Literal readings of multiword expressions: as scarce as hen's teeth

Agata Savary, Silvio Cordeiro

Literal readings of multiword expressions: as scarce as hen's teeth

Agata Savary
Université François Rabelais Tours, France
University of Düsseldorf, Germany
agata.savary@univ-tours.fr

Silvio Ricardo Cordeiro
Aix-Marseille Université, France
silvio.cordeiro@lif.univ-mrs.fr

Abstract

Multiword expressions can have both idiomatic and literal occurrences. Distinguishing these two cases is considered one of the major challenges in MWE processing. We suggest that literal readings should be considered in both semantic and syntactic terms, which motivates their study in a treebank. We propose heuristics to automatically pre-identify candidate sentences that might contain literal readings of verbal VMWEs, and we apply them to an existing Polish treebank. We also perform a linguistic study of the literal readings extracted by the different heuristics. The results suggest that literal readings constitute a rare phenomenon. We also identify some properties that may distinguish them from their idiomatic counterparts.

1 Introduction

Multiword expressions (MWEs) are word combinations, such as ***all of a sudden***, *a* ***hot dog***, *to* ***pay*** *a* ***visit*** or *to* ***pull one's leg***, which exhibit lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasies. They encompass closely related linguistic objects such as idioms, compounds, light verb constructions, rhetorical figures, institutionalised phrases or named entities. A prominent feature of many MWEs, especially of verbal idioms such as *to* ***pull one's leg***, is their non-compositional semantics, i.e. the fact that their meaning cannot be deduced from the meanings of their components, and from their syntactic structure, in a way deemed regular for the given language. For this reason, MWEs pose special challenges both to linguistic modeling (e.g. as linguistic objects crossing boundaries between lexicon and grammar) and to Natural Language Processing (NLP) applications, especially those which rely on semantic interpretation of text (e.g. information retrieval, information extraction or machine translation).

Another challenging property of many MWEs, as in example (1), is that we can encounter their literally understood counterparts, as in (2). However, it is not clear what should be considered an occurrence of a literal reading of an MWE. Should "coincidental" co-occurrences of its lexicalized components,[1] like in (4) as opposed to (3), also be considered its literal occurrences? Should variants like (6), which considerably change the "canonical" syntactic dependencies between the components, compared to (5), still be considered idiomatic occurrences? Finally, what should be the status of word plays which deliberately refer to both the idiomatic and the literal reading of an MWE, as in (7)?

(1) *The man was* ***pulling my leg*** *but I didn't believe him.*

(2) *The kid was pulling my leg to make me play with him.*

(3) *The preparations were not thoroughly planned* ***after all****.*

(4) *After all the preparations we finally left.*

(5) *The Samsung boss can still* ***pull*** *the* ***strings*** *from prison.*

(6) *The article addresses the political* ***strings*** *which the journalist claimed that the senator* ***pulled****.*

(7) (Polish) *Wyciągnięcie rąk uchroniło go od* ***wyciągnięcia nóg*** *'Stretching hands prevented him from stretching legs'* ⇒ STRETCHING HIS HANDS PREVENTED HIM FROM DYING

For a given MWE E with lexicalized components $e_1, \ldots, e_n$, we define its *literal reading occurrence*, or *literal reading* (LR) for short, as a co-occurrence of the lexemes $e_1, \ldots, e_n$ in a context in which:

[1] The lexicalized components of an MWE are those which are always realized by the same lexeme. For instance in *to* ***pay*** *a* ***visit*** the head verb is always a form of *pay* but the determiner *a* can be freely replaced, as in ***paid*** *many* ***visits***. In this paper the lexicalized components of MWEs are highlighted in boldface.

Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 64–72,
Prague, Czech Republic, January 23–24, 2018.

Keywords: multiword expressions, literal reading, idiomaticity rate, treebank, Polish

(i) it is not a MWE; and (ii) one of the typical senses of each of $e_1, \ldots, e_n$ is activated; and (iii) the syntactic constraints among $e_1, \ldots, e_n$ are preserved, i.e. either the same or equivalent dependencies hold between E's components as in its canonical (citation) form. Dependencies are equivalent if the syntactic variation can be neutralized while preserving the overall meaning. For instance (6) can be reformulated into: *The journalist claimed that the senator* ***pulled*** *political* ***strings****, and this article addresses them.* Therefore, the syntactic constraints between $e_1 = pull$ and $e_2 = strings$ visible in (5) are preserved in (6). According to this definition, only example (2) above is considered an LR.[2] Example (4) does not fulfill condition (iii), while (1), (3), (5) and (6) do not fulfill (i-ii), since their are idiomatic readings (IRs). In example (7), the expression *wyciągnięcie rąk* STRETCHING HANDS points to a typical meaning of the verb *wyciągnąć* STRETCH. By analogy, the reader is also induced to think of a literal meaning of the noun *nogi* LEGS. However, the idiomatic meaning of *wyciągnięcie nóg* '*stretching legs*' ⇒ DYING is still intact and thus it fails condition (i). Note that, due to the presence of condition (iii), the study of literal readings of MWEs is best done in a treebank.

The motivation to study the phenomenon of LRs of MWEs, and of its frequency in particular, is both of linguistic and of computational nature. Firstly, psycholinguistic studies put special interest in the interplay between LRs and IRs, as well as their distributional and statistical properties, when discovering how idioms are stored and processed in human mind (Cacciari and Corradini, 2015). Secondly, the links between LRs and IRs readings can inform us which morpho-syntactic variation is allowed or prohibited by some MWEs, and why (Sheinfux et al., 2017; Pausé, 2017). Additionally, an opposition of the contexts in which LRs and IRs readings occur may yield better methods to automatically distinguish them (Peng et al., 2014; Peng and Feldman, 2016).

This last task is considered one of the major challenges in automatic processing of MWEs (Constant et al., 2017). Its quantitative importance can be estimated by measuring the *idiomaticity rate*, i.e. the ratio of occurrences of an MWE with idiomatic reading to both its idiomatic and literal occurrences in a corpus (El Maarouf and Oakes, 2015). If the overall (i.e. aggregated for all MWEs) idiomaticity rate is relatively low, distinguishing IRs and LRs readings becomes, indeed, a major challenge, as claimed by Fazly et al. (2009). If, conversely, it is high, or even close to 100%, the task can be neglected for many applications. Also, as shown by (Waszczuk et al., 2016), a high idiomaticity rate can considerably speed up parsing, if appropriately taken into account by a parser's architecture.

In this paper we are interested in verbal MWEs (VMWEs), in which syntactic flexibility can be particularly rich. We exploit an existing multilingual corpus (Savary et al., 2017) in which VMWE annotations are accompanied by morphological and dependency annotations, but literal occurrences are not tagged (Sec. 2). We propose several heuristics to automatically detect possible literal occurrences of known, i.e. manually annotated, VMWEs (Sec. 3). Then we manually categorize the resulting occurrences using a typology which accounts for true and false positives, as well as for linguistic properties of LRs as opposed to IRs (Sec. 5). We report on results in a Slavic language: Polish (Sec. 5). Finally, we conclude and discuss perspectives for future work (Sec. 6).

2 Corpus

We use the openly available PARSEME corpus[3] manually annotated for VMWEs in 18 languages (Savary et al., 2017). Among its 5 VMWE categories, three are relevant to this Polish-dedicated study:

- *Idioms* (IDs) are verbal phrases of various syntactic structures, mostly characterized by non-compositional meaning, as in (8). Due to the fact that many idioms were conceived as metaphors, they maintain a large potential of LRs, as exemplified in (9).
  (8) *dawno już powinien był* ***wyciągnąć nogi*** '*long-ago already should-he have stretched legs*' ⇒ HE SHOULD HAVE DIED LONG AGO
  (9) *położyłem się na trawie i wyciągnąłem nogi* '*I-lay-down on the-grass and stretched legs*'
- *Light-verb constructions* (LVCs) are *VERB (PREP) (DET) NOUN* combinations in which the verb V is semantically void and the noun N is a predicate expressing an event or a state, as in (10). The

[2]Henceforth, we use wavy and dashed underlining for true and false LRs, respectively. Straight underlining denotes focus.
[3]http://hdl.handle.net/11372/LRT-2282

idiomatic nature of LVCs lies in the fact that the verb may be lexically constrained and does not contribute any semantics to the whole expression. LVCs are mostly semantically compositional, therefore the notion of a LR is less intuitively motivated for them. A LR of an LVC should be understood as a co-occurrence of its lexemes which does not have all the required LVC properties. This occurs, for instance, when N is not predicative or does not express and event or a state, as in (11), where *udziały* '*shares*' denotes an amount of financial assets. Figures 1a and 1b present another occurrence of this VMWE, and of its LR, respectively.

(10) ***mieć** swój **udział** w debacie* '*have one's share in debate*' ⇒ TO TAKE PART IN THE DEBATE

(11) *mieć udziały w spółce* '*have shares in company*' ⇒ TO HAVE SHARES IN A COMPANY

- *Inherently reflexive verbs* (IReflVs), pervasive in Romance and Slavic languages but not in English, are combinations of a verb V and a reflexive clitic $RCLI$, such that one of the 3 non-compositionality conditions holds: (i) V never occurs without $RCLI$ as is the case for the VMWE in (12); (ii) $RCLI$ distinctly changes the meaning of V, like in (13); (iii) $RCLI$ changes the subcategorization frame of V, like in (15) as opposed to (16). IReflVs are semantically non-compositional in the sense that $RCLI$ is not an argument of the verb. LRs never occur for type (i) but they do occur for types (ii) and (iii), due to homonymy with compositional V-RCLI combinations which express true reflexive or reciprocal meanings, as in (14), or impersonal or middle passive alternation, as in (17).

(12) ***bał się** wody* '*feared RCLI water*' ⇒ HE WAS AFRAID OF WATER

(13) *nie **oglądaj się** na innych* '*not watch RCLI on others*' ⇒ DO NOT COUNT ON THE OTHERS

(14) *oglądam się w lustrze* '*I-am-watching myself in the mirror*'

(15) ***spotykać się** z przyjaciółmi* '*meet RCLI with friends.INST*' ⇒ MEET FRIENDS

(16) *spotykać przyjaciół* '*to meet friends.ACC*'

(17) *nie spotyka się takich ludzi* '*not meets RCLI such people*' ⇒ SUCH PEOPLE ARE NEVER MET

The Polish part of the training corpus contains 11,578 sentences, for a total of 191,239 tokens and 3,149 annotated instances of VMWEs.[4] For most languages, including Polish, the VMWE annotation layer is accompanied by morphological and syntactic layers (ML and SL, respectively), as shown in Fig. 1a and 1b. In ML, a lemma, a POS and morphological features are assigned to each token. SL represents syntactic dependencies between tokens. For Polish, both ML and SL use the Universal Dependencies (UDs) tagsets.[5] ML was created partly manually and partly automatically, and SL automatically, using UDPipe[6] with its pre-trained Polish model. While the PARSEME corpus is manually annotated and categorized for IRs of VMWEs, it is not annotated for their LRs. Therefore, we developed several heuristics which allow us to identify them automatically.

3 Identifying literal readings

We use no external resources, therefore we can only identify LRs for VMWEs which are annotated at least once in the corpus. In order to fully reliably perform this task, we would have to ensure that conditions (i), (ii) and (iii) from page 1 hold. Condition (i) can be automatically fulfilled by discarding predictions that coincide with annotated VMWEs. Condition (ii) cannot be checked automatically, given that the available annotation layers do not account for semantics. It must, thus, be subject to manual verification. Condition (iii) is closely linked to the SL annotations but checking it fully reliably can be hindered by at least two factors. Firstly, some dependency annotations in SL can be incorrect, especially if SL was constructed automatically. Secondly, defining conditions under which two sets of dependency relations are equivalent seems challenging and highly language-dependent. Given the large number of possible syntactic structures of VMWEs, an exhaustive catalog of such equivalences would be huge, or

[4]The annotation was performed by a single native Polish annotator. The inter-annotator agreement (IAA) in VMWE identification was measured in terms of the F-measure and κ, with the scores of 0.529 and 0.434, respectively. The IAA in VMWE categorisation (based on the VMWE identified jointly by two annotators) assessed in terms of the F-measure, and equal to 0.939. All IAA scores were based on a small sample of the corpus, anotated in parallel by another Polish speaker who only had few experience with the guidelines and did not annotate the final corpus. Therefore, these IAA scores are rather weak indicators of the annotation quality.

[5]http://universaldependencies.org/guidelines.html

[6]https://ufal.mff.cuni.cz/udpipe

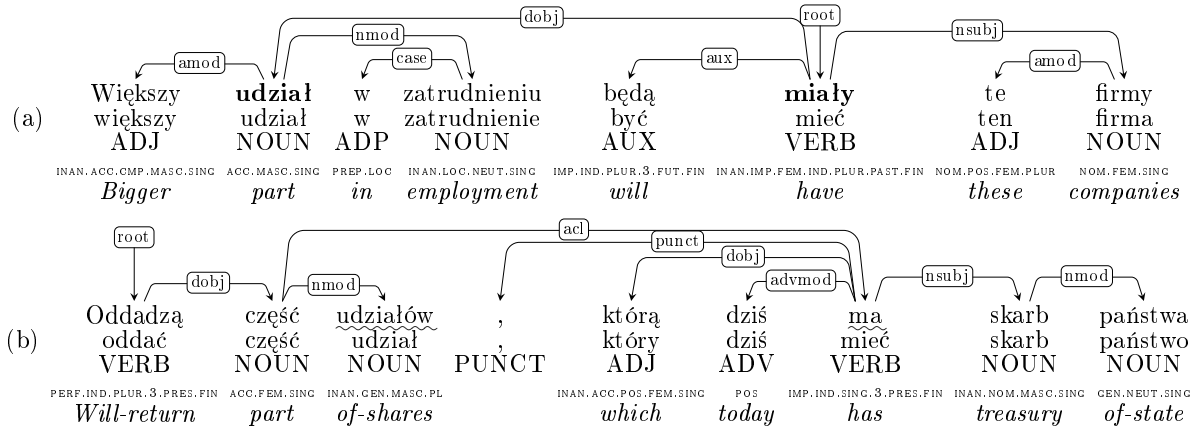

Figure 1: Morphosyntactic annotations for an occurrence context of the VMWE ***mieć udział*** '*have share*' ⇒ TAKE PART (a) and its LR (b). Translations: (a) *These companies will participate in employment more intensively.* (b) *They will return the part of the shares that the treasury has today.*

even potentially infinite, due to long-distance dependencies in recursively embedded relative clauses, as illustrated in example (6). In order to cope with these obstacles, we designed four heuristics which should cover a large majority of LRs in complementary ways, while maintaining the amount of false positives relatively low (i.e. the heuristics are skewed towards high recall). They rely on the following definitions.

Each *sequence* of words is a function $s : \{1, 2, \ldots, |s|\} \to W$, where W are word forms. The sequence s can be noted as $s := \{s_1, s_2, \ldots s_{|s|}\}$, where $s_i := (i, w_i)$ is a single *token*. A sequence can thus be denoted as a set of pairs: $s = \{(1, w_1), (2, w_2), \ldots, (|s|, w_{|s|})\}$. For example, the sentence in Fig. 1a can be represented as a sequence $s = \{(1, \text{Większy}), (2, \text{udział}), \ldots, (8, \text{firmy})\}$. For a given token $s_i = (i, w_i)$, $\text{lemma}(s_i)$ is its case-folded lemma form (or nil if unavailable in ML), and $\text{surface}(s_i)$ is its case-folded surface form. For instance in Fig. 1a, $\text{lemma}(s_6) = \text{mieć}$, $\text{surface}(s_6) = \text{miały}$, and $\text{surface}(s_1) = \text{większy}$. As not every token may have lemma information, we define $\text{lemmasurface}(s_i)$ as the lemma if available, and as the surface form otherwise. If s is a sentence, each token s_i is associated with its parent, denoted as $\text{parent}(s_i)$, through a syntactic label, denoted as $\text{label}(s_i)$. Some tokens may have parent nil (and label root). In Fig. 1a, $\text{label}(s_2) = \text{dobj}$, $\text{parent}(s_2) = s_6$, $\text{label}(s_6) = \text{root}$, and $\text{parent}(s_6) = \text{nil}$. For a given sequence s, its *subsequence* q is an injection defined as an order-preserving sequence over tokens of s, i.e. $q : \{1, 2, \ldots, |q|\} \to s$ such that, if $i < j$, $q(i) = s_k$ and $q(j) = s_l$, then $k < l$. The definitions of lemmas and surface forms extend straightforwardly to tokens of a subsequence: $\text{lemma}((i, s_k)) := \text{lemma}(s_k)$ and $\text{surface}((i, s_k)) := \text{surface}(s_k)$. For instance in Fig. 1a, the subsequence corresponding to the tokens in bold can be formalized as $q = \{(1, s_2), (2, s_6)\} = \{(1, (2, \text{udział})), (2, (6, \text{miały}))\}$, and $\text{lemma}(q_2) = \text{lemma}((2, s_6)) = \text{lemma}(s_6) = \text{mieć}$, etc.

In a subsequence q, the definition of a parent still relies on the dependencies in the underlying sequence s but is restricted to the tokens in q. Formally, for a given $1 \leqslant i \leqslant |q|$, if there exists $1 \leqslant j \leqslant |q|$ such that $\text{parent}(q(i)) = q(j)$, then $\text{parent}_{sub}(q_i) := q_j$. Otherwise $\text{parent}_{sub}(q_i) := nil$. For instance in Fig. 1a, $q_1 = (1, s_2)$, $q_2 = (2, s_6)$, $\text{parent}_{sub}(q_1) = q_2$ and $\text{parent}_{sub}(q_2) = \text{nil}$. In Fig. 1b, where the subsequence consisting of the underlined tokens forms a non-connected graph, the parents of both components are nil, i.e. $q_1 = (1, s_3)$, $q_2 = (2, s_7)$, and $\text{parent}_{sub}(q_1) = \text{parent}_{sub}(q_2) = \text{nil}$.

In the pre-processing step we extract each occurrence of an annotated VMWE in a sentence s as a subsequence of s, noted $m = \{m_1, m_2, \ldots, m_{|m|}\}$. For each known VMWE m extracted in this way, and for each sentence $s' = \{s'_1, s'_2, \ldots, s'_{|S|}\}$, we then look for literal matches of m in s'. We define a *literal match* as an injection $\phi : m \to s'$, where for every $t \in m$, we have $\text{lemmasurface}(t) \in \{\text{lemma}(\phi(t)), \text{surface}(\phi(t))\}$, and the image of m is not annotated as a VMWE itself. For instance, for the VMWE $m = \{(1, s_2), (2, s_6)\}$ from Fig. 1a, we obtain the following literal match in the sentence from Fig. 1b: $\phi = \{((1, s_2), s'_3), ((2, s_6), s'_7)\}$. The set of such bijections can be huge and include a large number of false positives, i.e. coincidental co-occurrences of m's components in the same sentence.

Figure 2: True and false LRs of *mieć udział* '*have share*' ⇒ TAKE PART, with extracts of SL.

Therefore, we restrain the set of such injections with the following criteria.

- **WindowGap**: Under this criterion, all matched tokens must fit into a sliding window with no more than g external elements. Formally, let J be the set of all matched indexes in the sentence s', i.e. $J = \{\, j \mid m_i \in m,\ s'_j = \phi(m_i) \,\}$. Then ϕ is only considered to match if $\max(J) - \min(J) + 1 \leqslant g + |m|$. For m in Fig. 1a and s' in Fig. 1b we have $J = \{3, 7\}$ and $|m| = 2$. Thus, the tokens corresponding to *udziałów ma* are a literal match only if $g \geqslant 3$. In the case of Fig. 2, every reading can be matched with $g \geqslant 2$.
- **BagOfDeps**: Under this criterion, a literal match must be a connected graph, but the directions and the labels of the dependencies are ignored. Formally, there must be a token $m_{\text{root}} \in m$ for which $\text{parent}(m_{\text{root}}) = \text{nil}$. Moreover, for every token $m_i \in m \setminus \{m_{\text{root}}\}$, there exists a token $m_k \in m$ such that $\text{parent}(\phi(m_i)) = \phi(m_k)$. For instance, the readings in Fig. 2a, 2b and 2d are matched under this criterion, but not those in Fig. 2c and Fig. 1b.
- **UnlabeledDeps**: Under this criterion, a literal match must be a connected directed graph in which the dependency labels are ignored but the parent relations are preserved. Formally, this criterion adds a restriction to BagOfDeps: m_k must be such that $m_k = \text{parent}_{sub}(m_i)$. For instance, the readings in Fig. 2b and 2d are matched under this criterion, but not those in Fig. 2a, 2c and Fig. 1b.
- **LabeledDeps**: Under this criterion, a literal match must be a connected directed graph in which both the parent relations and the dependency labels are preserved. Formally, this criterion adds a restriction to UnlabeledDeps: For every $m_i \in m \setminus \{m_{\text{root}}\}$, we must have $\text{label}(m_i) = \text{label}(\phi(m_i))$. Only the reading in Fig. 2b is matched under this criterion.

4 Results

The above heuristics, which are language-independent, were used to automatically pre-select LR candidates of VMWEs occurring in the training part of the Polish PARSEME corpus. For each of the 3,149 annotated VMWE instances, each of the four heuristics (with $g = 2$)[7] was used to extract literal matches, their POS sequences and the sentences in which they occur. We then performed a manual tagging of each LR candidate.[8] Out of the resulting 416 literal matches, 72 (17.3%) were manually tagged as true LRs, i.e. conforming to the definition in Sec. 1. These 72 occurrences correspond to 32 distinct VMWEs. The remaining 344 matches were due to one of these 3 reasons: (i) coincidental co-occurrences of VMWE components, as in example (4) and Fig. 2c–d, (ii) true VMWEs, wrongly omitted in the original annotation (29 such cases were detected), (iii) false VMWEs, which should have never been annotated (8 occurrences of 3 such expressions were detected).

Tab. 1 shows the per-category and the overall efficiency of the four heuristics from Sec. 3 in the task of finding LRs of VMWEs (the best results are highlighted in bold).[9] The overall F-scores (even if more than twice better for IDs than for other categories) indicate that automatic identification of LRs is a hard task. Obviously, mixing all heuristics gives optimal recall (since only those occurrences which were extracted by at least one of them are examined here). In particular, WindowGap and BagOfDeps are

[7]The average length of a gap in a VMWE in the Polish PARSEME corpus is equal to 0.53 and its mean absolute deviation (MAD) is equal to 0.77. Since the LRs had not been manually annotated, analogous data for the gaps contained in LRs were not available in advance. But when the LRs identified in this study (see below) are concerned, the average length of a gap and its MAD are equal to 1.1 and 1.2 respectively.

[8]One Polish native speaker, a co-author of this paper, participated in this task. She was also the main annotator of the VMWE layer in the Polish PARSEME corpus.

[9]Matches due to errors in the VMWE annotations were kept in Tab. 1. Correcting these errors would require a re-execution of the heuristics, which could bias our evaluation towards the underlying tool.

Category	WindowGap			BagOfDeps			UnlabeledDeps			LabeledDeps			All		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
ID	0.41	0.88	0.56	**0.50**	0.19	0.27	0.67	0.13	0.21	n/a	0.00	n/a	0.43	**1.00**	**0.60**
IReflV	0.15	1.00	**0.26**	0.14	0.63	0.23	**0.15**	0.63	0.25	0.14	0.37	0.20	0.13	**1.00**	0.23
LVC	0.17	0.73	0.28	**0.20**	0.65	**0.30**	0.15	0.38	0.21	0.14	0.19	0.16	0.17	**1.00**	0.29
ALL	**0.18**	0.88	**0.30**	0.17	0.54	0.26	0.16	0.43	0.23	0.14	0.22	0.17	0.17	**1.00**	0.30

Table 1: Precision, recall and F-measure of the four heuristics.

largely complementary: only 41,7% of LRs are extracted by both of these methods. Also expectedly, the WindowGap method outperforms each other individual method as far as recall is concerned. It also has optimal overall scores, even if it remains behind BagOfDeps and UnlabeledDep in precision for individual VMWE categories. Not surprisingly, the recall of the BagOfDeps is systematically higher than the recall of UnlabeledDeps, which in turn is systematically higher than the recall of LabeledDeps – since these heuristics rely on increasing degrees of syntactic constraints. However, this does not result in higher precision scores. To the contrary: BagOfDeps has the best precision of the three methods. This phenomenon may be partially explained by the presence of errors in SL.

All the results shown here rely on a maximum-coverage hypothesis (MCH), i.e. the assumption that the four heuristics, with $g = 2$, allow us to extract all LRs of the previously annotated VMWEs. This hypothesis is strong. Potentially, there could be a LR whose components have a gap longer than 2, and which was not extracted e.g. due to non-connectivity in the dependency graph as in Fig. 1b, or due to an error in the SL. Ideally, we should, thus, examine all co-occurrences of the lexicalized components of a given VMWE, whatever their distance in the sentence. However, we estimate that this would triple the number of exact matches and require a much higher manual annotation effort. We thus performed a less labor-intensive experiment to assess the reliability of MCH. We applied the WindowGap heuristic with a gap length of 9,999 (which exceeds all sentence lengths in the corpus) to the first 1,000 sentences of the corpus, which yielded 41 literal matches. Then, the matches previously seen (i.e. extracted by any of the four previously used heuristics) were eliminated, and the resulting 30 occurrences were manually labeled according to the same scenario as above. All of them were false positives, which suggests that the four heuristics would hardly ever miss any LRs among their literal matches.

As seen in Sec. 3, our heuristics are skewed towards high recall, which makes them practical for pre-identifying and manually validating LR candidates, but not optimal for automatic classification of IRs and LRs. Previous methods proposed for the latter task include (Fazly et al., 2009), where unsupervised MWE identification is based on statistical measures of lexical and syntactic flexibility of MWEs. The notion of a LR seems to have a much larger scope than in our approach: it notably includes variants stemming from replacement of lexicalized components by automatically extracted similar words, e.g. *spill corn* vs. ***spill** the **beans***. The test data are restricted to the 28 most frequent verb-object pairs, and their manually validated IRs and LRs, i.e. accidental co-occurrences of the MWE components are excluded from performance measures (unlike in our approach). Their precision and recall in LR identification range from 0.18 to 0.86 and from 0.11 to 0.61, respectively. These results are hard to compare to Tab. 1, due to the very different understanding of the task and its experimental settings.

5 Corpus study

Given the manually identified true LRs, we can estimate the idiomaticity rate ($IdRate$) as follows:

$$IdRate_{CAT} = \frac{|IR_{CAT}|}{|LR_{CAT}| + |IR_{CAT}|} \tag{18}$$

where IR_{CAT} is the set of (idiomatic) VMWE occurrences of category CAT[10], LR_{CAT} is the set of true LRs of VMWEs of category CAT, and $CAT \in \{\text{ID}, \text{IReflV}, \text{LVC}, \text{ALL}\}$. As shown in Tab. 2, LRs of VMWEs in Polish are rare: the overall $IdRate$ amounts to 0.978. This score is consistent with

[10]This number was updated by accounting for the VMWE annotation errors identified during the manual validation (cf. Sec. 4).

(Waszczuk et al., 2016), where the $IdRate$ of Polish verbal, nominal, adjectival and adverbial MWEs is estimated at 0.95. It is, however, in sharp contrast to (Fazly et al., 2009), where the proportion of LRs of the most frequent English verb-object MWEs was estimated at 40%. This is probably due to the different understanding of LRs by these authors, and their relatively restricted experimental scope (cf. Sec. 4). Important cross-language factors might also influence the $IdRate$, such as the pervasiveness of lexicalized determiners like *the/a* in Germanic and Romance languages vs. the lack of their equivalents in Slavic ones.

Tab. 2 also shows the per-category $IdRate$. Many IDs originated as metaphors, and this is reflected in the fact that IDs have the lowest $IdRate$, even if only slightly lower than other categories. IReflVs, conversely, have the highest $IdRate$, despite homonymy, shown in examples (14) and (17).

Category	# LRs		# IRs		IdRate
	tokens	types	tokens	types	
ID	16	5	322	219	0.953
IReflV	30	19	1547	368	0.981
LVC	26	8	1301	662	0.980
ALL	72	32	3170	1249	0.978

Table 2: Idiomaticity rate per VMWE category and overall.

Category	MORPH		SYNT		OTHER	
	tokens	types	tokens	types	tokens	types
ID	7	3	8	2	1	1
IReflV	8	3	1	1	21	16
LVC	18	2	2	1	6	5
ALL	(46%) 33	8	(15%) 11	4	28	22

Table 3: LRs distinguishable from VMWEs by constraints of various types

A close-up study of the 32 distinct VMWEs corresponding to the 72 LR tokens reveals that their individual $IdRate$ varies greatly: from 0.20 for ***daje się*** *(zauważyć X)* '*allows RCLI (notice X)*' ⇒ IT IS POSSIBLE (TO NOTICE X) to 0.94 for ***czuć się*** *(dobrze)* '*feel RCLI (well)*' ⇒ TO FEEL (WELL).

In view of automatically distinguishing LRs from IRs, we studied the morphological and syntactic constraints imposed by VMWEs. We manually tagged the 72 LRs with one of the following labels:

- **MORPH**: the LR does not respect the morphological constraints imposed by the corresponding VMWE on one of its lexicalized components. For instance, the VMWE in example (10) requires the nominal component *udział* '*share*' to occur in singular. If this constraint were known, the occurrence in (11) could be automatically classified as literal. Morphological constraints can also concern the head verb, e.g. the VMWE in (19) allows no overt subject and restricts the finite forms of its head verb *dać* '*allow*' to 3rd person singular. Knowing this constraint would allow us to automatically identify (20), where the verb is inflected in 2nd person imperative, as an LR.
- **SYNT**: the LR violates the syntactic constraints – other than the dependencies between its lexicalized components – imposed by the VMWE. This typically concerns dependencies between lexicalized components and external arguments or adjuncts. E.g., while the VMWE in (19) admits no overt subject, the LR in (21) does take a subject *pięćdziesięciolatka* '*50-year-old-woman*'. Also, the VMWE from (22) requires an infinitive complement and its noun *stan* '*state*' allows no modifier. If this constraint were known, the dependent of this noun in (23) would automatically imply a LRs.
- **OTHER**: in order to distinguish an LR from IRs, more advanced (e.g. semantic) constraints would have to be verifiable. E.g., an LVC with the light verb *mieć* '*to have*' in present tense and occurring under the scope of negation, as in (24), is homonymic with the existential *być* '*to be*', whose negation in present tense is realized in Polish precisely by *mieć* '*to have*', as in (25). Since Polish is a pro-drop language, the subject in (24) can be skipped, which makes both occurrences look identical. Also, IReflVs like in (26) are polysemic with reflexive, reciprocal, impersonal or middle alternation uses, as in (27), and divergences in syntactic constraints are inexistent or unverifiable (e.g. due to dropped arguments). Only powerful pragmatic mechanisms would allow these cases to be distinguished.

(19) *dokładnich kwot nie **da się** wyliczyć* '*exact amounts not allows.3.SING.FIN.PRES RCLI calculate*' ⇒ THE EXACT AMOUNTS CANNOT BE CALCULATED

(20) *nie daj się zbywać ogólnikami* '*not allow.2.SING.IMPER RCLI dispose-of with-commonplaces*' ⇒ DON'T BE DISPOSED OF WITH COMMONPLACES

(21) *Pięćdziesięciolatka nie da się na to złapać* '*50-year-old-woman not allows.3.SING.FIN.FUT RCLI on this catch*' ⇒ A 50-YEAR-OLD WOMAN WILL NOT FALL INTO THIS TRAP

(22) *więcej nie **jestem w stanie** dokonać* '*more not am in state to-do*' ⇒ I AM NOT ABLE TO DO MORE

(23) *trzech żołnierzy było w stanie krytycznym* '*three soldiers were in state critical*' ⇒ THREE SOLDIERS WERE IN A CRITICAL STATE

(24) *(klient) nie **ma powodów** do satysfakcji* '*(client) not has reasons for satisfaction*' ⇒ (THE CLIENT) HAS NO REASONS TO BE SATISFIED

(25) *nie ma powodów do satysfakcji* '*not has reasons for satisfaction*' ⇒ THERE ARE NO REASONS TO BE SATISFIED

(26) *kadydaci **znaleźli się** w trudnej sytuacji* '*candidates found RCLI in hard situation*' ⇒ THE CANDIDATES FOUND THEMSELVES IS A DIFFICULT SITUATION

(27) *kadydaci znaleźli się dopiero po tygodniu* '*candidates found RCLI only after week*' ⇒ CANDIDATES WERE FOUND ONLY A WEEK LATER

As shown in Tab. 3, 61% of the LRs can be automatically distinguished in the treebank from IRs if morphological and syntactic constraints imposed by VMWEs are known, e.g. encoded in a lexical resource (Przepiórkowski et al., 2017) or learned from a corpus. The remaining 39% of LRs call for powerful mechanisms which go beyond sentence boundaries and most lexical encoding frameworks. Note also that the percentage of the VMWE types which exhibit any literal readings is relatively low (32 types out of 1249, i.e. 2.6%). This suggests that methods for MWE identification might benefit from language-specific components explicitly targeting those few expressions.

6 Conclusions and future work

The main contribution of this paper is a close examination of several aspects of literal readings (LRs) of VMWEs. Firstly, we defined the notion of an LR in terms of both the semantics of their components, and of their syntactic dependencies, which motivates their study in a treebank. We proposed four language-independent heuristics, oriented towards high recall and a reasonable precision, for the task of automatically identifying LRs, given their manually performed annotations in a treebank. We applied these heuristics to Polish data stemming from a multilingual corpus annotated for VMWEs following universal guidelines, and we manually validated the extracted LR candidates. The resulting dataset, available under an open license[11], allowed us to show that automatic identification of LRs is a hard task, especially when syntactic annotations are created automatically. We also discovered that up to 61% of the LRs can be automatically distinguished from their idiomatic counterparts if data on morphological and syntactic constraints imposed by VMWEs are available (e.g. lexically encoded or learned from a corpus). Last but not least, we showed that LRs are relatively rare in Polish: the idiomaticity rate of VMWEs is equal to 0.978, and only 2.6% of all VMWE types exhibit literal readings in our corpus.

The proposed heuristics can also be used as part of MWE annotation methods. In the context of PARSEME, a similar tool was used to check the consistency of VMWE annotations in the corpus, and to detect VMWE occurrences that were possibly missed during the annotation phase.

Future work could investigate the extent to which the results from the different heuristics are statistically significant. The heuristics could also be extended to handle long-distance dependencies such as the one in (6). We also plan to apply this study to other languages from various languages families, concerned by the PARSEME corpus, so as to check the discovered tendencies. Preliminary studies in Portuguese show that the definition of an LR needs enhancements: not only the syntactic dependencies between the lexicalized components are to be preserved but also their POS. This condition is necessary to avoid ambiguities, notably between the reflexive pronoun *se* '*RCLI*' in IReflVs and the conjunction *se* '*if*'. Further enhancement, useful for Slavic languages, might consist in merging aspectual pairs (perfective/imperfective) of VMWEs such as ***da się*** '*let.PERF RCLI*' ⇒ IT WILL BE POSSIBLE (TO) vs. ***daje się*** '*let.IMP RCLI*' ⇒ IT IS POSSIBLE (TO). Finally, the findings on LRs may enhance MWE identification methods. They may for instance yield useful hints for feature engineering, or may be used in a post-processing step to eliminate LRs wrongly recognized as variants of VMWEs seen in the training corpus.

[11] http://clip.ipipan.waw.pl/MweLitRead

References

Cristina Cacciari and Paola Corradini. 2015. Literal analysis and idiom retrieval in ambiguous idioms processing: A reading-time study. *Journal of Cognitive Psychology* 27(7):797–811. https://doi.org/10.1080/20445911.2015.1049178.

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* to appear.

Ismail El Maarouf and Michael Oakes. 2015. Statistical Measures for Characterising MWEs. In *IC1207 COST PARSEME 5th general meeting*. http://typo.uni-konstanz.de/parseme/index.php/2-general/138-admitted-posters-iasi-23-24-september-2015.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1):61–103. https://doi.org/10.1162/coli.08-010-R1-07-048.

Marie-Sophie Pausé. 2017. *Structure lexico-sentaxique des locutions du français et incidence sur leur combinatoire*. Ph.D. thesis, Université de Lorraine, Nancy, France.

Jing Peng and Anna Feldman. 2016. Automatic idiom recognition with word embeddings. In *SIMBig (Revised Selected Papers)*. Springer, volume 656 of *Communications in Computer and Information Science*, pages 17–29.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 2019–2027. http://www.aclweb.org/anthology/D14-1216.

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.

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 EACL'17 Workshop on Multiword Expressions*.

Livnat Herzig Sheinfux, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner. 2017. *Representation and Parsing of Multiword Expressions*, Language Science Press, Berlin, chapter Verbal MWEs: Idiomaticity and flexibility, pages 5–38.

Jakub Waszczuk, Agata Savary, and Yannick Parmentier. 2016. Promoting multiword expressions in A* TAG parsing. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. pages 429–439. http://aclweb.org/anthology/C/C16/C16-1042.pdf.