Lexicon for multiword expression identification

Anonymous ACL submission

Abstract

Following the idea that lexicons are needed in order for automatic identification of multiword expressions(MWE) to handle the unpredictable nature of MWEs, this paper proposes a lexicon formalism, itself declined in multitudes of possible sub-formalisms depending on the linguistic features considered , along with an evaluation method which could be used to compare lexicon formalisms to each other. An exploration of the powerset of features is done in order to find the bests of such subset of features to be used. The impact of the proposed lexicon formalism on MWE identification is investigated, leading us to conjecture that lexicon indeed have the potential to help MWE identification.

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 MWE) 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 regular identification but also some of 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, shifting the burden of unseen MWEs as much as possible on discovery and using lexicon as the interface between discovery and identification.

2 Multiword Expression

We abide by PARSEME's definition of MWE [Savary et al., 2018a] (adapted from [Baldwin and Kim, 2010]) :

> multiword expressions (MWEs) are understood as (continuous or discontinuous) sequences of words which:
> • contain at least two component words which are lexicalised, i.e. always realised by the same lexemes [...], including a head word and at least one other syntactically related word,
> • display 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] the following 3 have a large impact on our view of MWE lexicons.

- variability
- discontinuity
- literal-idiomatic ambiguity

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 **vists** 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. We call the various forms of a MWE and the condition under which they are taken the paradigm of the MWE.

Discontinuity Discontinuity can be seen as a form of variability where component words of a MWE are not adjacent to one another and are separated by a word or group of words. 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***) ; 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[1]).

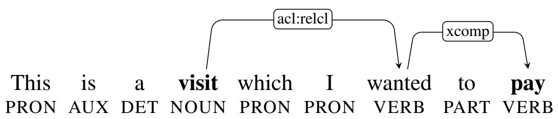

Figure 1: syntactic discontinuity

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 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.

[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 Lexicon for MWE

The word lexicon is colloquially used to describe something akin to a dictionary or a collection of words. In the literature, what exactly constitute the lexicon is not yet fixed. As shown by Di Sciullo and Williams [1987], the set of "words" described by morphology is distinct from the set of "words" used by syntax, and while both could be considered as the lexicon, they also are distinct from the set of "words" which Di Sciullo and Williams would consider as the lexicon. Moreover, lexicon definitions will vary on how they represent their set of lexical item. Extensional lexicon will tend to list their lexical item forms, while intentional lexicon will make use of tools such morphology, lexemes and inflectional paradigm to describe the lexical items in generative ways. While theses discussion are about **the** lexicon, we consider here that the MWE lexicon works in very similar fashion and the same reasoning can be held.

We will neither settle for one of those views nor introduce a new definition of what might be a MWE lexicon, but instead take from the notion of adequacy [Jackendoff, 1975, Chomsky, 1965] to set objectives, allowing us to compare different view of what is a MWE-lexicon.

First, we define the MWE-lexicon observational adequacy as the property of a MWE lexicon to characterize MWEs' forms. Observational adequacy can be evaluated from a generative standpoint (can the lexicon generate a list list of all imaginable forms) as well parsing standpoint (can the lexicon recognize all the forms encountered in a text). Either way, for a MWE lexicon to be observationally adequate, forms have to be characterized in such a way that variability and discontinuity are both taken into account and to make as few mistake as possible (if any).

Then, we define the MWE-lexicon descriptive adequacy as the property of MWE lexicon to accurately describe MWEs. This description should cover all the property associated to the MWE (Morphosyntactic property of its component, Syntactic relations of the components, Meaning of the MWE, Syntactic role of the MWE, ...) and must be in accordance with linguistic knowledge.

Both of these adequacies can be interpreted and measured in various ways. While observational adequacy can be relatively straight forward to measure in quantitative ways, descriptive adequacy will more readily be viewed from a qualitative or categorical standpoints where two descriptions of the same phenomenon might differs without even being comparable.

4 Related works

Numerous MWE-lexicon formalisms have been put forward through the years. Of all aspects of the descriptive adequacy, we find that how discontinuities, component order and component relations to each others are managed offer great insight into MWE-lexicon formalisms and will therefore describe lexicon formalisms through that lens.

Apart from lexicons which cannot represent MWE discontinuities. Probably one of the less descriptive way of handling discontinuities is to view MWEs' forms as ordered sequences of their component and to denote insertions in the list as either "*" or "+" indicating whether insertions are possible or mandatory respectively. [Al-Haj et al., 2014, Alegria et al., 2004]. Descriptiveness aside, one big issue of such formalism is that not restricting what can be inserted between MWE component results in high number of false positives (lexicon mistaking non-idiomatic occurrences for idiomatic occurrences.). More descriptive formalisms might choose to instead use phrase grammar to denote insertions and therefore restrict what phrases can be inserted in a given MWE. [Grégoire, 2010] Doing so will mechanically reduce the number of false positive, but also impede the lexicon power of generalisation.

Even more descriptive formalisms will use highly expressive grammar such as XMG or LFG in order to restrict as specifically as needed what can and cannot be inserted in a MWE. [Savary et al., 2018b, Dyvik et al., 2019] Such formalisms aim for comprehensive description of MWEs which is quite appealing and makes for great linguistic tool. Highly expressive grammar will however have a computational cost, and might be harder to interface to already existing MWE identifiers.

Other solutions which cannot necessarily be said to be of higher or lower descriptive power than non comprehensive formalism include Villavicencio et al. [2004] where constraint on insertions are based on semantic relations and logical role of the words and our proposed formalism (discussed latter) where syntactic dependencies are used in order to restrict insertion and component order.

A formalism being descriptive in its handling of discontinuities is not the be-all end-all of MWE-lexicon. We argue that observational adequacy could serve as a great tool to compare MWE-lexicons and by extension MWE-lexicon formalisms. There, however, do not seem to be any consensus on how to measure observational adequacy. Of all the works cited earlier not even two of them seem to measure observational adequacy (or similar concepts) in the same way. Some might look at the number of MWE covered by their lexicon, some the number of MWE occurrence covered. But in any case, it seems that no comparison can be done between any of them.

More than being a great comparison tool, we hope that lexicon with great observational adequacy can be used as MWE identifier to be used in conjunction to an traditional MWE-identifier. We will therefore evaluate our lexicon formalism from a observational adequacy standpoint.

Keeping in mind our goal to promote observational adequacy and remarks made on the handling of discontinuities, we introduce our lexicon formalism in the following section.

5 λ-CSS lexicon

5.1 litteral occurrences

Savary et al. [2019a] ask what exactly is a literal occurrence of a MWE and what distinguishe 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 <u>idiomatic occurrence</u> of the MWE. Whereas when they display no idiosyncrasy, we talk of an <u>non-idiomatic occurrence</u> of the MWE.

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

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

(2) I <u>paid</u> them a <u>visit</u> to the museum

(3) I <u>paid</u> for a <u>visit</u> of the museum

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

5.2 Coarse Syntactic Structure (CSS)

A CSS can be seen as a simplification of the dependency tree of a given MWE occurrence. More precisely, given set of word σ and given sentence S, a CSS can be defined as the minimal connected dependency tree covering the words σ in S, where words are either represented by a node containing their lemma and part of speech if they are in σ, or by a dummy node if not. Nodes are connected by their relational dependencies.

For example, for the sentence in figure 2 where bold words are components of a MWE, and figure 3 its dependency tree where word form are replaced by their lemma and part of speech (POS), figure 4 is the CSS of the bold words of figure 2. And in figure 9 the CSS of a MWE with syntactic discontinuities in bold in figure 5.

I **paid** them a **visit** at the hospital

Figure 2: A sentence S and its subsequence σ in bold, an occurrence of a MWE.

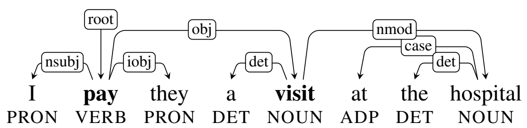

Figure 3: The dependency graph of the sentence from figure 2.

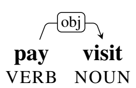

Figure 4: The coarse syntactic structure CSS(S, σ) of the subsequence σ in S.

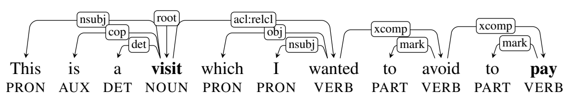

Figure 5: Dependency tree of a sentence with a syntactic discontinuity

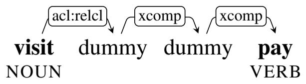

Figure 6: Coarse syntactic structure of the syntactically discontinuous subsequence in bold from figure 5

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 lexicons with hopefully great observational adequacy.

Such a lexicon would simply consist in a set of CSSs. We will however first question the relevancy of component word being represented by their lemma and POS and not some other features. Word being represented by their lemma and POS serves as an approximation of their lexemes, which seems like a perfectly reasonable choice in order for CSS to do what they were designed to do (which is help approximate our intuitive notion of literal occurrence). We however would like for our lexicon to be as observationally adequate as possible and we will therefore wonder if representing MWE by a different set of feature would be beneficial to our end goal.

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

5.3 λ-CSS

We define λ-CSSs as the minimal connected dependency tree covering a given set of word σ in a given sentence, where words are represented by their property in λ if they are part of σ and by dummy nodes if not. Words are still connected according to their syntactic dependencies, but these dependencies are only labeled if the corresponding feature is in λ. Insertions (word necessary for the tree to be connected but not in σ are represented by dummy. When a word of σ does not have a feature which is considered by the λ-CSS (such as a noun not having a tense), the feature is marked as null for the word.

6 Optimal set of feature

Considering all the features available in PARSEME corpora, our aim is to find if there are better set of features λ than $\{lemma, pos, deprel\}$ for MWE represen-

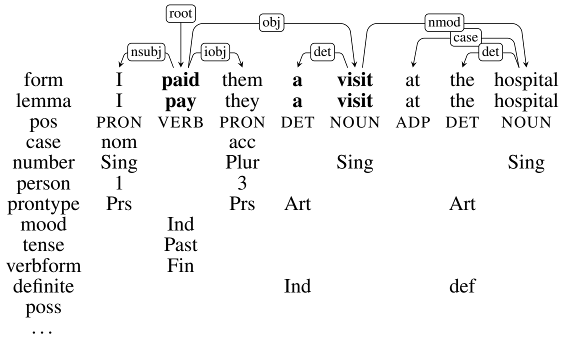

Figure 7: Dependency graph with all features of a sentence.

Figure 8: $\lambda = \{form, deprel, number\}$-CSS of the MWE in 7, and its simplified representation (on the right).

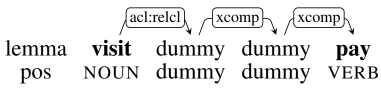

Figure 9: $\{ lemma, pos, deprel \}$-CSS of the syntactically discontinuous subsequence in bold from figure 5

tation, and hopefully to find the best of such λ.

As stated earlier, we've made the choice to focus on observational adequacy, more precisely, we will evaluate observational adequacy from a parsing standpoint. Given a lexicon and a sentence, we define a match as a subsequence of the sentence which is also characterized (can be recognized) by the lexicon. A match is said to be idiomatic if its subsequence is also an idiomatic occurrence of a MWE, and non-idiomatic otherwise.[2] Thus, given a lexicon and a corpus of sentences, we use the following two measures: precision, the ratio of number of idiomatic matches to total number of matches ; and recall, the ratio of the number of idiomatic matches to the number of idiomatic occurrences in the corpus. In both case the desired output is to maximise the measures.

Since we have not one, but two criteria to evaluate of observational adequacy, and because we wish to avoid making a priori choices on how those two criteria might be related to each other [Hwang and Masud, 2012], we only consider a solution A to be better than another solution B if A dominates B. Meaning that A is better than B on at least one criteria and better or equal on the other.

Both a corpus of sentence and a lexicon are needed in order to compute our measures. This means that our proposed measures do not actually evaluates lexicon formalism, but instances of those formalisms on specific corpora. In order for these measures to be applicable to lexicon formalisms, we propose that lexicon formalisms should be evaluated in conjunction with a instantiation method and a instantiation corpus. Lexicons formalisms can therefore be compared by fixing the instantiation method and the instantiation corpus.

In our case, we use the german (DE), basque (EL), french (FR), hebrew (HE), hindi (HI), italian (IT), polish (PL), portugese (PT), Swedish (SV), turkish (TR), and chinese (ZH) corpora used in PARSEME shared task 1.2 [Ramisch et al., 2020]. (The Greek, Irish and Romanian corpora have not been used for technical reasons.) For each corpus, and for each considered set of features λ, we instantiate our lexicon by collecting the λ-CSSs of every MWE idiomatic occurrences annotated in the corpus.

Depending on the language from 17 to 40 features are considered. Hence, the number of subset of features that can be used for MWE representation is at best quite large or at worst astronomical. A comprehensive exploration of the solution space is therefore out of the question.

Our solution space being the powerset of the considered features, it can be seen as a lattice (a graph) where each solution A is connected to solutions with either : all features from A plus one more feature, or all features from A but one. Each solution can therefore be seen as having a neighbourhood of similar solutions (with one feature of difference each). Hoping that this notion of neighbourhood in the solution space persists in some fashion in the objective space, we opted for a greedy exploration of the solution space considering non-dominated solution as to be explored.

We note here than, while not a explicit criterion, when two neighbouring solutions have equal precision and recall, we find the simplest of the neighbour to be a preferable solution. This crite-

[2]In order to reduce confusion as much as possible, we use the term match to refer to a subsequence of a sentence characterized by a lexicon, and the term occurrence to refer to a subsequence corresponding to a MWE.

rion is not explicitly evaluated, but only implicitly enforced by the exploration algorithm 1 described here :

Algorithm 1: Greedy Pareto bottom-up lattice exploration

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 $objective(f_i) \neq objective(s_i)$ **then**

$tmp \leftarrow tmp \cup \{ s_i \cup \{ f_i \} \}$

$last\_it\_res \leftarrow$
$\{ s_i \mid Pareto(res \cup tmp) \cap tmp \}$

$res \leftarrow res \cup last\_it\_res$

Result: res

With $objective(s)$ the function returning the position in the objective space of a given solution, and $Pareto(S)$ the function returning the set of non-dominated solutions of of a set of solutions.

Algorithm 1 was ran 2-fold using only the TRAIN datasets, half the dataset was used to generate lexicon, half for evaluation of the lexicon. This was done twice per corpus, once with $\{ lemma \}$ as the starting set of feature s, the other with $\{ form \}$. Solutions with neither of these feature where found to induces extreme number of matches, most of which were non-idiomatic, and not worthy of systematic exploration. All solutions generated by algortihm 1 were then re-evaluated by generating the lexicon from the TRAIN and DEV dataset, and scoring them against the TEST dataset. In the end 12, 14, 36, 7, 20, 22, 22, 16, 22, 16 solutions were selected for DE, FR, HE, HI, IT, PL, PT, SV, TR, ZH respectively.[3]

In table 1 the resulting solutions from algorithm 1 on the french corpus. A clear distinction between solutions can be made on whether a solutions uses $form$ or $lemma$. Solutions based on $form$ have high precision and low recall, while solution using $lemma$ and not $form$ have more balanced precision and recall, solutions using both $form$ and $lemma$ act as solutions using $form$. Not shown here, however the same can be said on other languages, as for all languages the highest precision solution all uses $form$ while the highest recall solutions all use $lemma$.

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

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

While tempting to aggregate our precision and recall into a F-score, as is often done, in order to be able to choose a preferred solution, we need to remember that we only aim for observationally adequate lexicon formalism in order to help automatic MWE-identifier. Whether good recall, good precision or some balance of the two is needed in order to accomplish our goal is yet unknown. Nevertheless, we still need to choose a preferred solution, we will therefore look at our solution's arithmetic, harmonic (F-score), and geometric means of their precision and recall in hope to get a better overview of our solutions than with a simple F-score. In table 2, the best performing solutions found through algorithm 1 for each languages according to each of these means. Except in Hebrew and Chinese (and perhaps Basque) where $form$ is used, in almost all other cases, the solution $\{ lemma, deprel \}$ is the best performing solution according to all three means. Only in Swedish is $\{ lemma, deprel, upos \}$ preferred to $\{ lemma, deprel \}$.

In table 3, we compare the solutions $\{ lemma, deprel \}$, $\{ form, deprel \}$ and $\{ lemma, deprel, upos \}$ (original CSS) on their arithmetic, harmonic, and geometric means. Once again, except for Basque, Hebrew and Chinese, all three means tend agree and place $\{ lemma, deprel \}$ as the best performing of the three solution. The difference between $\{ lemma, deprel \}$ and $\{ lemma, deprel, upos \}$ are however very slight. Still, since they are

[3] technical issues prevented algorithm 1 for basque to be run in timely fashion.

	arithmetic (%)	harmonic (%)	geometric (%)
DE	–	lemma deprel	–
FR	–	lemma deprel	–
HE	–	form hebSource	–
HI	–	lemma deprel	–
IT	–	lemma deprel	–
PL	–	lemma deprel	–
PT	–	lemma deprel	–
SV	–	lemma deprel upos	–
TR	–	lemma deprel	–
ZH	–	form deprel upos	–

Table 2: Best performing solution according to Precison and Recall means (agreement between means shown by "–")

two neighbouring solution we will prefer the simpler of the two, consolidating our preference for the solution $\{lemma, deprel\}$. As of now, will therefore consider $\{lemma, deprel\}$ as the overall best performing solution across our corpora. (While keeping in mind that this does not hold for Basque, Hebrew and Chinese.)

7 Impact of lexicon on identification

If and how lexicon based upon $\{lemma, deprel\}$-CSS can help an automatic identifier is still to be determined. Lexicon and identifier both production MWE annotations, these annotations can be merged in two main fashion. Either only the intersection of their annotations is considered as annotated (annotation must be approved by both the lexicon and the identifier), either the union of their annotations is considered as annotated.

In table 4 we compare one of the best performing identifier in PARSEME shared task 1.2, MTLB-STRUCT [Taslimipoor et al., 2020, Savary et al., 2018a], our $\{lemma, deprel\}$-CSS based lexicon (still trained on both the TRAIN and DEV datasets), the union, and the intersection of both predictions.

Despite the fact that in almost all cases MTLB-STRUCT outperforms our lexicon, the union of their respective prediction actually slightly improve on MTLB-STRUCT prediction for 5 corpora out of 11. While not a clear improvement, we need to remember that our lexicon was automatically generated in a rather naive fashion and actually did not possess any information not at disposition of MTLB-STRUCT since they were trained on the same dataset. The argument for the use of lexicon in MWE identification is that lexicon can help for unseen MWE, but in this case our lexicon could not help on MWE not seen by MTLB-STRUCT.

In table 5 we cheat and generate our lexicon using the TEST dataset. Results are to be interpreted with utmost care, but still give an interesting peek on how much a lexicon could help traditional identifier. Here the union of our lexicon to MTLB-STRUCT clearly improve on MTLB-STRUCT results. Once again, these result are far from being proof of the benefits of using lexicon for MWE-identification, but are still encouraging. Better ways of quantifying the benefits of a lexicon on MWE-identification of unseen MWE must be found.

8 Concluding remarks

In this paper we proposed a generalisation of the concept of Coarse Syntatic Structure called λ-CSS and argued that it could be the basis of MWE lexicon. Argument for $\{lemma, deprel\}$-CSSs to be used as MWE lexicon is put forward. A first attempt to measure the impact of lexicon formalism on MWE identification is put forward, however no clear conclusion is drawn from this attempt.

References

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

I. Alegria, O. Ansa, X. Artola, N. Ezeiza, K. Gojenola, and R. Urizar. Representation and treatment of multiword expressions in basque. In Proceedings of the Workshop on Multiword Expressions: Integrating Processing, pages 48–55, 2004.

T. Baldwin and S. N. Kim. Multiword expressions. In N. Indurkhya and F. J. Damerau, editors, Handbook of Natural Language Processing, Second edition, pages 267–292. CRC Press, Boca Raton, 2010.

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

M. Constant, G. Eryiğit, J. Monti, L. Van Der Plas, C. Ramisch, M. Rosner, and A. Todirascu. Multiword expression processing: A survey. Computational Linguistics, 43(4):837–892, 2017.

A.-M. Di Sciullo and E. Williams. On the definition of word, volume 14. Springer, 1987.

H. J. J. Dyvik, G. S. Losnegaard, and V. Rosén. Multiword expressions in an lfg grammar for norwegian. 2019.

		DE	EL	FR	HE	HI	IT	PL	PT	SV	TR	ZH
arithmetic (%)	form deprel	66.11	58.75	68.11	**37.79**	61.15	54.48	70.48	62.72	67.35	50.75	**47.89**
	lemma deprel	**72.38**	60.00	**74.84**	15.70	**69.38**	**64.07**	**83.07**	**74.83**	76.68	**61.30**	32.94
	lemma deprel upos	71.80	**60.12**	74.75	21.49	69.19	**64.07**	81.94	74.70	**77.20**	61.11	34.52
harmonique (%)	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
geometric (%)	form deprel	61.74	54.80	65.15	**35.13**	53.72	51.06	65.79	55.74	61.83	44.29	**47.40**
	lemma deprel	**70.71**	59.86	**74.75**	10.85	**67.05**	**64.03**	**82.32**	**73.84**	75.94	**61.19**	22.08
	lemma deprel upos	69.83	**59.96**	74.65	20.91	66.83	**64.03**	80.99	73.61	**76.20**	60.96	26.73

Table 3: Precision and Recall's arithmetic, harmonic, and geometric means of selected λ-CSS based lexicon

	DE	EL	FR	HE	HI	IT	PL	PT	SV	TR	ZH
MTLB-STRUCT	76.17	**72.62**	**79.42**	**48.30**	**73.62**	**63.76**	81.02	73.34	71.58	69.46	**69.63**
inter	56.61	51.80	62.16	36.19	46.89	47.43	60.48	46.29	52.75	36.69	55.91
union	**76.45**	71.12	78.87	44.29	73.29	62.92	81.41	**74.76**	73.74	**69.92**	58.43
$\{lemma, deprel\}$-lexicon	69.07	59.71	74.65	7.50	64.80	64.00	**81.58**	72.86	**75.21**	61.08	14.81

Table 4: F-score (%) of MTLB-STRUCT, our lexicon, and the union and intersection of their predictions.

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

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

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

C. Ramisch, S. Cordeiro, A. Savary, V. Vincze, V. Mititelu, A. Bhatia, M. Buljan, M. Candito, P. Gantar, V. Giouli, et al. 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. Association for Computational Linguistics, 2018.

C. Ramisch, A. Savary, B. Guillaume, J. Waszczuk, M. Candito, A. Vaidya, V. B. Mititelu, A. Bhatia, U. Iñurrieta, V. Giouli, et al. 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, 2020.

A. Savary, C. Ramisch, S. R. Cordeiro, F. Sangati, V. Vincze, B. Qasemi Zadeh, M. Candito, F. Cap, V. Giouli, I. Stoyanova, et al. The parseme shared task on automatic identification of verbal multiword expressions. In The 13th Workshop on Multiword Expression at EACL, pages 31–47, 2017.

A. Savary, M. Candito, V. B. Mititelu, E. Bejček, F. Cap, S. Čéplö, S. R. Cordeiro, G. C. Eryiğit, V. Giouli, M. van Gompel, et al. Parseme multilingual corpus of verbal multiword expressions. In Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop, 2018a.

A. Savary, S. Petitjean, T. Lichte, L. Kallmeyer, and J. Waszczuk. Object-oriented lexical encoding of multiword expressions: Short and sweet. arXiv preprint arXiv:1810.09947, 2018b.

A. Savary, S. Cordeiro, T. Lichte, C. Ramisch, U. Iñurrieta, and V. Giouli. Literal occurrences of multiword expressions: rare birds that cause a stir. The Prague Bulletin of Mathematical Linguistics, 2019a.

A. Savary, S. Cordeiro, and C. Ramisch. 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), 2019b.

	DE	EL	FR	HE	HI	IT	PL	PT	SV	TR	ZH
MTLB-STRUCT.open	76.17	72.62	79.42	48.30	73.62	63.76	81.02	73.34	71.58	69.46	69.63
union	**89.22**	**87.44**	**89.99**	**81.36**	**86.38**	**85.65**	**92.16**	**88.21**	**85.04**	**85.42**	**78.77**
$\{lemma, deprel\}$-lexicon	96.97	98.21	94.07	73.41	95.45	92.58	98.49	97.82	94.77	92.48	31.20

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

S. Taslimipoor, S. Bahaadini, and E. Kochmar. Mtlb-struct@ parseme 2020: Capturing unseen multi-word expressions using multi-task learning and pre-trained masked language models. arXiv preprint arXiv:2011.02541, 2020.

A. Villavicencio, A. Copestake, B. Waldron, and F. Lambeau. Lexical encoding of mwes. In Proceedings of the Workshop on Multiword Expressions: Integrating Processing, pages 80–87, 2004.