A Multiword Expression Lexicon Formalism Optimised for Observational Adequacy

Anonymous ACL submission

Abstract

Past research advocates that, in order to handle the unpredictable nature of multiword expressions (MWEs), their identification should be assisted with lexicons. The choice of the format for such lexicons, however, is far from obvious. We propose the first – to our knowledge – method to quantitatively evaluate MWE lexicon formalisms based on the notion of observational adequacy. We apply it to derive a simple yet adequate MWE-lexicon formalism, dubbed λ-CSS, based on syntactic dependencies. It proves competitive with lexicons based on sequential representation of MWEs, as well as with a state-of-the art MWE identifier.

1 Introduction

Multiword expressions (MWEs), such as ***by and large***, ***carbon footprint*** or *to* ***pull*** *one's* ***leg*** 'to tease someone', exhibit irregularities which are challenging for text processing. Most notably, their meaning cannot be straightforwardly deduced from the meanings of their components, which is an obstacle for semantically-oriented applications. To help such applications process MWEs correctly, one solution is to pre-identify MWEs in text, so as to later apply dedicated procedures to them.

Recognizing MWEs occurrences in texts (henceforth referred to as MWE *identification*) is, according to Constant et al. (2017), one of the two main subtasks of MWE processing (the other being MWE *discovery*, the task of generating sets of MWEs) and still represents quite a challenge despite having been the focus of many works. Notably, PARSEME shared tasks on identification of verbal MWEs (Savary et al., 2017; Ramisch et al., 2018, 2020) have provided a controlled environment and focused challenges for MWE identification. Each edition of the task trying to put in focus those facets of the identification task which are the hardest.

One thing that PARSEME shared tasks definitely highlighted is that identification of MWEs unseen during training proves to be significantly harder than identification of seen MWEs. This can be seen in the results of editions 1.1 and 1.2 of the shared tasks when comparing the scores of various identifiers on seen vs unseen MWEs. The difficulty of identifying unseen MWE should not come as a surprise as this task can be seen as presenting the challenges of both identification and discovery.

Seeing this discrepancy between identification of seen and unseen MWEs, Savary et al. (2019b) argue that the use of MWE lexicons is key to high-quality MWE identification. Thus, shifting the burden of unseen MWEs on discovery and using lexicon as the interface between discovery and identification.

In accordance with this argument, this paper investigates MWE-lexicon formalisms, how they can be compared and introduce one such MWE-lexicon formalisms.

2 Multiword Expression

We abide by PARSEME's definition of a MWE (Savary et al., 2018a), adapted from (Baldwin and Kim, 2010), as a (continuous or discontinuous) sequence of words, at least two of which are lexicalized (always realised by the same lexemes), which displays some degree of lexical, morphological, syntactic and/or semantic idiosyncrasy.

MWEs happen to present quite a few interesting properties. Of all the properties listed by (Savary et al., 2018a; Baldwin and Kim, 2010; Constant et al., 2017) we will only mention the following 3 for the impact they have on how MWEs can and should be represented and what MWE-lexicons need to accomplish.

Variability MWEs can appear under a variety of *forms* depending on the morphosyntactic context in which they occur (e.g. ***I pay*** *him a* ***visit*** / *The* ***visits***

*she **pays** me*), their components can be found in different orders, forms, or even differently syntactically related. This makes simple representations such as sequences of forms insufficiently descriptive and pushes us to more complex representations capturing all the forms under which a MWE could appear.

Discontinuity Discontinuity can be seen as a form of variability where component words of a MWE are not adjacent to one another but separated by a word or group of words named the *insertion*. We define two types of discontinuity: *linear discontinuity* where the component words of the MWE are not next to each other in the sentence (e.g. ***pay** someone a **visit***, where '*someone a*' is the insertion between '*pay*' and '*visit*') ; and *syntactic discontinuity* where a component of the MWE is not directly related by a syntactic dependency to any other component of the MWE (e.g. figure 1 where '*wanted*' is the insertion between '*visit*' and '*pay*'[1]).

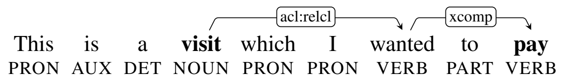


Figure 1: Syntactic discontinuity

Not all MWEs can be discontinued and anything cannot be inserted between MWE components. What can and cannot be inserted in a MWE depends on the MWE and should be described for a MWE representation to be complete.

Literal-idiomatic ambiguity While MWEs are defined as groups of words displaying some form of idiosyncrasy, sometimes the very group of words composing a given MWE can appear in a sentence without displaying any idiosyncrasy. In this case, we say that the occurrence is non-idiomatic (e.g. *I paid them a visit to the museum*) as opposed to idiomatic occurrences (e.g. *I **paid** them a **visit** at the hospital*). This very fact is the reason behind the need for MWE identification. Non-idiomatic occurrences can further be divided into literal and coincidental occurrence, (sec. 6.1), the former denoted by wavy underline, the latter by dashed underline.

[1]All syntactic analyses in this paper follow the Universal Dependencies formalism and are generated according to UDPipe 2.6 (english-ewt-ud-2.6-200830).

3 MWE-lexicon Formalisms

Numerous MWE-lexicons (MWE-Ls) have been put forward in the past. Each of them follows a MWE-L formalism, henceforth simply called *formalism*, which determines what kind of information can be stored and how. Unfortunately, formalisms are often only an afterthought, as a result, works on MWE-Ls often focus on MWE extraction and only touch upon how MWEs are represented in the MWE-L. Nevertheless, formalisms can be loosely categorized based on the kind of representation used to store their lexical entries.

Probably one of the biggest categories of MWE-L formalisms would be those based on phrase grammars. We further divide this category into two smaller: (i) formalisms based on list-like or regex-like structures (Breidt et al., 1996; Alegria et al., 2004; Oflazer et al., 2004; Sailer and Trawiński, 2006; Spina, 2010; Quochi et al., 2012; Al-Sabbagh et al., 2014; Al-Haj et al., 2014; Walsh et al., 2019), component words are listed in the order in which they can appear and discontinuities are most often denoted by special symbols imposing constraints on the types of insertions allowed (either by limiting the number of insertions or the words which can be inserted); (ii) formalisms based on more expressive phrase grammars (CFGs, TAGs, LFGs, HPSGs, ...) (Grégoire, 2010; Przepiórkowski et al., 2017; Savary et al., 2018b; Dyvik et al., 2019), here component words are usually terminals appearing in grammar rules, and discontinuities are denoted by non-terminals.

Less frequent are dependency-based formalisms, like PDT-Dep (Pecina, 2008), in which only bigrams of syntactically dependent words are considered.[2]

Other popular categories are driven by semantics (Villavicencio et al., 2004; Borin et al., 2013) or relational databases (Vondřička, 2019).

These categories do not cover all possibilities and whether a specific MWE-L belongs to one category over another could be disputed.

4 Evaluation of MWE-lexicon Formalisms

Seeing all these different MWE-Ls and formalisms, one might ask which one is best in order to assist MWE Identification. One part of

[2]Some other MWE-Ls encode syntactic dependencies as auxiliary data.

the answer comes us from Savary et al. (2019b) which recommend that MWE-Ls aiming to assist MWE identification should be distributed in extensional and standard format, and that the lemmas and POS of MWEs' component words, as well as the least syntactically marked dependency structure and some other morphosyntactic variants judged relevant should be accessible. The other part of the answer comes us from looking at how MWE-Ls have been compared up until now.

To our knowledge, there are only few studies comparing MWE-Ls. PARSEME's survey (Losnegaard et al., 2016) references more than fifty MWE lexicons and lists in dozens of languages, and compares their accessibility, languages represented, size, and capacity to encode discontinuous MWEs. Savary (2008) compares a few lexicons of continuous MWEs showing how their formalisms allow one to encode salient MWE properties.

Such comparisons are relevant to our work but are mostly qualitative in nature. Formalisms are compared on what they can and cannot express and quantitative comparisons are almost exclusively reserved to compare MWE-Ls' sizes. To our knowledge, MWE-L formalisms themselves have not yet been compared quantitatively. This brings us to the question of how MWE-L formalisms can be quantitatively compared.

5 Adequacy

In order to evaluate MWE-Ls, we borrow the notion of adequacy, first defined for grammars (Chomsky, 1965) then adapted to lexicons (Jackendoff, 1975). Adequacy can be divided into three levels, which, in the context of MWE-Ls, can be summarized as follows: (i) *observational adequacy*, which evaluates the coverage of MWE observations accounted for in a MWE-L; (ii) *descriptive adequacy*, which estimates whether a MWE-L accurately and exhaustively describes all the properties of the covered MWEs; (iii) *explanatory adequacy*, relating to how well a MWE-L explains the reasons behind MWE behavior.

In this paper, we focus on observational adequacy (OA) since it is the easiest to quantify and strongly relevant to MWE identification. This choice coincides with recommendations by Savary et al. (2019b), who advocate that MWE identification be assisted by MWE-Ls which use a relatively simple dependency-based formalism.

Perfect OA can more accurately be defined as the MWE-L accounting for all possible observations of MWEs and only those. In other words, the MWE-L must contain entries which match all possible MWEs observations (here understood as surface forms). It follows that OA can be measured from the standpoint of generation or parsing. More precisely, MWE-Ls are evaluated on their capacity to either generate all possible MWE forms, or to recognize all MWE forms encountered in text.

OA can be measured in a multitude of ways. In this study we keep ourselves to precision and recall, which measure the proportion of actual MWE observations in those matched by the lexicon and in those existing in text, respectively. Note that the measure of precision from a generative standpoint causes issues, since MWE occurrences can be literal (cf. Sec. 6.1).

Finally, in order for OA to be applicable to formalisms, we propose that they should be evaluated in conjunction with an instantiation method and corpus. Thus, two formalisms can be compared provided that their respective MWE-Ls are instantiated on the same data, in similar fashion, and that OA is measured on the same corpus.

6 λ-CSS Lexicons

Now that we have suggested criteria for an optimal format of MWE-Ls, let us see how this format could look like.

6.1 Literal occurrences

Savary et al. (2019a) ask what exactly is a literal occurrence of a MWE and what distinguishes it from an idiomatic or coincidental occurrence. Roughly, when all the lexemes of a MWE appear in a sentence and they together display some form of idiosyncrasy, then we talk of an *idiomatic occurrence* of the MWE. Whereas when they display no idiosyncrasy, we talk of a *non-idiomatic occurrence* of the MWE. Non-idiomatic occurrences are furthermore divided into *literal occurrences* and *coincidental occurrences*. Savary et al. (2019a) define the former as an occurrence which appears in a syntactic configuration in which could have been idiomatic. The latter is then simply defined as a non-idiomatic occurrence which is not literal.

In the following: in **bold** in (1) an idiomatic occurrence, in wavy underline in (2) a literal occurrence, and in dashed underline in (3) a coincidental occurrence :

(1) I **paid** them a **visit** at the hospital 'I visited

them at the hospital‘

(2) I paid them a visit to the museum

(3) I paid for a visit of the museum

In order to judge whether a non-idiomatic occurrence is in a syntactic configuration that could be idiomatic, it is compared to syntactic configurations of known idiomatic occurrences. To compare syntactic configurations, Savary et al. define the *Coarse Syntactic Structure* (*CSS*).

6.2 Coarse Syntactic Structure (CSS)

A CSS can be seen as a simplification of the dependency tree of a given MWE occurrence. More precisely, given a set of words σ and a sentence S, a CSS is the minimal connected dependency tree covering σ in S, where a word is either represented by a node containing its lemma and part of speech, if it is in σ, or by a dummy node otherwise. Nodes are connected by their relational dependencies.

For instance, for sentence (1), figure 2 shows its dependency tree, where word forms are replaced by their lemmas and parts of speech (POS). Then, figure 4a is the CSS of the MWE ***paid visit***, and figure 4b the CSS of the MWE with syntactic discontinuities from figure 3.

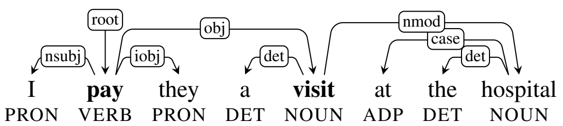


Figure 2: A dependency graph.

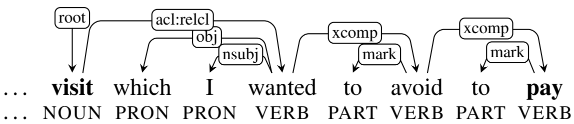
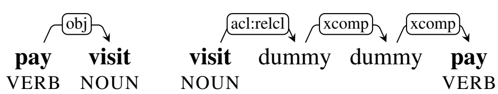


Figure 3: A dependency tree with syntactic discontinuities



(a) (b)

Figure 4: Coarse syntactic structure Figures 2 and 3

CSSs were originally designed in order to put an applicable definition to the notion of a literal occurrence of a MWE. However, since literal occurrences of MWE are relatively infrequent (Savary et al., 2019a), we argue that CSSs could be used as the basis of MWE-L formalisms with hopefully great observational adequacy.

MWE-Ls following such a formalism would simply consist in a set of CSSs of MWE occurrences. We will however first question the relevancy of component words being represented by their lemmas and POS and not some other features. Lemmas and POS do provide an approximation of lexemes, which lets CSSs do what they were designed to do (help approximate our intuitive notion of literal occurrence). We however would like for our lexicon to be as observationally adequate as possible, therefore we will wonder if representing MWEs by a different set of features would be beneficial.

For this reason, we propose a generalisation of CSSs, dubbed λ-CSS, where λ is the set of features used to describe MWEs.

6.3 λ-CSSs

We define a λ-CSS as the minimal connected dependency tree covering a given set of words σ in a given sentence S, where words in σ are represented not necessarily by their lemmas and POS, but by a set of properties λ. Words are still connected according to their syntactic dependencies, but these dependencies are only labeled if the corresponding feature (noted ‘deprel’) is in λ. Insertions (words necessary for the tree to be connected but not in σ) are represented by dummies. When a word in σ does not have a certain feature from λ (such as a noun not having a tense), the feature is marked as null for the word.

For instance, if figure 5 is the morphosyntactic analysis of sentence (1), then figure 6 is the $\{form, deprel, number\}$-CSS of the MWE component words. Similarly, figure 7 is the $\{lemma, pos, deprel\}$-CSS of the MWE in figure 3.

We will now ask which combination of features λ gives the best basis for a MWE-L formalism. We only consider formalisms where a unique set of features λ is used to describe all MWEs. While a formalism where each MWE is represented by its optimal set of features could be very interesting, we find that: (i) this would greatly increase the complexity of the experimental setup; (ii) results on

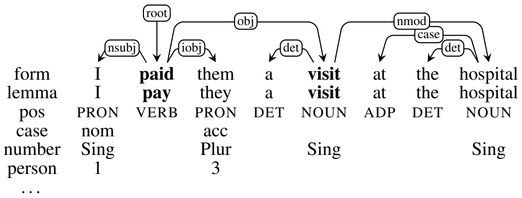


Figure 5: Dependency graph with all features of a sentence.

Figure 6: $\{form, deprel, number\}$-CSS of the MWE in 5, and its simplified representation (on the right).

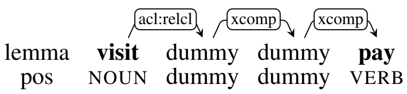


Figure 7: $\{\,lemma, pos, deprel\,\}$-CSS of the syntactically discontinuous subsequence in bold from figure 3

less frequent MWEs would be dubious at best; (iii) it is still interesting to know which set of features is best on average;

7 Results

We use the German (DE), Greek (EL), French (FR), Hebrew (HE), Hindi (HI), Italian (IT), Polish (PL), Portuguese (PT), Swedish (SV), Turkish (TR) and Chinese (ZH) PARSEME shared task 1.2 corpus (Ramisch et al., 2020).[3]

Given a lexicon and a sentence, we define a *match* as a subsequence of the sentence which is accounted for (recognized by) the lexicon. A match can correspond to an idiomatic MWE occurrence or not. In the former case, it is called an *idiomatic match*. Then, given a lexicon and a corpus of sentences, we define: *precision* as the ratio of idiomatic matches to the total number of matches; and *recall* as the ratio of idiomatic matches to the number of idiomatic occurrences in the corpus. The aim is to maximise both measures.

As proposed earlier, formalisms will be evaluated in conjunction with a given instantiation method and instantiation corpus. To that end, during instantiation phase, we collect the λ-CSSs of all idiomatic occurrences annotated in the instantiation corpus. This method has the advantage of being very simple to implement and to introduce very little variation during the instantiation process. Its one downside (beside needing annotated data) is that some properties of MWEs cannot be deduced from single observations, i.e. the descriptive adequacy of the instantiated lexicon is limited.

[3]Basque, Irish and Romanian are skipped for technical reasons.

7.1 Optimal set of features λ

In this section we aim to find the optimal set of features λ for MWE representation in MWE-L based on λ-CSS or λ-CSS lexicons for short.

Since we have not one, but two evaluation criteria (precison and recall), and because we wish to avoid making a priori choices on how they should be combined (Hwang and Masud, 2012) (at least during the exploration of the solution space), we will for now only consider a solution A to be better than another solution B if A dominates B. That means that A is considered better than B on at least one criterion and better or equal on the others.

Depending on the language, from 17 to 40 features are considered. Some features such as $lemma$, $form$, $upos$ or $deprel$ are available in all language and for all words, while others such as $Number$ or $Aspect$ are only occur for some words and languages. The number of subsets of features that can be used for MWE representation always is quite large. A comprehensive exploration of the solution space is therefore out of the question.

Since our solution space is the powerset of the considered features, it can be seen as a lattice, i.e. a graph where each solution is represented by a node. Then, a solution A is connected to solutions with all features in A plus or minus one. Each solution therefore has a neighbourhood of similar solutions (with one feature of difference each). We then perform a greedy exploration of the solution space that considers non-dominated solutions as those to be explored. When two neighbouring solutions have equal precision and recall, we consider the simplest of the two neighbours to be the preferable solution. This criterion is not explicitly evaluated, but enforced by the exploration algorithm 1 (line 8), where $score(s)$ returns the position of a given solution in the objective space, and $Pareto(S)$ returns the set of non-dominated solutions.

This algorithm was run 2-fold using TRAIN+DEV datasets, half of the dataset was used to generate MWE-Ls, and another half for OA evaluation. This was done twice per corpus, once with $\{\,lemma\,\}$, and once with $\{\,form\,\}$,

Algorithm 1: Bottom-up Greedy Pareto

Data:
$features$: the set of all considered features
s: a subset of $features$

Initialization
- $last\_it\_res \leftarrow \{ s \}$
- $res \leftarrow \{ s \}$

while $last\_it\_res \neq \emptyset$ **do**
- $tmp \leftarrow \emptyset$
- **foreach** $s_i \in last\_it\_res$ **do**
  - **foreach** $f_i \in features \setminus s_i$ **do**
    - **if** $score(s_i \cup \{ f_i \}) \neq score(s_i)$
      - $tmp \leftarrow tmp \cup \{ s_i \cup \{ f_i \} \}$
- $last\_it\_res \leftarrow Pareto(res \cup tmp) \cap tmp$
- $res \leftarrow res \cup last\_it\_res$

Result: res

as the starting set of features s.[4] All solutions generated in this way were then re-evaluated by instantiating the lexicon from TRAIN+DEV, and scoring it against the TEST dataset. In the end, 12, 142, 14, 36, 7, 20, 22, 22, 16, 22, 16 solutions were selected for DE, EL, FR, HE, HI, IT, PL, PT, SV, TR, ZH respectively.[5]

Table 1 presents the solutions provided by algorithm 1 on the French corpus. A clear distinction between solutions can be made depending on whether they use $form$ or $lemma$. The former have high precision and low recall, while the latter have more balanced precision and recall. Solutions using both act as the former.

As shown in table 2, the solutions with the highest precision always use $form$ and most of them use $deprel$. The solutions with the highest recall systematically use $lemma$. The most harmonious solutions (i.e. those with the highest F-scores) almost always use $deprel$, , $lemma$ or both. However, Greek (EL), skipped in the table due to the large size of its optimal solution, Hebrew (HE), and Chinese (ZH) act in quite unique ways. On the Greek corpus, features such as the $case$ and the $voice$ are used in both the most precise and the most harmonious solutions. In Hebrew and Chinese, $form$ is used instead of $lemma$ in the most harmonious solutions. However, the solutions with the highest recall still use $\{ lemma \}$ with both languages.

P (%)	R (%)	solution features
71.78	75.06	lemma
73.18	74.91	lemma, upos
78.60	71.08	lemma, deprel
84.08	52.47	form
85.42	52.17	form, lemma
85.27	51.95	form, upos
85.54	51.80	form, lemma, upos
87.94	48.27	form, deprel
88.02	48.12	form, lemma, deprel
87.84	47.83	form, upos, deprel
87.94	47.76	form, lemma, upos, deprel
87.16	47.46	form, lemma, upos, deprel, Number
87.16	47.46	form, upos, deprel, Number
86.93	47.46	form, lemma, deprel, Number

Table 1: Precision(P) and Recall(R) for selected solution for French

	P	R	F
DE	lem+form+deprel	lem	lem+deprel
FR	lem+form+deprel	lem	lem+deprel
HE	form+upos+Voice	lem	form
HI	form+deprel	lem	lem+deprel
IT	form+deprel+upos	lem	lem+deprel
PL	form+deprel	lem	lem+deprel
PT	lem+form+deprel	lem	lem+deprel
SV	form,+deprel	lem	lem+deprel+upos
TR	lem+form+upos+ deprel	lem	lem+deprel
ZH	form+deprel+upos+lem	lem	form+deprel+upos

Table 2: Best performing solutions according to Precision (P) and Recall (R) and F-score (F); lem stand for $lemma$.

Table 3 presents the F-scores of the solutions $\{ lemma, deprel \}$, $\{ form, deprel \}$, $\{ lemma, deprel, upos \}$ and, when necessary, the solutions with the best F-score in order to: (i) get a better understanding of the impact of using $lemma$ over $form$ (used in conjunction with $deprel$ since this leads to more precise and more harmonious solutions), (ii) to compare the score of the original CSS ($\{ lemma, deprel, upos \}$) to what appears to be the most harmonious CSS for most languages: $\{ lemma, deprel \}$.

As expected, the scores of $form$ based solution in Hebrew and Chinese are well above those of $lemma$ based solution (but not particularly higher than the scores of $form$ based solutions in other languages). Conversely, for all other languages, $lemma$ based solution perform much better than $form$ based solutions. As for the differences between $\{ lemma, deprel \}$ and $\{ lemma, deprel, upos \}$, we can see that in

[4] Solutions with neither of these features resulted in huge numbers of mostly non-idiomatic matches, not worthy of systematic exploration.

[5] Technical issues prevented algorithm 1 to be run in reasonable time on Greek with $\{ form \}$.

	DE	EL	FR	HE	HI	IT	PL	PT	SV	TR	ZH
form, deprel	57.66	51.12	62.33	**32.66**	47.21	47.85	61.41	49.54	56.77	38.66	**46.92**
lemma, deprel	**69.07**	59.71	**74.65**	7.49	**64.80**	**64.00**	**81.58**	**72.86**	**75.21**	**61.08**	14.81
lemma, deprel, upos	67.92	**59.80**	74.55	20.35	64.54	**64.00**	80.05	72.54	**75.21**	60.82	20.70
highest F		**60.93**		**37.65**							**47.44**

Table 3: F-score(%) of selected λ-CSS based lexicon

most languages adding *upos* slightly deteriorate F-scores. This deterioration is however quite noticeable in German (DE) and Polish (PL). On the other side, in Greek (EL) and Swedish (SV), the results are only marginally better with *upos*.[6] In short, apart from Hebrew (HE) and Chinese (ZH), the solution $\{lemma, deprel\}$ is either the one with best F-score or very close to be so, while it is also one of the simplest solutions.

7.2 Sequential discontinuity based lexicon and non-verbal MWE

We now compare our $\{lemma, deprel\}$-CSS lexicon format to various list-like formalisms analogous to those discussed in Sec. 3. To cover MWE of all syntactic types, we use the French Sequoia corpus (Candito et al., 2021) annotated for both verbal and non-verbal MWEs, along with the French corpus of PARSEME shared task 1.2, annotated for verbal MWEs only.

As earlier, MWE-Ls are instantiated by looking at the MWEs annotated in the TRAIN+DEV corpora, then OA is evaluated on the TEST corpora.

All the list-like MWE-Ls considered here operate in similar fashion. Once an annotated MWE occurrence is encountered in the instantiation corpus, a lexical entry is created storing the *lemma*s of the MWE components in the sequential order in which they appear. Discontinuities are handled with 4 different methods with varying details about the inserted elements, stored in between the components. Below, each method explained and illustrated with the lexical entries instantiated from sentence (1):

1. contiguous: discontinuous MWEs are ignored, e.g. example (1) yields $\emptyset$
2. $[lemma]$: the list of *lemma* of the insertions is stored, here: [**pay**, [they, a], **visit**]
3. $[upos]$: the list of *upos* of the insertions is stored, here: [**pay**, [PRON, DET], **visit**]
4. *: insertions are represented by the special character '*', meaning that any insertion (or none) can happen, here: [**pay**, *, **visit**]

A common practice is to limit the maximum size of discontinuities, in order both to reduce the computational cost of identification and to possibly improve precision. To mimic such a practice, we run our list-like MWE-Ls in 4 different configurations. With $n = [1, 2, 3]$, only insertions of n words or less are considered, occurrences with larger insertions are ignored during instantiation and identification. In the 4th configuration the size of insertions is ignored.

	FR Sequoia			FR PARSEME		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
λ-CSS	90.74	67.74	**77.57**	78.60	71.08	**74.65**
contiguous	91.76	56.45	69.90	71.63	48.49	57.83
$[lemma]$						
1	91.12	63.82	75.07	71.90	60.63	65.79
2	90.94	64.75	75.64	72.17	61.44	66.38
3	91.00	65.21	75.97	72.09	61.59	66.43
∞	91.00	65.21	75.97	72.08	61.74	66.51
$[pos]$						
1	90.85	64.06	75.14	72.10	63.50	67.53
2	90.68	64.98	75.70	72.52	65.05	68.58
3	90.73	65.44	76.04	72.47	65.27	68.68
∞	90.73	65.44	76.04	72.45	65.42	68.75
*						
1	86.42	64.52	73.88	67.26	66.37	66.81
2	79.56	66.36	72.36	63.13	71.82	67.19
3	74.23	67.05	70.46	58.20	73.66	65.02
∞	33.22	67.97	44.63	26.05	75.86	38.78

Table 4: Precision, Recall and F-score of λ-CSS MWE-L and list-like MWE-L on french corpora (with and without non verbal MWE respectively)

In table 4 we find the OA, measured by way of precision (P), recall (R) and F-score (F), of MWE-Ls based on $\{lemma, deprel\}$-CSS, and the 4 methods above. Results of the last three MWE-L formalisms are decomposed according to the maximal size of insertions.

We chose to ignore the MWE ***de le*** 'of the', annotated 34 times in the Sequoia's TRAIN+DEV and 2 times in the TEST. If not for this, the precision of the list-like MWE-Ls would go from around 90%

[6] Swedish results with and without *upos* are so close that they appear equal with 4 significant figures.

	DE	EL	FR	HI	IT	PL	PT	SV	TR	HE	ZH
MTLB-STRUCT	76.17	**72.62**	**79.42**	**73.62**	63.76	81.02	73.34	71.58	69.46	**48.30**	**69.63**
union	**76.45**	71.12	78.87	73.29	62.92	<u>81.41</u>	**74.76**	<u>73.74</u>	**69.92**	44.29	58.43
$\{lemma, deprel\}$-lexicon	69.07	59.71	74.65	64.80	**64.00**	**81.58**	72.86	**75.21**	61.08	7.50	14.81

Table 5: F-score (%) of MTLB-STRUCT, our lexicon, and the union of their predictions.

to around only 45% since *de le* is an extremely frequent combination of words which is almost never idiomatic. This choice only barely affects the results of the $\{lemma, deprel\}$-CSS lexicon but allows for a much fairer comparison.

The first thing to notice is that precision is on the whole higher on Sequoia corpus than on the FR PARSEME corpus. This is somewhat expected since verbal MWEs are often harder to identify than non-verbal MWEs. Our takeaway, is that even though the $\{lemma, deprel\}$ was optimised for OA of verbal MWEs, $\{lemma, deprel\}$-CSS lexicon perform correctly (or even better) on MWEs not restricted to verbal MWEs. The second conclusion is that our MWE-L is more observationally adequate than any of the list-like MWE-Ls tested here. This seems especially true on verbal MWEs where the advantages of dependency representation are crucial.

7.3 Impact of lexicon on identification

In this section we investigate how $\{lemma, deprel\}$-CSS lexicons compare to a traditional a MWE identifier, and we estimate their usefulness in assisting MWE identification.

Riedl and Biemann (2016) showed that using MWE resources during MWE identifier training can improve the quality of the identification. Alas, going as far as intervening during the training of MWE identifier is outside the scope of this study. As a alternative we propose a naive a posteriori approach where we simply compare the MWE identifier scores to those of the union of the identifier and MWE-Ls annotations.

In table 5 we compare the F-score of MTLB-STRUCT (Taslimipoor et al., 2020) – a deep learning based MWE identifier, the winner of the PARSEME shared task 1.2 – to our $\{lemma, deprel\}$-CSS lexicon and to the union of predictions. F-scores on the Hebrew (HE) and Chinese (ZH) corpora are isolated since on these corpora $lemma$ based solutions were outperformed by $form$ based solutions. Since even $form$ based solution were largely underperforming for these corpora we do not expect $\{lemma, deprel\}$-CSS to compare to MTLB-STRUCT on those corpora.

We do not expect a $\{lemma, deprel\}$-CSS lexicon to out-perform a state-of-the-art MWE identifier if both have access to the same training data. The lexicon's potential in rather to be seen in its future use with data discovered in large non-annotated corpora. Still, it is interesting to notice that even now: (i) MWE-Ls annotations do outperform MTLB-STRUCT on 3 out of the 11 corpora, (ii) the union of both annotations outperforms MTLB-STRUCT on 5 out of the 11 corpora.

This shows that the OA of $\{lemma, deprel\}$-CSS lexicons is reasonable, i.e. comparable to the state-of-the-art in MWE identification. We also obtained encouraging results towards assisting MWE identification with lexicons.

8 Concluding Remarks

In this paper we proposed, to our knowledge, the first method of quantitatively evaluating MWE-lexicon formalisms through observational adequacy. We also presented a MWE-lexicon formalism based on a generalisation of the concept of a Coarse Syntatic Structure, which we call $\{lemma, deprel\}$-CSS. We brought evidence that this specific set of features allows for higher observational adequacy than alternative sets of features on verbal MWEs in most of the 11 languages studied. Furthermore, we compared this formalism to MWE-lexicons based on sequential representation of MWEs. We showed that our formalism achieves higher observational adequacy on French regardless of the fact that only verbal or all types of MWEs are considered. Finally, we showed the observational adequacy of our formalism holds its own even when compared to annotations produced by a state-of-the-art MWE identifier. While this study focuses on MWE-lexicon formalisms instantiated on annotated corpora, our vision is that such lexicons should be instantiated through MWE discovery in large non-annotated corpora or through extraction from other MWE resources.

References

Hassan Al-Haj, Alon Itai, and Shuly Wintner. 2014. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2):130–170.

Rania Al-Sabbagh, Roxana Girju, and Jana Diesner. 2014. Unsupervised construction of a lexicon and a repository of variation patterns for Arabic modal multiword expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 114–123, Gothenburg, Sweden. Association for Computational Linguistics.

Iñaki Alegria, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. 2004. Representation and treatment of multiword expressions in Basque. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 48–55, Barcelona, Spain. Association for Computational Linguistics.

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

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. Saldo: a touch of yin to wordnet's yang. *Language resources and evaluation*, 47(4):1191–1211.

Elisabeth Breidt, Frederique Segond, and Giuseppe Valetto. 1996. Formal description of multi-word lexemes with the finite-state formalism IDAREX. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2021. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2).

Noam Chomsky. 1965. Aspects of the theory of syntax.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Helge J Jakhelln Dyvik, Gyri Smørdal Losnegaard, and Victoria Rosén. 2019. Multiword expressions in an lfg grammar for norwegian.

Nicole Grégoire. 2010. Duelme: a dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1):23–39.

C-L Hwang and Abu Syed Md Masud. 2012. *Multiple objective decision making—methods and applications: a state-of-the-art survey*, volume 164. Springer Science & Business Media.

Ray Jackendoff. 1975. Morphological and semantic regularities in the lexicon. *Language*, pages 639–671.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. Parseme survey on mwe resources. In *9th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2299–2306.

Kemal Oflazer, Özlem Çetinoğlu, and Bilge Say. 2004. Integrating morphology with multi-word expression processing in Turkish. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 64–71, Barcelona, Spain. Association for Computational Linguistics.

Pavel Pecina. 2008. Reference data for czech collocation extraction. In *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 11–14.

Adam Przepiórkowski, Jan Hajič, Elżbieta Hajnicz, and Zdeňka Urešová. 2017. Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography*, 30(1):1–38.

Valeria Quochi, Francesca Frontini, and Francesco Rubino. 2012. A MWE acquisition and lexicon builder web service. In *Proceedings of COLING 2012*, pages 2291–2306, Mumbai, India. The COLING 2012 Organizing Committee.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Martin Riedl and Chris Biemann. 2016. Impact of MWE resources on multiword recognition. In *Pro-*

ceedings of the 12th Workshop on Multiword Expressions*, pages 107–111, Berlin, Germany. Association for Computational Linguistics.

Manfred Sailer and Beata Trawiński. 2006. The collection of distributionally idiosyncratic items: A multilingual resource for linguistic research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Agata Savary. 2008. Computational inflection of multiword units. *Linguistic Issues in Language Technology*, 1.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten van Gompel, et al. 2018a. Parseme multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Agata Savary, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019a. Literal occurrences of multiword expressions: rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics*.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019b. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*.

Agata Savary, Simon Petitjean, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2018b. Object-oriented lexical encoding of multiword expressions: Short and sweet. *arXiv preprint arXiv:1810.09947*.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Stefania Spina. 2010. The dictionary of Italian collocations: Design and integration in an online learning environment. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. Mtlb-struct@ parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. *arXiv preprint arXiv:2011.02541*.

Aline Villavicencio, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical encoding of MWEs. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 80–87, Barcelona, Spain. Association for Computational Linguistics.

Pavel Vondřička. 2019. Design of a multiword expressions database. *The Prague Bulletin of Mathematical Linguistics*, 112(1):83–101.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2019. Ilfhocail: A lexicon of Irish MWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 162–168, Florence, Italy. Association for Computational Linguistics.