseen	N/A	51%	86.94	65.22	74.25	90.40	69.14	78.35

Table 11 presents the performance of the $BEST_2$ configuration for nominal MWEs (first row) and $BEST_3$ configuration for general MWEs (fourth row) on the full FTB-test corpus. For each expression (nominal and general), we also present the results for the sub-corpus containing a higher density of variants and of unseen MWEs. The numbers in columns %var and %unseen indicate the proportion of variant/unseen MWEs in each sub-corpus. Notice that, in the case of general MWEs, these proportions are quite low (32% and 51%), indicating that sentences containing variant and unseen general MWEs often contain seen ones too. When focusing on variants (Nom. variants and Gen. variants sub-corpora), the proportion of unseen MWEs is very small and not relevant (N/A), and vice-versa.

If we focus on variants, we can observe relatively stable results with respect to the full FTB-test corpus. For nominal MWEs, precision increases by 8-10%, whereas recall decreases by about 3% for both MWE-based and token-based measures. Results for general MWEs follow a similar pattern: around 4-6% improvement in precision at the cost of around 4-5% decrease in recall. The precision of general MWE identification in the variants sub-corpus is particularly impressive, reaching 92.77%.

The variants sub-corpora contain less unseen MWEs than the full FTB-test corpus, so the predicted MWEs are more reliable (better precision), showing that our model is robust to morphological variability. On the other hand, the fact that recall drops indicates that it is indeed slightly harder to recognise variants of MWEs than those seen identically in training and test corpora. In short, we infer that variants can be correctly handled and identified by our model, provided that a good lemmatiser is available (results presented here are based on gold lemmas, their substitution by predicted lemmas should be studied in the future).

On the other hand, predicting unseen MWEs is considerably harder. Recall drops drastically by about 23-25% for nominal MWEs and by about 9% for general MWEs, and the improvements in precision do not compensate for this, yielding much lower F-measure values, specially for nominal MWEs where the concentration of unseen MWEs in the sub-corpus is higher (72%). The improvements in precision are probably due to the fact that some seen and variant MWEs are still present in the sub-corpora. AMs could also have some predictive power to identify unseen MWEs, and we intend to verify their contribution for unseen MWE identification in the future. These results show that our model is limited in the identification of unseen MWEs, and can probably only identify some of those that appear in the AM lexicons.

6 Conclusions and future work

We have described and evaluated a simple and fast CRF tagger that is able to identify several categories of continuous multiword expressions in French. We have reported feature selection studies and shown that, for AQ constructions and for general MWEs, our results are almost as good as those obtained by parsers, even though we do not rely on syntactic trees. This was not true for DD constructions, though, which seem to require parsing-based methods to be properly analysed. Based on these results, we conclude that, when treebanks are not available, sequence models such as CRFs can obtain reasonably good results in the identification of continuous MWEs. On the other hand, when MWE-annotated treebanks exist, parsing-based models seem to obtain better results, especially for expressions whose high ambiguity requires syntax to be resolved.

An interesting direction of research would be to study the interplay between automatic POS tagging and MWE identification. We recall that our results were obtained with an off-the-shelf POS tagger and lemmatizer. Potentially, performing both tasks jointly could help obtaining more precise results (Constant & Sigogne 2011). Moreover, we could explore CRFs' ability to work with lattices in order to pre-select the most plausible MWE identification (and POS tagging) solutions, and then feed them into a parser which would take the final decision.

Another idea for future work would be an investigation of the features themselves. For example, in this work, we were not fully satisfied with the quality of the representation of lexical features. We would like to investigate the reason why lexical features were not always useful for the task of MWE identification, which could be done by performing an error analysis on the current systems. Another interesting question is whether annotated corpora are at all necessary: could hand-crafted and/or automatically built lexicons be employed to identify MWEs in context in a fully unsupervised way?

While these experiments shed some light on the nature of MWEs in French, the feature selection methodology is highly empirical and cannot be easily adapted to other contexts. Therefore, we would like to experiment different techniques for generic automatic feature selection and classifier tuning (Ekbal & Saha 2012). This could be performed on a small development set, and would ease the adaptation of the tagger to other contexts.

Finally, we would like to experiment with other sequence tagging models such as recurrent neural networks. In theory, such models are very efficient to perform feature selection and can also deal with continuous word representations able to encode semantic information. Moreover, distributed word representations could

be helpful in building cross-lingual MWE identification systems.

Acknowledgments

This work has been funded by projects PARSEME (Cost Action IC1207), PARSEME-FR (ANR-14-CERA-0001), and AIM-WEST (FAPERGS-INRIA 1706-2551/13-7). We also thank the anonymous reviewers for their valuable suggestions.

Abbreviations

AQ	adverb+*que*	DD	*de*+determiner
AM	association measure	FTB	French Treebank
BIO	begin-inside-outside	LF	lexicon feature
CRF	conditional random field	MWE	multiword expression

References

Abeillé, Anne, Lionel Clément & François Toussenel. 2003. Building a treebank for French. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, 165–187. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Baroni, Marco & Silvia Bernardini (eds.). 2006. *Wacky! Working papers on the web as corpus*. Bologna, Italy: GEDIT.

Boukobza, Ram & Ari Rappoport. 2009. Multi-word expression identification using sentence surface features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 468–477. August 6-7, 2009.

Buchholz, Sabine & Erwin Marsi. 2006. CoNLL-x shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning* (CoNLL-X '06), 149–164. http://dl.acm.org/citation.cfm?id=1596276.1596305.

Candito, Marie & Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 743–753. Association for Computational Linguistics. http://www.aclweb.org/anthology/P14-1070.

Carpuat, Marine & Mona Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in Statistical Machine Translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 242–245. Association for Computational Linguistics. http://www.aclweb.org/anthology/N10-1029.

Constant, Matthieu & Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 161–171. Association for Computational Linguistics. http://www.aclweb.org/anthology/P16-1016.

Constant, Matthieu & Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the ALC Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (MWE 2011), 49–56. Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-0809.

Cordeiro, Silvio, Carlos Ramisch & Aline Villavicencio. 2016. UFRGS & LIF at SemEval-2016 task 10: Rule-based MWE identification and predominant-supersense tagging. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 910–917. Association for Computational Linguistics. http://www.aclweb.org/anthology/S16-1140.

Diab, Mona & Pravin Bhutada. 2009. Verb noun construction MWE token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, 17–22. Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W09/W09-2903.

Ekbal, Asif & Sriparna Saha. 2012. Multiobjective optimization for classifier ensemble and feature selection: An application to named entity recognition. *International Journal on Document Analysis and Recognition (IJDAR)* 15(2). 143–166. DOI:10.1007/s10032-011-0155-7

Fazly, Afsaneh, Paul Cook & Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1). 61–103. http://aclweb.org/anthology/J09-1005.

Finlayson, Mark & Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the ACL Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (MWE '11), 20–24. Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-0805.

Green, Spence, Marie-Catherine de Marneffe & Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics* 39(1). 195–227. DOI:10.1162/COLI_a_00139

Lafferty, John D., Andrew McCallum & Fernando C. N. Pereira. 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. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=645530.655813.

Laporte, Éric, Takuya Nakamura & Stavroula Voyatzi. 2008. A French corpus annotated for multiword expressions with adverbial function. In *Proceedings of the 2nd Linguistic Annotation Workshop*, 48–51. https://halshs.archives-ouvertes.fr/halshs-00286541.

Le Roux, Joseph, Antoine Rozenknop & Matthieu Constant. 2014. Syntactic parsing and compound recognition via dual decomposition: Application to French. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers* (COLING 2014), 1875–1885. Dublin, Ireland: Dublin City University & Association for Computational Linguistics. http://www.aclweb.org/anthology/C14-1177.

Nasr, Alexis, Carlos Ramisch, José Deulofeu & André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1116–1126. Association for Computational Linguistics. http://www.aclweb.org/anthology/P15-1108.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 1659–1666. European Language Resources Association (ELRA). 23-28 May, 2016.

Nivre, Joakim & Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Workshop on methodologies and evaluation of multiword units in real-world applications* (MEMURA 2004), 39–46. http://stp.lingfil.uu.se/~nivre/docs/mwu.pdf.

Okazaki, Naoaki. 2007. *CRFsuite: A fast implementation of conditional random fields (CRFs)*. http://www.chokkan.org/software/crfsuite/.

Ramisch, Carlos. 2014. *Multiword expressions acquisition: A generic and open framework* (Theory and Applications of Natural Language Processing XIV). Cham, Switzerland: Springer. 230. http://link.springer.com/book/10.1007%2F978-3-319-09207-2.

Ramshaw, Lance & Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *3rd Workshop on Very Large Corpora*, 82–94. http://aclweb.org/anthology/W95-0107.

Riedl, Martin & Chris Biemann. 2016. Impact of MWE resources on multiword recognition. In *Proceedings of the 12th Workshop on Multiword Expressions* (MWE '16), 107–111. Association for Computational Linguistics. http://anthology.aclweb.org/W16-1816.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI:10.5281/zenodo.1469555

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 31–47. Association for Computational Linguistics. DOI:10.18653/v1/W17-1704

Schneider, Nathan, Emily Danchik, Chris Dyer & Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics* 2(1). 193–206. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/281.

Schneider, Nathan, Spencer Onuffer, Nora Kazour, Nora Emily Danchik, Michael T. Mordowanec, Henrietta Conrad & Smith Noah A. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014), 455–461. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/521_Paper.pdf.

Scholivet, Manon & Carlos Ramisch. 2017. Identification of ambiguous multiword expressions using sequence models and lexical resources. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 167–175. Association for Computational Linguistics. http://www.aclweb.org/anthology/W17-1723.

Seddah, Djamé, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska & Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 Shared Task: a cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically-Rich Languages*, 146–182. Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-4917.

Shigeto, Yutaro, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung & Yuji Matsumoto. 2013. Construction of English MWE dictionary and its application to POS tagging. In *Proceedings of the 9th Workshop on Multiword Expressions* (MWE '13), 139–144. Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-1021.

Silberztein, Max, Tamás Váradi & Marko Tadić. 2012. Open source multi-platform NooJ for NLP. In *Proceedings of the 24th International Conference on Computational Linguistics: Demonstration Papers* (COLING-12), 401–408. The Coling 2012 Organizing Committee. http://www.aclweb.org/anthology/C12-3050.

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

Vincze, Veronika, István Nagy T. & Gábor Berend. 2011. Detecting noun compounds and light verb constructions: A contrastive study. In *Proceedings of the ALC Workshop on Multiword Expressions: From Parsing and Generation to the*

Real World (MWE '11), 116–121. Portland, OR, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-0817.

Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, ÇağrÄ± Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj & Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, 1–19. Vancouver, Canada: Association for Computational Linguistics. http://www.aclweb.org/anthology/K17-3001.