Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir

Agata Savary, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, Voula Giouli

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 112 APRIL 2019 5–54

Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir

Agata Savary,[a] Silvio Ricardo Cordeiro,[b] Timm Lichte,[c] Carlos Ramisch,[d] Uxoa Iñurrieta,[e] Voula Giouli[f]

[a] University of Tours, France
[b] Paris-Diderot University, France
[c] University of Tübingen, Germany
[d] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[e] University of the Basque Country, Spain
[f] Athena Research Center, Greece

Abstract

Multiword expressions can have both idiomatic and literal occurrences. For instance *pulling strings* can be understood either as making use of one's influence, or literally. Distinguishing these two cases has been addressed in linguistics and psycholinguistics studies, and is also considered one of the major challenges in MWE processing. We suggest that literal occurrences 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 occurrences of verbal VMWEs, and we apply them to existing treebanks in five typologically different languages: Basque, German, Greek, Polish and Portuguese. We also perform a linguistic study of the literal occurrences extracted by the different heuristics. The results suggest that literal occurrences constitute a rare phenomenon. We also identify some properties that may distinguish them from their idiomatic counterparts. This article is a largely extended version of Savary and Cordeiro (2018).

1. Introduction

A multiword expression (MWE) is a combination of words which exhibits lexical, morphosyntactic, semantic, pragmatic and/or statistical idiosyncrasies (Baldwin and Kim, 2010). MWEs encompass diverse linguistic objects such as idioms (*to* ***pull*** *the*

 Corresponding author: agata.savary@univ-tours.fr
Cite as: Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, Voula Giouli. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. The Prague Bulletin of Mathematical Linguistics No. 112, 2019, pp. 5–54. doi: 10.2478/pralin-2019-0001.

strings 'make use of one's influence to gain an advantage'), compounds (*a* ***hot dog***), light-verb constructions (*to* ***pay*** *a* ***visit***), rhetorical figures (***as busy as a bee***), institutionalized phrases (***traffic light***) and multiword named entities (***European Central Bank***). A prominent feature of many MWEs, especially of verbal idioms such as *to* ***pull*** *the* ***strings***, is their non-compositional semantics, that is, 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 to those which rely on semantic interpretation of text (e.g. information retrieval, information extraction or machine translation).

Another outstanding property of many MWEs, as illustrated in Example (1), is that we can encounter their literally understood counterparts, as in (2).

(1) The boss was **pulling** the **strings** from prison. (EN)
'The boss was making use of his influence while in prison.'

(2) You control the marionette by pulling the strings. (EN)

This phenomenon, also called *literal-idiomatic ambiguity* (Savary et al., 2018), has been addressed in linguistic and psycholinguistic literature, and is considered a major challenge in MWE-oriented NLP tasks (Constant et al., 2017), as will be discussed in Section 10. Despite this considerable attention received from the scientific community, the notion of literal occurrence has rarely been formally defined. It is, thus, often unclear whether uses such as the following should be regarded as literal occurrences:

- "Coincidental" co-occurrences of components of a given MWE or of their homographs, as in Examples (3) and (4) respectively,[1]

  (3) As an effect of pulling, the strings broke. (EN)

  (4) He strings paper lanterns on trees without pulling the table. (EN)

- Variants, like (5), (6), (7) and (8), which change the syntactic dependencies between the components, as compared to (1),

  (5) Determine the maximum force you can pull on the string so that the string does not break. (EN)

  (6) My husband says no **strings** were **pulled** for him. (EN)

  (7) She moved Bill by **pulling** wires and **strings**. (EN)

[1] See below for an explanation of the different styles of highlighting and underlining used in this article.

(8) The article addresses the **strings** which the journalist claimed that the senator **pulled**. (EN)

- Co-occurrences exhibiting substantial changes in semantic roles, as in (9),

(9) The strings pulled the bridge. (EN)

- Uses like (10), where idiomatic and literal meanings are wittingly combined.

(10) He was there, **pulling** the **strings**, literally and metaphorically. (EN)

In this article, we put forward a definition of a literal occurrence which is not only semantically but also syntactically motivated. Intuitively, for a given MWE e with components $e_1, \ldots, e_n$, we conceive a *literal occurrence* (LO) of e as a co-occurrence e' of words $e'_1, \ldots, e'_n$ fulfilling the following conditions:

1. $e'_1, \ldots, e'_n$ can be attributed the same lemmas and parts of speech as $e_1, \ldots, e_n$.
2. The syntactic dependencies between $e'_1, \ldots, e'_n$ are the same or equivalent to those between $e_1, \ldots, e_n$ in a canonical form of e.[2]
3. e' is not an idiomatic occurrence of a MWE

When Conditions 1 and 3 are fulfilled but Condition 2 is not, we will speak of a *co-incidental occurrence* (CO) of e. Formal definitions of these conditions and notions will be provided in Section 2. What we eventually want to capture is that only Example (2) above is considered an LO. Examples (3), (5) and (9) are COs since they do not fulfill Condition 2. Examples (1), (6), (7), (8) and (10) do not fulfill Condition 3, since they are *idiomatic occurrences* (IOs). Finally, Example (4) is considered out of scope (not an IO, an LO or a CO), since it involves a lemma (*string*) with a different part of speech than the the MWE e, and therefore does not fulfill Condition 1. Because of Condition 2, the study of literal occurrences of MWEs is best carried out when explicit syntactic annotation is available, that is, in a treebank.

Assuming the above understanding of LOs as opposed to IOs and COs, this article focuses on verbal MWEs (VMWEs), which exhibit particularly frequent discontinuity, as well as syntactic ambiguity and flexibility (Savary et al., 2018). Henceforth, we use wavy and dashed underlining for LOs and COs, respectively. Straight underlining denotes emphasis. Lexicalized components of MWEs are shown in **bold**. Section 2.4 provides more details on the notation of examples used in this article.

We propose to study two main research questions. Firstly, we wish to quantify the LO phenomenon, that is, to estimate the relative frequency of LOs with respect to IOs

[2] As formally defined in Section 2, a canonical form of a VMWE is one of its least marked syntactic forms preserving the idiomatic meaning. A form with a finite verb is less marked than one with an infinitive or a participle, the active voice is less marked than the passive, etc. For instance, a canonical form of (1) is *the boss **pulled strings***. Dependencies are equivalent if the syntactic variation can be neutralized while preserving the overall meaning. For instance, (8) can be reformulated into *The journalist claimed that the senator **pulled** the **strings**, and this article addresses them.*

and COs, as well as the distribution of this frequency across different VMWE types and categories. Secondly, we are interested in cross-lingual aspects of LOs. To this aim, we focus on five languages from different language genera:[3] Basque (Basque genus), German (Germanic genus), Greek (Greek genus), Polish (Slavic genus) and Portuguese (Romance genus). We try to discover possible cross-lingual reasons that may favour the use of LOs, and, conversely, those reasons which are language specific.

The contributions of these efforts are manifold. We provide a normalized and cross-lingual terminology concerning the LO phenomenon. We pave the way towards a better understanding of the nature of ambiguity in VMWEs. We show that ambiguity between an idiomatic and a literal occurrence of a sequence is a challenge in MWE processing which is qualitatively major but quantitatively minor. We put forward recommendations for linguistically informed methods to automatically discover LOs in text. Last but not least, we provide an annotated corpus of positive and negative examples of LOs in five languages. It is distributed under open licenses and should be useful for linguistic studies, for example, on idiom transparency or figurativeness, as well as for data-driven NLP methods, for example, on MWE identification (Savary et al., 2017; Ramisch et al., 2018) or compositionality prediction (Cordeiro et al., 2019).

The article is organized as follows. We provide the necessary definitions, and in particular we formalize the notions of LOs and COs (Section 2). We exploit an existing multilingual corpus in which VMWE annotations are accompanied by morphological and dependency annotations, but literal occurrences are not tagged (Section 3). We propose heuristics to automatically detect possible LOs of known, that is, manually annotated, VMWEs (Section 4). We manually categorize the resulting occurrences using a typology which accounts for true and false positives, as well as for linguistic properties of LOs as opposed to those of IOs (Section 5). We report on the results in the five languages under study (Section 6), discussing characteristics of LOs (Section 7), of COs (Section 8) and of erroneous occurrences (Section 9). Finally, we present related work (Section 10), draw conclusions and discuss future work (Section 11).

This work is a considerably extended version of Savary and Cordeiro (2018). Compared to the previous article, we expanded our scope to five languages instead of one (Polish). We enhanced and formalized the definition of LOs. We enlarged the annotation typology and designed unified annotation guidelines, which were then used by native annotators to tag LOs, COs and annotation errors in their native languages. Finally, we produced results of both the automatic and the manual annotation for the five languages under study. Thanks to these extensions, the conclusions have a broader significance than in our previous work.

[3]The genus for each language is indicated according to the WALS (Dryer and Haspelmath, 2013).

2. Definitions and notations

In this section we formalize the nomenclature related to sequences and dependency graphs, and we summarize basic definitions concerning VMWEs and their components, adopted from previous work. We also formally define the central notions which are required in this work: VMWE tokens, variants and types, as well as idiomatic, literal and coincidental occurrences. Finally, we explain the notational conventions used throughout this article to gloss and translate multilingual examples.

2.1. Sequences, subsequences, graphs, subgraphs and coarse syntactic structures

Each *sequence* of word forms is a function $s : \{1, 2, \ldots, |s|\} \rightarrow W$, where the domain contains all integers between 1 and $|s|$, and W is the set of all possible word forms (including punctuation). A sequence s can be noted as $s := \{s_1, s_2, \ldots, s_{|s|}\}$, where $s_i := (i, w_i)$ is a single *token*. In other words, a sequence can be denoted as a set of pairs: $s = \{(1, w_1), (2, w_2), \ldots, (|s|, w_{|s|})\}$. For example, the sentence in Example (6), whose morphosyntactic annotation is shown in Figure 1(b), can be represented as a sequence $s = \{(1, \text{My}), (2, \text{husband}), (3, \text{says}), \ldots, (9, \text{him}), (10, .)\}$. Sequences can be seen as perfectly tokenized sentences, because they ignore orthographic conventions regarding spaces between word forms (e.g. before commas), compounding (e.g. *snowman* counts as two word forms), contractions (e.g. *don't* counts as two word forms), etc.

A sentence is a particular sequence of word forms for which the corpus used in our study provides lemmas, morphological features, dependency relations and VMWE annotations. For a given token $s_i = (i, w_i)$, let $\text{surface}(s_i)$, $\text{lemma}(s_i)$ and $\text{pos}(s_i)$ be its surface form, lemma and part of speech.[4] Consider Figure 1, which shows simplified morphosyntactic annotations of Examples (1), (6) and (7) from page 6. In Figure 1(a), $\text{surface}(s_6) = \textit{strings}$ and $\text{lemma}(s_6) = \textit{string}$.

A *dependency graph* for a sentence s is a tuple $\langle V_s, E_s \rangle$, where $V_s = \{\langle 1, \text{surface}(s_1), \text{lemma}(s_1), \text{pos}(s_1)\rangle, \ldots, \langle |s|, \text{surface}(s_{|s|}), \text{lemma}(s_{|s|}), \text{pos}(s_{|s|})\rangle\}$ and E_s is the set of labeled edges connecting nodes in V_s. For instance, Figure 1(a) shows a graphical representation of the dependency graph of sentence (1). Each token s_i of s is associated in the dependency graph 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 Figure 1(a), $\text{label}(s_2) = \text{nsubj}$, $\text{parent}(s_2) = s_4$, $\text{label}(s_4) = \text{root}$, and $\text{parent}(s_4) = \text{nil}$.

Given two sequences p and q over the same word forms, p is a *subsequence* of q iff there is an injection $\text{sub}_p^q : \{1, 2, \ldots, |p|\} \rightarrow \{1, 2, \ldots, |q|\}$, such that: (i) word forms are preserved, that is, for $i \in \{1, 2, \ldots |p|\}$, the condition $p(i) = q(\text{sub}_p^q(i))$ holds; and (ii) order is preserved, that is, for $i, j \in \{1, 2, \ldots |p|\}$, if $i < j$, then $\text{sub}_p^q(i) < \text{sub}_p^q(j)$. Thus, every subsequence is a sequence, and the definitions of lemmas, parts of speech and

[4] Morphological features are not used in our formalization of LOs and are further ignored, although they could be useful to improve our treatment of agglutinative languages like Basque in the future.

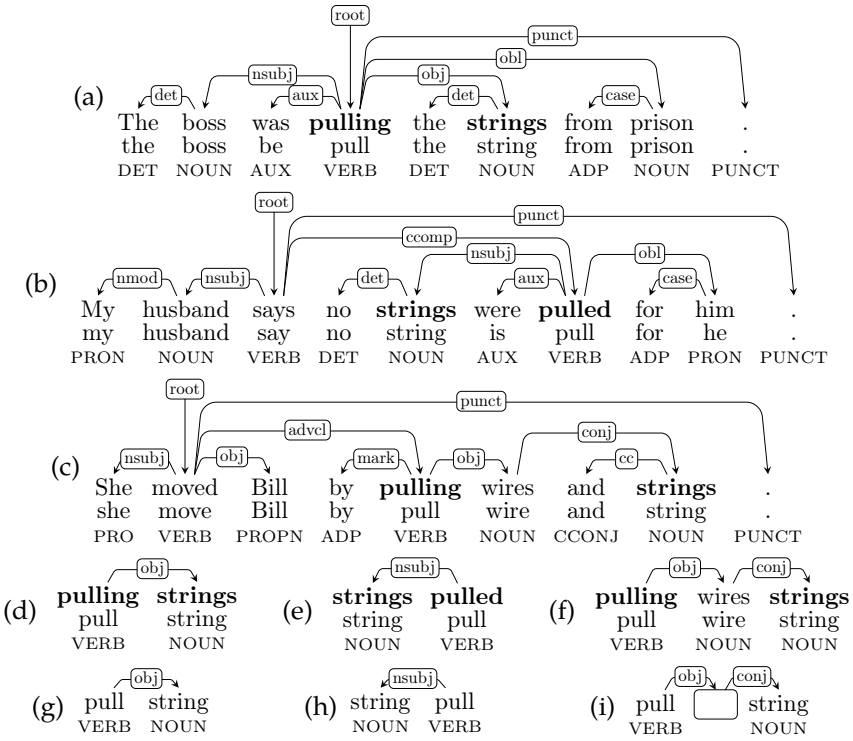

Figure 1. Dependency graphs (a-b-c) for the sentences in Examples (1), (6) and (7), the dependency subgraphs (d-e-f) corresponding to the VMWE tokens in bold, and the coarse syntactic structures (g-h-i) of these tokens. All examples use Universal Dependencies v2.

surface forms of sequence tokens apply straightforwardly to subsequence tokens. For instance, in Figure 1(a), the subsequence corresponding to the tokens in bold can be formalized as $p = \{p_1, p_2\} = \{(1, \text{pulling}), (2, \text{strings})\}$ and $\text{sub}^s_p(1) = 4, \text{sub}^s_p(2) = 6$. We also have $\text{lemma}(p_2) = \text{lemma}((\text{sub}^s_p(2), \text{strings})) = \text{lemma}(s_6) = \textit{string}$, etc.

A subsequence p of a sentence s defines a *dependency subgraph* $\langle V_p, E_p \rangle$ as a minimal weakly connected graph[5] containing at least the nodes corresponding to the tokens in p. In other words, only those edges from $\langle V_s, E_s \rangle$ are kept in $\langle V_p, E_p \rangle$ which appear in the dependency chains connecting the elements of p. If nodes not belonging to p appear in these chains, they are kept in the dependency subgraph for the sake of connectivity. Such nodes are called *intervening nodes*. For instance, Figures 1(d-e-f) show

[5]A directed graph is weakly connected if there is a path between every pair of vertices when the directions of edges are disregarded.

the dependency subgraphs corresponding to two-token subsequences (highlighted in bold) from the sentence graphs from Figures 1(a-b-c). Note that Figure 1(f) corresponds to a subsequence with words *pulling* and *strings* only but its subgraph also contains the intervening node for *wires*.

In a dependency subgraph of a subsequence p we can further abstract away from surface forms and their positions in the sentence, as well as from intervening nodes. In this way, we obtain the *coarse syntactic structure (CSS)* of p. Formally, if p contains k intervening nodes, then $\mathrm{css}(p) = \langle V_{\mathrm{css}(p)}, E_{\mathrm{css}(p)}\rangle$ is a directed graph where $V_{\mathrm{css}(p)} = \{\langle \_, \_, \mathrm{lemma}(p_1), \mathrm{pos}(p_1)\rangle, \ldots, \langle \_, \_, \mathrm{lemma}(p_{|p|}), \mathrm{pos}(p_p)\rangle\}_{\mathrm{ms}} \cup \{\mathtt{dummy}_1, \ldots, \mathtt{dummy}_k\}$, ms denotes a multiset, and $\mathtt{dummy}_i$ are dummy nodes replacing the intervening words.[6] All dependency arcs from E_p are reproduced in $E_{\mathrm{css}(p)}$. Figures 1 (g-h-i) show the CSSes of the subsequences highlighted in bold in Figures 1 (a-b-c).

In a subsequence p, the definition of a parent still relies on the dependencies in the underlying sentence s, but is restricted to the tokens in p. Formally, for a given $1 \leqslant i \leqslant |p|$ and $k = \mathrm{sub}^s_p(i)$, if there exists $1 \leqslant j \leqslant |p|$ and $l = \mathrm{sub}^s_p(j)$ such that $\mathrm{parent}(s_k) = s_l$, then $\mathrm{parent}^s_p(p_i) := p_j$. Otherwise $\mathrm{parent}^s_p(p_i) :=$ nil. For instance, in Figure 1(a), if we take $p = \{p_1, p_2\} = \{(1, \mathrm{pulling}), (2, \mathrm{strings})\}$ and $\mathrm{sub}^s_p(1) = 4, \mathrm{sub}^s_p(2) = 6$, then $\mathrm{parent}^s_p(p_1) = \mathrm{nil}$ and $\mathrm{parent}^s_p(p_2) = p_1$.

Note that, in Figure 1(c), where the subsequence *pulling strings* forms a non connected graph, the parents of both components are nil, that is, taking $\mathrm{sub}^s_p(1) = 5$ and $\mathrm{sub}^s_p(2) = 8$, we have $\mathrm{parent}^s_p(p_1) = \mathrm{parent}^s_p(p_2) = \mathrm{nil}$, although *strings* is dominated by *wires* in the dependency subgraph in Figure 1(f).

2.2. VMWE occurrences, variants and types

Concerning VMWEs, we adapt and extend the PARSEME corpus definitions from (Savary et al., 2018). Namely, if a sentence s is a sequence of syntactic words (i.e., elementary units linked through syntactic relations), then a *VMWE occurrence* (*VMWE token*) e in s is a subsequence of s (in the sense defined in Section 2.1) of length higher than one[7] which fulfills four conditions.

First, all components $e_1, \ldots, e_n$ of e must be *lexicalized*, that is, replacing them by semantically related words usually results in a meaning shift which goes beyond what is expected from the replacement. For instance, replacing *pulling* or *strings* in Example (1) by their synonyms *yanking* or *ropes*, respectively, leads to the loss of the idiomatic meaning: the sentence no longer alludes to using one's influence. Conversely, the determiner *the* can be interchanged with *some*, *many*, etc. with no harm to the idiomatic meaning. Therefore, *pulling* and *string* are lexicalized in (1) but *the* is not.

[6] The first two empty slots denote unspecified positions and surface forms.

[7] The PARSEME guidelines assume the existence of multiword tokens, some of which can be VMWEs, e.g. (DE) *aus-machen* 'out-make'⇒'open'. They consist of at least two words which occur as single tokens due to imperfect tokenization. Our definition of sequences excludes multiword tokens.

Second, the head of each of e's *canonical forms* must be a verb v. A canonical form of a VMWE is one of its least marked syntactic forms preserving the idiomatic meaning. A form with a finite verb is less marked than one with an infinitive or a participle, a non-negated form is less marked than a negated one, the active voice is less marked than the passive, a form with an extraction is more marked than without, etc. For most VMWEs, the canonical forms are equivalent to the so-called *prototypical verbal phrases*, that is, minimal sentences in which the head verb v occurs in a finite non-negated form and all its arguments are in singular and realized with no extraction. For some VMWEs, however, the prototypical verbal phrase does not preserve the idiomatic meaning, and then the canonical forms can be, for example, with nominal arguments in plural. This is the case in Example (11), which shows a canonical form of the VMWE occurrences from Examples (1), (6) and (7)[8], with a direct object in plural (for brevity, subjects are replaced by *he*).

(11) he **pulled** the **strings** (EN)

Other examples of canonical forms which are not prototypical verbal phrases include passivized phrases, as in (EN) ***the die is cast*** 'the point of no retreat has been passed' vs. (EN) *someone cast the die*.

Third, all lexicalized components other than v in a canonical form of e must form phrases which are syntactically directly dependent on v. In other words, $e_1, \ldots, e_n$ and the dependency arcs which connect them in s must form a weakly connected graph. This condition heavily depends on a particular view on syntax and, more specifically, on representing dependency relations. In this article, we follow the conventions established by the Universal Dependencies (UD) initiative (Nivre et al., 2016), which assume, in particular, that syntactic relations hold between content words, and function words depend on the content words which they specify. One of the consequences of this stance is that inherently adpositional verbs, composed of a verb and a selected preposition such as *rely on*, do not form connected graphs (the preposition is a *case* marker of the verb's object). Therefore, they are not considered VMWEs.

Finally, e in s must have an idiomatic meaning, that is, a meaning which cannot be deduced from the meanings of its components in a way deemed regular for the given language.[9] Semantic idiomaticity is hard to estimate directly, but has been approximated by lexical and syntactic tests defined in the PARSEME annotation guidelines (version 1.1).[10] These tests are applied to a canonical form of any VMWE candidate.

[8] As well as from Examples (8) and (10), which are further neglected.

[9] Morphological and/or syntactic idiomaticity of MWEs is also mentioned by some works. However, it implies semantic idiomaticity, because regular rules concern regular structures only. Thus, if an MWE is morphologically or syntactically irregular, its meaning cannot be derived by regular rules.

[10] `http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/`

Recall that a VMWE token e is a subsequence of a sentence s and is associated with a CSS $\text{css}(e) = \langle V_{\text{css}(e)}, E_{\text{css}(e)} \rangle$, as shown in Figures 1 (g-h-i).[11] We define a VMWE *syntactic variant*, or *variant* for short, v as a set of all VMWE occurrences having the same CSS and the same meaning. Formally, let $\sigma_{\text{ID}}(e)$ be the idiomatic meaning contributed by the VMWE token e in sentence s. Then, the VMWE variant associated with e is defined as $v(e) := \{e' \mid \text{css}(e') = \text{css}(e), \sigma_{\text{ID}}(e') = \sigma_{\text{ID}}(e)\}$. Note that VMWE variants as such are not ambiguous: they always come with one meaning. What can be ambiguous, however, is their CSS. For instance, the CSS in Figure 1(g) can have both the idiomatic meaning conveyed in Example (1) and a literal meaning, present in Example (2). Different VMWE occurrences may correspond to the same variant. For instance, the VMWE token from Example (1) and its canonical form in (11) correspond to the variant whose CSS is shown in Figure 1(g).

Finally, collections of VMWE variants form *VMWE types*. Formally, a *VMWE type*, or a *VMWE* for short, is an *equivalence class* of all VMWE variants having the same component lemmas and parts of speech, and the same idiomatic meaning. For each such equivalence class, its *canonical variant* is the variant stemming from its canonical forms, as defined above. The CSS of this canonical representative is called the *canonical structure* of the VMWE. For instance, Figure 1(g) contains the canonical structure of the VMWE type whose occurrences are highlighted in bold in Figures 1(a-c).

2.3. Idiomatic, literal and coincidental occurrences

Given the definitions from the previous section, consider a VMWE type t with n components and $|t|$ variants. Formally, $t = \{\langle \text{css}_1, \sigma_{\text{ID}} \rangle, \langle \text{css}_2, \sigma_{\text{ID}} \rangle, \ldots, \langle \text{css}_{|t|}, \sigma_{\text{ID}} \rangle\}$, and $\text{css}_i = \langle V, E_i \rangle$, where $V = \{\langle \_, \_, \text{lemma}_1, \text{pos}_1 \rangle, \ldots, \langle \_, \_, \text{lemma}_n, \text{pos}_n \rangle\}_{\text{ms}}$. Let s be a sentence of length $|s|$. A *potential occurrence* p of t in s is defined as a subsequence of s whose lemmas and parts of speech are those in (any of the CSSes of) t. Formally, p is a subsequence of length n of s (in the sense of the definitions in Section 2.1) and $\{\langle \_, \_, \text{lemma}(p_1), \text{pos}(p_1) \rangle, \ldots, \langle \_, \_, \text{lemma}(p_n), \text{pos}(p_n) \rangle\}_{\text{ms}} = V$.

Then, we assume the following definitions:

- p is an *idiomatic reading occurrence*, or *idiomatic occurrence* (IO) for short, of t iff
  - The CSS of p is identical to one of the CSSes in t.
  - p occurs with the meaning σ_{ID}, or with any other idiomatic meaning[12].
- p is a *literal reading occurrence*, or *literal occurrence* (LO) for short, of t iff

[11] Since $\text{css}(e)$ only specifies the lemmas of e's components, it might lack morphosyntactic constraints associated with e, e.g., the nominal object must be plural in *pull strings*. This motivates the annotation categories LITERAL-MORPH and LITERAL-SYNT presented in Section 5.

[12] This alternative condition covers cases of VMWE variants with the same CSS but different idiomatic meanings, for instance (EN) *to* ***take in*** 'to make a piece of clothing tighter', (EN) *to* ***take in*** 'to include something', (EN) *to* ***take in*** 'to remember something that you hear', etc. Note that, in this case, even if p is an idiomatic occurrence of t, it does not belong to any of t's variants, because of its different meaning. In other words, an IO of t is not necessarily an occurrence of t. It is rather an IO of t's CSS.

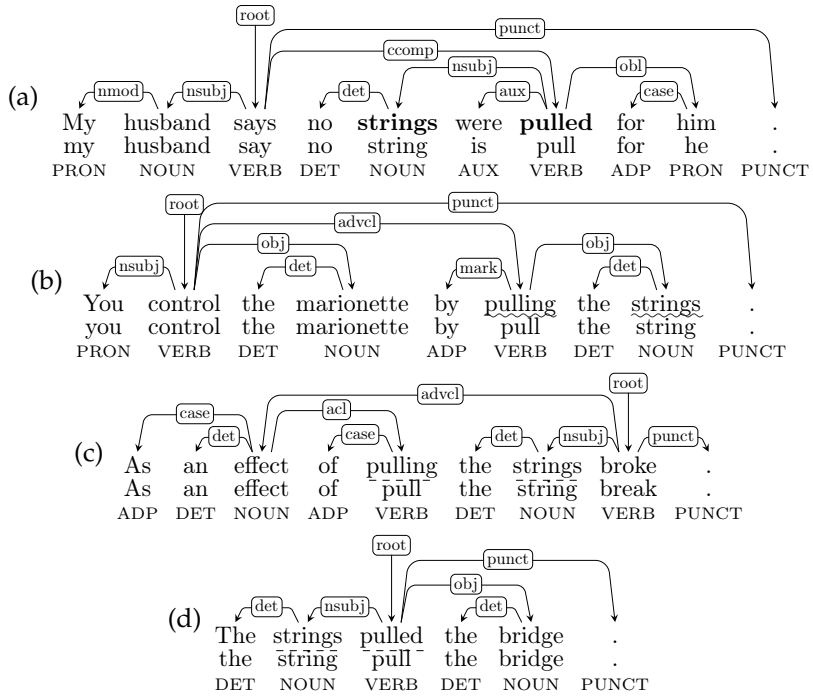

Figure 2. Morphosyntactic annotations (disregarding morphological features) for occurrence contexts of the VMWE (EN) ***pull strings****: (a) idiomatic occurrence, (b) literal occurrence, (c–d) coincidental occurrences.*

  - There is a rephrasing s' of s (possibly identical) such that: (i) s' is synonymous with s, (ii) there is a subsequence p' in s' such that the CSSes of p and p' have identical sets of vertexes ($V_{css(p)} = V_{css(p')}$), (iii) the CSS of p' is equal to the canonical structure of t.
  - p occurs with no idiomatic meaning (i.e not with the meaning σ_{ID} in particular), or it is a proper subsequence of a longer VMWE occurrence[13].
- p is a *coincidental occurrence* (CO) of t iff
  - there is no rephrasing s' of s which fulfills conditions (i–iii) describing an LO above.

For instance, consider the VMWE type t with the three variants whose CSSes are shown in Figure 1(g-h-i), and whose meaning is σ_{ID} = 'to make use of one's influ-

[13]This alternative condition covers cases like (EN) *He* ***pulled the string*** 'In baseball, he threw a pitch that broke sharply', which has one more lexicalized component (*the*) than the VMWE tokens in Figures 1(a-b-c).

ence'. Then, t occurs idiomatically, literally and coincidentally in the sentences from Figure 2(a), (b) and (c–d), respectively. In particular, the CO in Figure 2(d) has the same CSS as the IO in Figure 2(a). Still, the former is not an LO, since it cannot be rephrased in such a way that *strings* becomes the direct object of *pulling*, which is required in the canonical structure of t.

2.4. Notations for multilingual examples

Multilingual aspects of VMWEs addressed in this article are illustrated with examples which follow the notational conventions put forward in Markantonatou et al. (2018). A numbered example like (12) contains a sample VMWE in the original script followed by an ISO 639-1 language code,[14] a transcription (if any), a gloss, as well as a literal and an idiomatic translation. The inline version of the same example is: (EL) κάτι τέτοιο θα **ανοίξει την πόρτα** σ τη διαφθορά (kati tetio tha anixi tin porta s ti diaphthora) 'this will open the door to corruption'⇒'this will enable corruption'. The transliteration and the literal or idiomatic translations may sometimes be omitted for the sake of brevity or focus, as in (EL) κάτι τέτοιο θα **ανοίξει την πόρτα** σ τη διαφθορά 'this will open the door to corruption'.

(12) Κάτι τέτοιο θα **ανοίξει την πόρτα** στη διαφθορά. (EL)
Kati tetio tha anixi tin porta sti diafthora.
something such will open the door to-the corruption
This will open the door to corruption. 'This will enable corruption.'

These conventions also determine that segmentable morphemes are separated by a hyphen, as in the detachable verb-particle construction *ab-gesteckt* 'off-stuck' in Example (13), while one-to-many correspondences between the example and the gloss are marked by dots, as for *vom* 'by.the.DAT' in the same example.

(13) Der Rahmen für diese Verhandlungen soll vom Minister-rat
The framework for these negotiations should by.the.DAT Minister-council
ab-gesteckt werden. (DE)
off-stuck be.
The framework for these negotiations should be stuck off by the Council of Ministers. 'The framework for these negotiations should be set by the Council of Ministers.'

3. Corpus

We use the openly available PARSEME corpus, annotated for VMWEs in 19 languages (Savary et al., 2018; Ramisch et al., 2018).[15] Among its five major VMWE cat-

[14] DE for German, EL for Greek, EU for Basque, PL for Polish and PT for Portuguese

[15] Downloadable from the LINDAT/CLARIN infrastructure at: `http://hdl.handle.net/11372/LRT-2842`

egories, four are relevant to this study, dedicated to Basque, German, Greek, Polish and Portuguese:

- *Inherently reflexive verbs* (IRV) are pervasive in Romance and Slavic languages, present in German, but absent or rare in English or Greek. An IRV is a combination of a verb V and a reflexive clitic RCLI,[16] such that one of the 3 non-compositionality conditions holds: (i) V never occurs without RCLI, as is the case for the VMWE in (14); (ii) RCLI distinctly changes the meaning of V, like in (15); (iii) RCLI changes the subcategorization frame of V, like in (16) as opposed to (17). IRVs are semantically non-compositional in the sense that the RCLI does not correspond to any semantic role of V's dependents.

  (14) O aluno **se** **queixa** do professor. (PT)
  The student RCLI complains of.the teacher.
  'The student complains about the teacher.'

  (15) O jogador **se** **encontra** em campo. (PT)
  The player RCLI finds/meets on field.
  The player finds/meets himself on the field. 'The player is on the field.'

  (16) Eu **me** **esqueci** do nome dele. (PT)
  I RCLI forgot of.the name of.him.
  I forgot myself of his name. 'I forgot his name.'

  (17) Eu esqueci o nome dele. (PT)
  I forgot the name of.him.
  'I forgot his name.'

- *Light-verb constructions* (LVCs) are VERB(-ADP)(-DET)-NOUN[17] combinations in which the verb V is semantically void or bleached, and the noun N is a predicate expressing an event or a state. Two subtypes are defined:
  - *LVC.full* are those LVCs in which the subject of the verb is a semantic (i.e. compulsory) argument of the noun, as in Example (18),
  - *LVC.cause* are those in which the subject of the verb is the cause of the noun (but is not its semantic argument), as in (19).

  The idiomatic nature of LVCs lies in the fact that the verb may be lexically constrained and contributes no (or little) meaning to the whole expression.

[16] Some languages, e.g. German and Polish, use the term *reflexive pronoun* instead of *reflexive clitic*.

[17] Parentheses indicate optional elements. ADP stands for adposition, i.e. either a preposition or a postposition, spelled separately or together with the noun. The order of components may vary depending on the language, and intervening words (gaps) may occur.

(18) Ikasle hori-k ez **du interes**-ik ikasgai-a-n. (EU)
Student this-ERG no has interest-PART subject-the-LOC
This student has no interest in the subject. 'This student is not interested in the subject.'

(19) Kolpe-a-k **min eman** dio. (EU)
punch-the-ERG pain.BARE give AUX
The punch gave him/her pain. 'The punch hurt him/her.'

- *Verbal idioms* (VIDs) are verb phrases of various syntactic structures (except those of IRVs and VPCs), mostly characterized by metaphorical meaning, as in (20).

(20) Dawno już powinien był **wyciągnąć nogi**. (PL)
long.ago already should.3SG was stretch legs
He should have stretched his legs long ago. 'He should have died long ago.'

- *Verb-particle constructions* (VPC), pervasive in Germanic languages but virtually absent in Romance or Slavic ones, are semantically non-compositional combinations of a verb V and a particle PRT. Two subtypes are defined:
  - *VPC.full* in which the V without the PRT cannot refer to the same event as V with the PRT, as in Example (21),
  - *VPC.semi* in which the verb keeps its original meaning but the particle is not spacial, as in (22).

(21) Ein Angebot von Dinamo Zagreb hat Kovac bereits **aus-geschlagen**. (DE)
an offer of Dinamo Zagreb has Kovac already knocked-out
Kovac has already knocked out an offer from Dinamo Zagreb. 'Kovac has already refused an offer from Dinamo Zagreb.'

(22) Ende März **wertete** eine unabhängige Jury die Bilder **aus**. (DE)
end March evaluated an independent jury the paintings off
Late March, an independent jury evaluated the paintings off. 'Late March, an independent jury evaluated the paintings'

For all languages in the PARSEME corpus, the VMWE annotation layer is accompanied by morphological and syntactic layers, as shown in Figure 3. In the morphological layer, a lemma, a part of speech and morphological features are assigned to each token. The syntactic layer includes syntactic dependencies between tokens. For

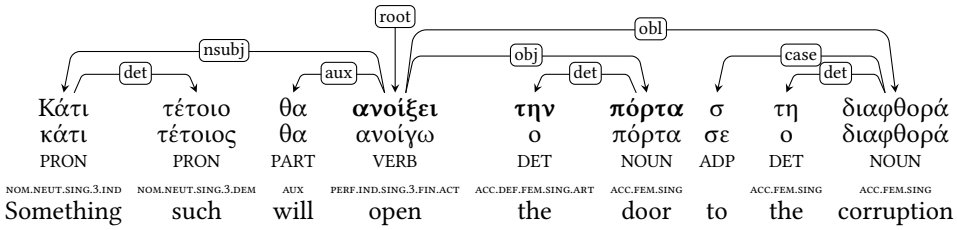

Figure 3. Morphosyntactic annotation for an occurrence context of the VMWE (EL) ***ανοίξει την πόρτα*** *(anixi tin porta) 'open the door'⇒'enable'.*

Language	Sentences	Tokens	VMWEs	Morphological layer		Syntactic layer	
				Tagset	Annotation	Tagset	Annotation
Basque	11,158	157,807	3,823	UD	partly manual	UD	partly manual
German	8,996	173,293	3,823	UD	automatic	UD	automatic
Greek	8,250	224,762	2,405	UD	automatic	UD	automatic
Polish	16,121	274,318	5,152	UD	partly manual	UD	partly manual
Portuguese	27,904	638,002	5,536	UD	partly manual	UD	partly manual

Table 1. Statistics of the PARSEME corpora used to extract LO candidates.

each language, this study combined the training, development and test sets into a single corpus whose sizes, tagsets and annotation methods are shown in Table 1.[18]

While the PARSEME corpus is manually annotated and categorized for IOs of VMWEs, it is not annotated for their LOs. Therefore, we developed several heuristics which allow us to identify them automatically, as discussed in the following section.

4. Automatic pre-identification of literal occurrences

We now consider the task of automatically identifying candidates for LOs in the corpora described in the previous section. In this work, we do not use any external resources. This allows us to compare all languages in a similar manner, but it also means that we can only automatically identify LO candidates for VMWEs which were annotated at least once in the corpus.

Moreover, in order to reliably perform the identification of LOs, we need to ensure that conditions 1, 2 and 3 from page 7 hold. To this aim, we may benefit from the

[18] UD stands for the Universal Dependencies tagset (`http://universaldependencies.org/guidelines.html`). For Basque, the PARSEME corpus uses both the UD tagset and a Basque-specific tagset. For this study, we unified the Basque corpus so that only the UD tagset is used.

morphological, syntactic and VMWE annotation layers present in the corpus. While checking Condition 1, we can rely on the underlying morphological annotation, which contains lemmas and parts of speech. However, as shown in Table 1, most of this annotation was performed automatically, and the risk of errors is relatively high. Therefore, the heuristics defined below rely only on lemmas but not on POS.[19] Condition 2 is closely linked to the syntactic annotations, but checking it fully reliably can be hindered by at least two factors. First, some dependencies can be incorrect, especially if determined automatically. Second, defining conditions under which two sets of dependency relations are equivalent is challenging and highly language-dependent because it requires establishing an exhaustive catalog of all CSSes for a VMWE type. Such a catalogue can be huge, or even potentially infinite, due to long-distance dependencies in recursively embedded relative clauses, as illustrated in Example (8) p. 7. Therefore, the heuristics defined below approximate VMWE types by abstracting away either from the dependency relations or from their directions and/or labels. Finally, Condition 3 can be automatically fulfilled by discarding all LO candidates that coincide with annotated VMWEs. Nonetheless, even if performed manually, VMWE annotations may still contain errors.

In order to cope with these obstacles, we design four *heuristics* which should cover a large part of LOs in complementary ways, while keeping the amount of false positives relatively low (i.e., the heuristics are skewed towards high recall). In the preprocessing step, we extract each occurrence of an annotated VMWE in a sentence s as a subsequence $e = \{e_1, e_2, \ldots, e_{|e|}\}$. For each VMWE e extracted in this way, and for each sentence $s' = \{s'_1, s'_2, \ldots, s'_{|s'|}\}$, we then look for relaxed non-idiomatic occurrences of e in s'. A relaxed non-idiomatic occurrence is a relaxed version of a potential occurrence (cf. Section 2.3), which applies to a VMWE occurrence rather than type, neglects POS and letter case, and is robust to missing lemmas. We first extend the definitions from Section 2 so as to account for missing or erroneous annotations. Namely, for a token s_i in sentence s, we define $\mathrm{lemmasurface}(s_i)$ as $\mathrm{lemma}(s_i)$, if available, and as $\mathrm{surface}(s_i)$ otherwise. Additionally, for any string x, $\mathrm{cf}(x)$ denotes its case-folded version. For instance, in Figure 1(a), $\mathrm{cf}(\mathrm{surface}(s_1)) = \mathrm{the}$. Finally, we say that r is a *relaxed non-idiomatic occurrence* (RNO) of e in s', if r is a subsequence of s' (cf. Section 2.1), $|r| = |e|$, and there is a bijection $\mathrm{rno}^r_e : \{1, 2, \ldots, |e|\} \to \{1, 2, \ldots, |e|\}$, such that: (i) for $i \in \{1, 2, \ldots, |e|\}$ and $j = \mathrm{rno}^r_e(i)$, we have $\mathrm{cf}(\mathrm{lemmasurface}(e_i)) \in \{\mathrm{cf}(\mathrm{lemma}(r_j)), \mathrm{cf}(\mathrm{surface}(r_j))\}$; and (ii) r has not been annotated as a VMWE. For instance, for the VMWE occurrence $e = \{(1, s_5), (2, s_7)\}$ from Figure 2 (a), we obtain the following RNO in sentence s' from Figure 2 (b): $r = \{(1, s'_6), (2, s'_8)\}$, with $\mathrm{rno}^r_e(1) = 2$ and $\mathrm{rno}^r_e(2) = 1$. Note that we do not require the POS tags in r to be the same as in e. In this way, we avoid sensitivity of the heuristics to tagging errors.

[19] Automatically determined lemmas may also be erroneous but we have to rely on them if LOs of previously seen VMWEs are to be found.

The set of such occurrences can be huge, and include a large number of false positives (that is, coincidental occurrences of e's components). Therefore, we restrain the set of *LO candidates* to the RNOs with the following criteria.

- **WindowGap**: Under this criterion, all matched tokens must fit into a sliding window with no more than g external elements (gaps). Formally, let J be the set of all matched indexes in sentence s', that is, $J = \{j \mid \text{sub}_r^{s'}(i) = j\}$. Then r is only considered to match if $\max(J) - \min(J) + 1 \leqslant g + |e|$. For the subsequences e in Figure 2(a) and the RNO r in Figure 2(b), we have $J = \{6, 8\}$ and $|e| = 2$. Thus, the RNO *pulling strings* would be proposed as an LO candidate only if $g \geqslant 1$. The RNO in Figure 2(c) would also be proposed if $g \geqslant 1$. In the case of Figure 1(a), if this VMWE had not been annotated, it could also be proposed as an LO candidate with $g \geqslant 1$, while the occurrence in Figure 1(c) would require $g \geqslant 2$. In this article, WindowGap uses $g = 2$ unless otherwise specified.
- **BagOfDeps**: Under this criterion, an RNO must correspond to a weakly connected unlabeled subgraph with no dummy nodes, that is, the directions and the labels of the dependencies are ignored. For the VMWE in Figure 2(a), the RNO from Figure 2(b) would be proposed, as it consists of a connected graph of the lemmas *pull* and *string*, but the RNO in Figure 2(c) would not be suggested, as the tokens *pulling* and *strings* correspond to a subgraph with a dummy node.
- **UnlabeledDeps**: Under this criterion, an RNO r must correspond to a connected unlabeled graph with no dummy nodes, that is, the dependency labels are ignored but the parent relations are preserved. Formally, this criterion adds a restriction to BagOfDeps: r must be such that, if $\text{parent}_e^s(e_k) = e_l$, $\text{rno}_e^r(k) = i$, and $\text{rno}_e^r(l) = j$, then $\text{parent}_r^{s'}(r_i) = r_j$. For the VMWE in Figure 2(a), the RNO *pulling strings* in Figure 2(b) would be proposed, as it defines a connected subgraph with an arc between the lemmas *pull* and *string*.
- **LabeledDeps**: Under this criterion, an RNO must be a connected labeled graph with no dummy nodes, in which both the parent relations and the dependency labels are preserved. Formally, this criterion adds a restriction to UnlabeledDeps: For every $e_k \in e \setminus \{e_{\text{root}}\}$, if $\text{rno}_e^r(k) = i$ then $\text{label}(e_k) = \text{label}(r_i)$. For the VMWE in Figure 2(a), differently from the heuristic UnlabeledDeps, the RNO *pulling strings* in Figure 2(b) would not be proposed because the label of the arc going from *pulled* to *strings* is not the same in both cases (*obj* vs. *nsubj*).

The heuristics defined by these criteria are language independent and were applied uniformly in the five languages: every RNO covered by at least one of the four heuristics was proposed as an LO candidate.

5. Manual annotation of literal occurrences

The sets of LO candidates extracted automatically were manually validated by native annotators. To this aim, we designed a set of guidelines which formalize the

methodology proposed for Polish in Savary and Cordeiro (2018), with some adaptations. We do not annotate the full corpus, but only the LO candidates retrieved by one of the heuristics, to save time and help annotators focus on potential LOs. As part of the morphological and syntactic layers in our corpora are automatically generated by parsers (Table 1), annotation decisions are taken based on ideal lemmas, POS tags and dependency relations (regardless of the actual dependency graphs in the corpora).

5.1. Annotation labels

We use the labels below for a fine-grained annotation of the phenomena. Each LO candidate is assigned a single label. The label set covers not only the target phenomena (LOs and COs of VMWEs) but also errors due to the original annotation or to the automatic candidate extraction methodology:[20]

- *Errors* can stem from the corpus or from the candidate extraction method.
  1. ERR-FALSE-IDIOMATIC: LO candidates that should not have been retrieved, but have been found due to a spurious VMWE annotation in the original corpus (error in the corpus, false positive):
     - *She […] brought back a branch of dill.* is retrieved as a candidate because *bring back* was wrongly annotated as an IO in ***bringing*** *the predator* ***back*** *to its former home*.
  2. ERR-SKIPPED-IDIOMATIC: LO candidates that should have been initially annotated as IOs in the corpus, but were not (error in the corpus, false negative).
     - *Bring down* was inadvertently forgotten in *Any insult […]* ***brings*** *us all* ***down***, although it is an IO.
  3. NONVERBAL-IDIOMATIC: LO candidates that are MWEs, but not verbal, and are thus out of scope (not an error, but a corpus/study limitation).
     - *Kill-off* functions as a NOUN in *After the major* ***kill-offs****, wolves […].*
  4. MISSING-CONTEXT: more context (e.g. previous/next sentences) would be required to annotate the LO candidate (genuinely ambiguous).
     - Without extra context, *blow up* is ambiguous in *Enron is blowing up.*
  5. WRONG-LEXEMES: The LO candidate should not have been extracted, because the lemmas or POS are not the same as in an IO (errors in the corpus' morphosyntactic annotation, or in the candidate extraction method).
     - The lexemes of *take place* do not occur in *Then take your finger and place it under their belly* because *place* is a VERB rather than a NOUN.
- *Coincidental* and *literal* occurrences are our focus. In the latter case, we also wish to check if an LO might be automatically distinguished from an IO, given additional information provided e.g., in VMWE lexicons.
  6. COINCIDENTAL: the LO candidate contains the correct lexemes (i.e., lemmas and POS), but the dependencies are not the same as in the IO.

[20] Although English is not part of this study, examples were taken from the PARSEME 1.1 English corpus.

  - The lexemes of *to **do the job*** 'to achieve the required result' co-occur incidentally in *[…] why you like the job and do a little bit of […]*, but they do not form and are not rephrasable to a connected dependency tree.
7. LITERAL-MORPH: the LO candidate is indeed an LO that could be automatically distinguished from an IO by checking morphological constraints.
   - The VMWE ***get going*** 'continue' requires a gerund *going*, which does not occur in *At least you get to go to Florida […]*
8. LITERAL-SYNT: the LO candidate is indeed an LO that could be automatically distinguished from an IO by checking syntactic constraints.
   - The VMWE *to **have** something **to do** with something* selects the preposition *with*, which does not occur in *[…] we have better things to do*.[21]
9. LITERAL-OTHER: the LO candidate is indeed an LO that could be automatically distinguished from an IO only by checking more elaborate constraints (e.g. semantic, contextual, extra-linguistic constraints).
   - *[…] we've come out of it quite good friends* is an LO of the VMWE *to **come of it*** 'to result', but it is unclear what kind of syntactic or morphological constraint could be defined to distinguish this LO from an IO.

5.2. Decision trees

Annotators label each automatically identified LO candidate using the decision tree below. Let $e = \{e_1, e_2, \ldots, e_{|e|}\}$ be a VMWE occurrence annotated in a sentence s and cs the canonical structure of e's type. Let $c = \{c_1, c_2, \ldots, c_{|c|}\}$ be e's LO candidate, i.e. an RNO extracted by one of the 4 heuristics from Section 4 in sentence s'.

Phase 1 – initial checks The automatic candidate extraction from Section 4 tries to maximize recall at the expense of precision, retrieving many false positives (e.g., annotation errors or wrong lexemes). Also, sometimes more context is needed to classify c. In this phase, we perform initial checks to discard such cases.

Test 1. [**FALSE**] Should e have been annotated as an IO of an MWE at all?
- NO → annotate c as ERR-FALSE-IDIOMATIC
- YES → go to the next test

Test 2. [**SKIP**] Is c actually an IO of an MWE that annotators forgot/ignored?
- YES, it is a verbal MWE → annotate c as ERR-SKIPPED-IDIOMATIC
- YES, but a non-verbal MWE → annotate c as NONVERBAL-IDIOMATIC
- UNSURE, not enough context → annotate c as MISSING-CONTEXT
- NO → go to the next test

Test 3. [**LEXEMES**] Do c's components have the same lemma and POS as cs's? That is, is c a potential occurrence (as defined in Section 2.3) of e?

[21] Here, the outcome depends on the PARSEME annotation conventions, in which selected prepositions are not considered as lexicalized components of VMWEs.

- NO → annotate c as WRONG-LEXEMES
- YES → go to the next test

Phase 2 – classification Once we have ensured that it is worth looking at the LO candidate c, we will (a) try to determine whether it is a CO or an LO, and (b) if it is the latter, then try to determine what kind of information would be required for an automatic system to distinguish an LO from an IO.

Test 4. **[COINCIDENCE]** Are the syntactic dependencies in c *equivalent* to those in cs? As defined in Section 2.3, dependencies are considered *equivalent* if a rephrasing (possibly an identity) of c is possible, keeping its original sense and producing dependencies identical to those in cs.[22]
- NO → annotate c as COINCIDENTAL
- YES → go to the next test

Test 4. **[MORPH]** Could the knowledge of morphological constraints allow us to automatically classify c as an LO?
- YES → annotate c as LITERAL-MORPH
- NO or UNSURE → go to the next test

Test 4. **[SYNT]** Could the knowledge of syntactic constraints allow us to automatically classify c as an LO?
- YES → annotate c as LITERAL-SYNT
- NO or UNSURE → annotate c as LITERAL-OTHER

5.3. Known limitations

As mentioned above, a precise definition of an LO, as proposed here, can only be done with respect to a particular syntactic framework. This is because we require the syntactic relations within an LO to be equivalent to those occurring in the canonical structure of a VMWE's type. The equivalence of the syntactic relations heavily depends on the annotation conventions of the underlying treebank. Here, we adopt UD, designed mainly to homogenize syntactic annotations across languages.

Suppose that the LVC in *the* ***presentation*** *was* ***made*** is annotated as an IO and that the heuristics propose the LO candidates (a) *his presentation made a good impression* and (b) *we made a surprise at her presentation*. In both LO candidates, the words *make* and *presentation* have a direct syntactic link, so we must base our decision on the relation's label. For Example (a), we cannot compare the labels between the LO candidate and the IO directly (both are `nsubj`), but we must first find the canonical structure of the IO (in which the label is `obj`) to conclude that this candidate is a CO rather than an LO. For candidate (b), the relation is `obl` and cannot be rephrased as `obj`, so this should

[22] Notice that we always compare the dependencies of c (or its rephrasing) with those in a canonical structure cs, never with those in an idiomatic occurrence e.

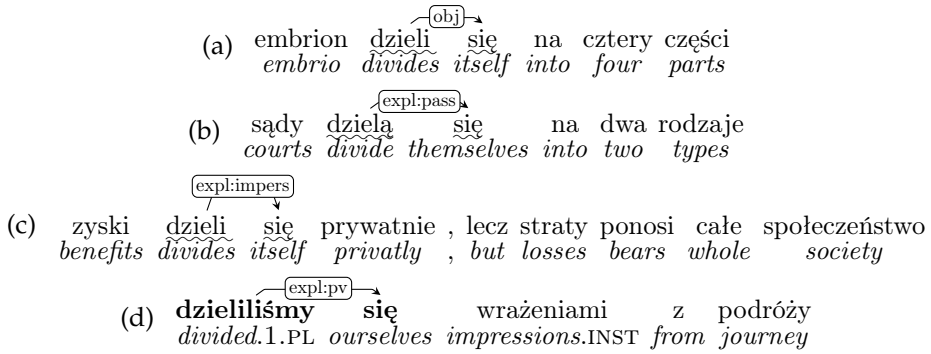

Figure 4. Four UD relations between a verb and a RCLI*. Translations: (a) 'the embryo splits into 4 parts', (b) 'there are 2 types of courts', (c) 'one shares benefits privately but loses are incurred by the whole society', (d) 'we shared our impressions from the journey'*

also be annotated as a CO. Notice that the outcomes could have been different in other syntactic frameworks, e.g., if `obj` and `obl` complements were treated uniformly.

The UD conventions are sometimes incompatible with our intentions. A notable example are verbs with reflexive clitics RCLI. According to UD, each RCLI should be annotated as *obj*, *iobj*, or as an expletive,[23] with one of its subrelations: *expl:pass*, *expl:impers* or *expl:pv* (Patejuk and Przepiórkowski, 2018), as shown in Figure 4. This means that the (semantic) ambiguity between the uses of the RCLI is supposed to be solved in the syntactic layer. Therefore, we ignore the (mostly language specific and often unstable) UD subrelations, so that the uses in Figure 4(b) and (c) are considered LOs of the IO in Figure 4(d). However, the use in Figure 4(a) has to be considered a CO, as we strictly cross our definition of an LO with this UD convention. Still, our intuition is that the (a) vs. (d) opposition in Figure 4 is one of the most challenging types of LOs and should be annotated as such. We postulate a future unification of the UD guidelines at this point, so that all examples in Figures 4(a-b-c-d) are annotated with the same dependency relation in the future. We argue that the distinction between purely reflexive and other uses of the RCLI should be avoided in the syntactic layer and be delegated to the semantic layer instead.

6. Results

In this section, we analyze the distribution of annotations across languages, and the suitability of heuristics (described in Section 4) to find genuine LOs.

[23] `http://universaldependencies.org/u/dep/expl.html#reflexives`

		DE	EL	EU	PL	PT
	Annotated IOs	3,823	2,405	3,823	4,843	5,536
	LO candidates	926	451	2,618	332	1,997
Distribution of labels	ERR-FALSE-IDIOMATIC	21.5% (199)	12.0% (54)	9.4% (246)	0.0% (0)	3.8% (76)
	ERR-SKIPPED-IDIOMATIC	27.0% (250)	47.5% (214)	17.3% (453)	5.4% (18)	10.7% (213)
	NONVERBAL-IDIOMATIC	0.0% (0)	0.0% (0)	0.2% (6)	0.0% (0)	0.5% (9)
	MISSING-CONTEXT	0.3% (3)	0.2% (1)	0.5% (12)	2.1% (7)	0.7% (13)
	WRONG-LEXEMES	40.1% (371)	0.9% (4)	26.7% (700)	1.8% (6)	38.1% (760)
	COINCIDENTAL (COs)	**2.6%** (24)	**27.9%** (126)	**42.4%** (1110)	**61.1%** (203)	**33.5%** (668)
	LITERAL (LOs)	**8.5%** (79)	**11.5%** (52)	**3.5%** (91)	**29.5%** (98)	**12.9%** (258)
	↪ LITERAL-MORPH	0.8% (7)	5.5% (25)	1.9% (51)	1.2% (4)	3.7% (73)
	↪ LITERAL-SYNT	1.5% (14)	2.0% (9)	0.7% (19)	8.1% (27)	2.2% (44)
	↪ LITERAL-OTHER	6.3% (58)	4.0% (18)	0.8% (21)	20.2% (67)	7.1% (141)
	Idiomaticity rate	**98%**	**98%**	**98%**	**98%**	**96%**

Table 2. General statistics of the annotation results. The idiomaticity rate is (#IOs)/(#IOs+#LOs), and #IOs include skipped idiomatic, e.g. $\frac{3823+250}{3823+250+79}$ *for DE.*

6.1. Annotation results

The general statistics of the (openly available) annotation results are shown in Table 2.[24] The VMWE annotations from the original corpus contained between 2.4 (EL) and 5.5 (PT) thousand annotated IOs of VMWEs (row 2).[25] The heuristics from Section 4 were then applied to these VMWEs to find LO candidates. An LO candidate was retained if it was extracted by at least one heuristic. The number of the resulting LO candidates (row 3) varies greatly from language to language, mainly due to language-specific reasons discussed in Sections 7–9. All LO candidates were annotated by expert native speakers (authors of this article) using the guidelines described in Section 5. The next rows (4–13) represent the distribution of annotation labels, documented in section 5.1, among the annotated candidates, across the five languages.

In most languages, a considerable fraction of the candidates turned out to be a result of incorrect annotations in the original corpus. These candidates may be false positives (row 4), or instances of false negatives (row 5).[26] In German, Basque and

[24]The annotated corpus is openly available at `http://hdl.handle.net/11372/LRT-2966`.

[25]In Polish, the reported number of annotated VMWEs is lower in Table 2 (4,843) than in Table 1 (5,152) because the former excludes VMWEs of the IAV (inherently adpositional verb) category, which were annotated only experimentally, and were disregarded in the present study.

[26]A point of satisfaction is that the number of errors of this kind dropped for Polish with respect to our previous work in (Savary and Cordeiro, 2018), performed on edition 1.0 of the PARSEME corpus. This indicates a better quality of the corpus in version 1.1.

	DE					EL				EU			PL				PT			
	IRV	LVC	VID	VPC	All	LVC	VID	VPC	All	LVC	VID	All	IRV	LVC	VID	All	IRV	LVC	VID	All
IdRate	99	100	99	97	**98**	99	95	100	**98**	99	93	**98**	98	99	96	**98**	93	99	88	**96**
EIR	99	100	97	97	**98**	94	92	100	**94**	86	58	**78**	95	94	90	**94**	85	92	73	**86**
ECR	0.6	0.3	1	.1	**.6**	5	3	0	**5**	14	37	**20**	3	5	7	**4**	9	7	18	**10**
ELR	1	0	1	3	**2**	1	5	0	**2**	1	5	**2**	2	1	3	**2**	6	1	10	**4**

Table 3. Extended idiomaticity (EIR), coincidentality (ECR) and literality (ELR). The numbers indicate percentages.

Portuguese, many of the incorrect candidates are also due to wrong lexemes, which results from two factors: (i) the fact that the heuristics rely on lemmas but not on parts of speech (Section 4), and (ii) incorrect lemmas in the underlying morphological layer.

The fraction of actual LOs among the extracted LO candidates (row 10) ranges from 3.5% (EU) to 29.5% (PL). This contrasts with a considerably higher number of COs (row 9) in almost all languages, with the exception of German. This might be partially explained by the fact that 30% of all German candidates stem from annotated multiword-token VPCs, e.g., (DE) *ab-geben* 'submit', which cannot have COs. The distribution of LITERAL-MORPH, LITERAL-SYNT and LITERAL-OTHER (rows 11–13) is addressed in sections 7–9.

The overall quantitative relevance of LOs can be estimated by measuring the *idiomaticity rate* (row 14), that is, the ratio of a VMWE's idiomatic occurrences (initially annotated IOs in the corpus or LO candidates annotated as ERR-SKIPPED-IDIOMATIC) to the sum of its idiomatic and literal occurrences in a corpus (El Maarouf and Oakes, 2015). If the overall idiomaticity rate is relatively low, distinguishing IOs and LOs becomes, indeed, a major challenge, as claimed by Fazly et al. (2009). However, as shown at the bottom of Table 2, the idiomaticity rate is very high (at least 96%) in all languages. In other words, whenever the morphosyntactic conditions for an idiomatic reading are fulfilled, this reading almost always occurs. This is one of the major findings of this work, especially from the point of view of linguistic considerations, given that most VMWEs could potentially be used literally.

From the point of view of NLP, however, more interesting is the proportion of IOs, COs and LOs with respect to the sum of these 3 types of occurrences. This is because a major MWE-oriented task is the automatic identification of MWEs in running text, where COs may play a confounding role. We call these the *extended idiomaticity rate* (EIR), *extended coincidentality rate* (ECR), and *extended literality rate* (ELR), respectively. Rows 4–6 in Table 3 show these three rates across languages and VMWE categories. EIR varies from language to language. In German, Greek and Polish, with total EIR over 94%, our heuristics become a powerful tool for identifying occurrences of previously seen VMWEs. In Basque and Portuguese, the proportion of IOs is much lower, notably due to language-specific CO-prone phenomena, discussed in Section 8. If

	DE		EL		EU		PL		PT	
	tokens	**types**	**tokens**	**types**	**tokens**	**types**	**tokens**	**types**	**tokens**	**types**
IOs	4 073	2 094	2 619	1 270	4 276	856	4 861	1 690	5 749	2 118
COs	24	0.9% (19)	126	5.5% (75)	1 110	18.0% (196)	203	4.7% (85)	668	10.7% (264)
LOs	79	2.4% (51)	52	2.0% (27)	91	3.6% (39)	98	2.6% (48)	258	3.2% (78)

Table 4. Distribution of IOs, LOs and COs across VMWE tokens and types. IO counts are updated to include err-skipped-idiomatic *cases.*

	IOs				COs				LOs			
	IRVs	**LVCs**	**VIDs**	**VPCs**	**IRVs**	**LVCs**	**VIDs**	**VPCs**	**IRVs**	**LVCs**	**VIDs**	**VPCs**
DE	9	8	34	49	8	4	79	8	4	0	27	70
EL	0	72	26	2	0	82	18	0	0	31	69	0
EU	0	79	21	0	0	50	50	0	0	24	76	0
PL	47	43	10	0	33	49	18	0	59	21	19	0
PT	16	64	21	0	14	43	43	0	25	15	60	0

Table 5. Distribution of IOs, LOs and COs, across VMWE categories (values are reported as percentages, adding up to 100 except for rounding).

those phenomena were treated as special cases (e.g., imposing additional morphological constraints) then the heuristics would also be effective for identifying previously seen VMWEs in these languages.

We also looked at the distribution of LOs and COs across VMWE types. Table 4 shows the number of IO, LO and CO tokens and types updated with respect to the initial VMWE annotation statistics, still considering ERR-SKIPPED-IDIOMATIC cases as IOs. Row 4 shows that the proportion of VMWE types which exhibit COs varies greatly among languages: from 0.9% in German to 10.7% in Portuguese and 18.0% in Basque. In Section 8, we further analyze the reasons for these particularities. Row 5 shows that the percentage of VMWE types with LOs is much more uniform, ranging from 2.0% for Greek to 3.6% for Basque. These LOs have a Zipfian distribution, as demonstrated by Figure 5: very few VMWEs have an LO frequency over 5, whereas a large majority of them has only one LO. The top-10 VMWE types with the highest individual LO frequency cover between 39% (in German) and 66% (in Greek) of all LOs. The appendix further shows the 10 VMWE types with the highest ELR and the 10 VMWE types with the highest frequency of LOs in each language. More in-depth language-specific studies might help understand why these precise VMWEs are particularly LO-prone.

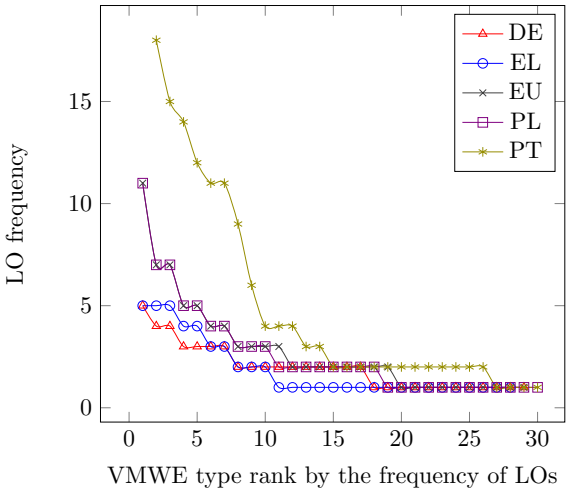

Figure 5. Frequency of LOs of the top-30 VMWE types per language. The VID (PT) ***já era*** *'already was.3SG.IPRF'⇒'it is over' (68 LOs) exceeds the vertical axis and is not shown.*

Table 5 shows the distribution of IOs, COs and LOs across VMWE categories. German has VMWEs of all 4 categories (with almost half of them being VPCs), while the other four languages are missing either IRVs or VPCs (or both). The distribution of COs and LOs across categories varies greatly across languages. The proportion of IOs to COs (excluding the cases of 0 occurrences) varies from 0.43 for German VIDs to 2 for German LVCs, except for German VPCs, with many IOs and LOs but few COs (probably due to the high percentage of mutiword tokens, as mentioned above). We also notice a pattern between LVCs and VIDs in Greek, Basque and Portuguese: LVCs are 2.8 to 3.8 times more frequent than VIDs, but their LOs exhibit roughly the inverse proportions. Interestingly, German seems to have no LOs for LVCs; while in Polish, most LOs stem from IRVs, with other occurrences almost evenly distributed between LVCs and VIDs.

6.2. Results of the heuristics in the task of finding literal occurrences

Once the candidates have been manually annotated, we can verify how well the four heuristics from section 4 solve the task of automatically identifying LOs of previously seen candidates. Table 6 presents precision (P), recall (R) and F-measure (F) in this task for each individual heuristic.

The precision represents the fraction of candidates that were then labeled as LITERAL. As expected, the most restrictive heuristic, LabeledDeps, obtains the highest precision, as its candidates are the ones that resemble the most the morphosyntactic

structure of the annotated VMWEs. In this work we were particularly interested in high recall, since the extracted candidates were further manually validated. The recall is the fraction of all candidates that were retrieved by a given heuristic. This definition of recall does not account for all of the LOs that could possibly have been found, but only for those which have been predicted by at least one heuristic, yielding a recall of 1.00 when the union of all heuristics is considered. We previously showed for Polish that this approximation proves accurate: these heuristics did not miss a single LO in the first 1,000 sentences of the corpus (Savary and Cordeiro, 2018).[27]

The recall for WindowGap is often quite high (91%–98%), suggesting that $g = 2$ is a good number of gaps in the common case, except for German (78%) and Greek (87%). This is consistent with Savary et al. (2018), in which German is an outlier concerning the average gap length within VMWEs (2.96), notably due to the frequency of long-distance dependencies in VPCs, which also occur in LOs, as in (DE) *Mutter Jasmin hielt ihn in letzter Sekunde fest* 'Mother Jasmin held him firmly till the last second'. Similarly, long-distance dependencies (i.e. those exceeding $g = 2$), due notably to the relatively free word order, especially in LVCs, may account for the 13% of LOs not found in Greek, as in (EL) έχει πολλές σπάνιες και αξιόλογες εικόνες (echi poles spanies ke aksiologes ikones) 'has many rare and valuable pictures'.

Through recall, we can attest that the heuristics are complementary, in the sense that no single heuristic is able to predict all of the LOs. For example, for German, WindowGap has R=78%, thus the other 22% of LOs were predicted through BagOfDeps (and possibly the other two more restrictive heuristics as well). Similarly, BagOfDeps has R=90%, implying that the other 10% were predicted only by WindowGap. This means that only 68% (i.e., $100\% - (22\% + 10\%)$) of the actual LOs were predicted by the intersection of both heuristics. Similar numbers are found for other languages, ranging from an intersection of 60% for Portuguese to 80% for Basque.

As expected, 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). These constraints are often valuable in filtering out false literal candidates, which is why the precision of these 3 methods mostly shows an inverse behavior.

7. Characteristics of literal occurrences

This section provides a qualitative analysis of LOs. The goal is to identify both cross-lingual and language-specific reasons for LOs to occur. Additionally, we show examples of morphosyntactic constraints which, if known in advance, e.g., from MWE lexicons (Przepiórkowski et al., 2017), may help automatically distinguish LOs from IOs in the VMWE identification task. Because the morphosyntactic behavior varies

[27] It might be worth repeating the same experiment for German, where long-distance dependencies in LOs are more pervasive.

Language	WindowGap			BagOfDeps			UnlabeledDeps			LabeledDeps			All (union)		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Basque	3	**91**	7	6	89	**11**	5	58	9	**6**	22	10	3	100	7
German	8	78	14	12	**90**	22	13	**90**	22	**14**	77	**23**	9	100	16
Greek	11	87	20	15	**90**	26	**16**	83	**27**	16	52	24	12	100	21
Polish	33	**96**	49	43	81	56	49	73	**59**	**52**	23	32	30	100	46
Portuguese	14	**98**	25	17	62	27	20	59	30	**34**	37	**36**	13	100	23

Table 6. Precision, recall and F-measure of the heuristics (all reported as percentages).

greatly across VMWE categories, this analysis is performed separately for each category.

7.1. IRVs

IRVs exhibit LOs due to homography with compositional VERB + RCLI combinations with true reflexive, reciprocal, impersonal and middle-passive uses. Recall from Section 5.3 and Figure 4 that these uses of RCLIs are supposed to be syntactically distinguished in UD via subrelations. However, due to their language-specific definition and inconsistent usage, subrelations are ignored in our annotation. Thus, examples like (23) are considered middle passive counterparts of the IRVs in (15), page 16.

(23) Nesse rio se encontraram muitos tipos de peixe. (PT)
In.this river RCLI found/met many kinds of fish.
'Many kinds of fish were found in this river.'

This large potential for LOs is displayed mainly in Portuguese and Polish (Table 5). Most of these LOs were annotated as LITERAL-OTHER, i.e., no explicit morphosyntactic hints can help automatically distinguish them from IOs, notably because the RCLI has a weak and infrequent inflection. Still, some LOs were labeled LITERAL-SYNT because they differ from the corresponding IOs by their valency frames. For instance, the IRV in Example (24) requires a genitive object, while the LO in (25) occurs with an accusative object.

(24) Polityk **dopuszczał się** bezprawia. (PL)
Politician allowed RCLI crime.GEN.
The politician allowed himself crime. 'The politician perpetrated crimes'

(25) Dopuszcza się inną działalność niż gastronomiczna. (PL)
Allows RCLI another activity.ACC than gastronomic.
'Activities other than gastronomic are allowed.'

7.2. LVCs

LVCs are mostly semantically compositional, in the sense that the light verb only contributes a bleached meaning (mostly stemming from morphological features, such as tense and aspect) to the whole expression. Therefore, the notion of an LO is less intuitively motivated for them. An LO of an LVC should be understood as a co-occurrence of the LVC's lexemes that does not have all the required LVC properties. This occurs, for instance, when a noun has both a predicative and a non-predicative meaning, i.e., it does or does not express an event or state. In Examples (26) and (27), the noun *zezwolenie* 'permission' means either the fact of being allowed to do something, or a concrete document certifying this fact (i.e. a permit), which yields an LVC and its LOs.

(26) Nie **mają** wymaganego **zezwolenia** na pracę. (PL)
Not have.3rd.PL required permission for work.
'They have no permission to work.'

(27) Kierowcy mieli sfałszowane zezwolenia. (PL)
Drivers had falsified permissions.
'The drivers had falsified permissions.'

The LVC in (26), like most other LVCs, exhibit a totally regular morhosyntactic behavior, therefore their LOs are usually classified as LITERAL-OTHER. Still, a few frequent LVCs do impose morphosyntactic constraints, like the LVC in (28), which prohibits modification of its direct object *miejsce* 'place'. Conversely, in the LO in (29), the same noun receives a nominal modifier, which makes it fall into the LITERAL-SYNT class.

(28) Zdarzenie **miało miejsce** w minioną sobotę. (PL)
Event had place in last Saturday.
'The event took place last Saturday.'

(29) Łódż miała stałe miejsce postoju na przystani. (PL)
Boat had permanent place of.parking on harbor.
'The boat had its permanent parking lot in the harbor.'

7.2.1. Polish-specific phenomena

Polish additionally exhibits a particular syntactic phenomenon which triggers a number of LOs. Namely, given the existential *być* 'to be' in present tense, e.g., in *są powody* 'are reasons.NOM'⇒'there are reasons', its negation is realized by the verb *mieć* 'to have' with the subject shifted to the object position, e.g., *nie ma powodów* 'not has reasons.ACC'⇒'there are no reasons'. Thus, an LVC occurring in present tense under the scope of negation, as in (30), is homonymic with a negated existential construction, as in (31).

(30) (Klient) nie **ma powodów** do satysfakcji. (PL)
Client not has reasons for satisfaction.
'(The client) has no reasons to be satisfied'

(31) Nie ma powodów do satysfakcji. (PL)
Not has reasons for satisfaction.
'There are no reasons to be satisfied'

Since Polish is a pro-drop language, the subject in (30) can be skipped, which makes both occurrences look identical. This clearly implies their labelling as LITERAL-OTHER.

7.2.2. Portuguese-specific phenomena

The Portuguese verb *ter* 'to have' exhibits two interesting language-specific phenomena which trigger LOs of LVCs: resultatives and secondary predication.

The structure of resultative constructions, illustrated by Example (32), may be very similar to some LVCs, as in (33). In both cases, the noun is the direct object of the verb *ter* 'to have' and it governs a participle. Because of the well known ambiguity of participles, in (32) the participle *renovada* 'renewed' depends on the noun via the *acl* relation, while in (33) *equilibrada* 'balanced' it is a plain adjectival modifier (one cannot specify the agent of *balance*).

(32) Ele tem sua força renovada quando descansa. (PT)
He has his strength renewed when rests.
'His strength gets renewed when he rests.'

(33) A criança **tem** uma **alimentação** equilibrada. (PT)
The child has a diet balanced.
'The child has a balanced diet.'

This subtle syntactic constraint might make (32) fall into the LITERAL-SYNT class, but it is unclear whether the presence of an outgoing *acl* relation is sufficient to distinguish an IO from an LO. Therefore, cases of this kind were labeled LITERAL-OTHER.

Secondary predication is illustrated in Example (34). There, the verb *ter* 'to have' has both a direct object (*obj*) and an indirect object (*iobj*) introduced by *como/por* 'as/by', the latter being a predicative of the former.

(34) João tem [seu irmão]$_{obj}$ [como um demônio]$_{iobj}$. (PT)
John has his brother as a demon.
'João considers his brother a demon.'

The indirect object can contain an abstract predicative noun, in which case its combination with *ter* 'have' is annotated as LVC.full, as in (35) and (36).

(35) Ela **tem** [**como objetivo**]$_{\text{iobj}}$ [a difusão de informações]$_{\text{obj}}$. (PT)
she has as goal the dissemination of information.
'Her goal is the dissemination of information.'

(36) Eles **tem** [essa atividade]$_{\text{obj}}$ [**como** uma **opção**]$_{\text{iobj}}$. (PT)
they have this activity as an option.
'This activity is a possible option for them.'

However, the opposite may also happen, that is, a predicative noun may appear in the *obj* position, as in (36). In this case, *tem atividade* 'has activity' is not an LVC.full, as it does not pass the V-REDUC test from the PARSEME guidelines.[28] Since the underlying CSS is identical to the canonical structure of this VMWE, this occurrence is annotation as LIT-OTHER.

7.3. VIDs

The origin of many VIDs lies in the metaphorical interpretation of semantically compositional constructions. Such VIDs are figurative (their literal meaning is easy to imagine) and naturally have a potential of LOs, as exemplified in (37)–(38).

(37) Gaixo dago eta ez **da** joateko **gauza**. (EU)
Sick is and no is going thing
He/She is sick and is no thing to go. 'He/She is sick and is unable to go.'

(38) Horiek beste garai bat-eko gauza-k dira. (EU)
These other time one-GEN thing-PL AUX
These are things from the past. 'These things belong to the past.'

Many of such cases, especially in Basque, Greek and Portuguese, can be distinguished by checking morphological or syntactic constraints (i.e. they are labelled LITERAL-MORPH or LITERAL-SYNT). Unlike in (37), the noun *gauza* 'thing' is in plural in (38). Since the noun inside the VID *gauza izan* 'be able (to)' is never used in the plural form, this feature indicates that the occurrence is literal.

Some LOs, however, fall into the LITERAL-OTHER class, notably when they are strong collocations or domain-specific terms. For instance, the LO in (40) is an institutionalized term, and has the same, both incoming and outgoing, syntactic dependencies as its corresponding IO in (39).

(39) Służenie nam **mają** **we krwi**. (PL)
serving us have.3rd.PL in blood
They have serving us in blood. 'Serving us is their innate ability.'

[28] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=lvc#test-lvc4

(40) Miał we krwi ponad 1,5 promila alkoholu (PL)
had.3rd.SING in blood over 1.5 per-mille alcohol
'His blood alcohol level was 1.5.'

7.3.1. Basque-specific phenomena

Basque, unlike the four other languages, is both postpositional and agglutinative, meaning that adpositions (which are separate words in the other four languages) are suffix-like (Inurrieta et al., 2018). Words decorated with different postpositions lemmatize to bare forms in which the postpositions are omitted. For instance, *kontu-a-n* 'account-ART-LOC' in Example (41) and *kontu-tik* 'account-ABL' in (42) both lemmatize to *kontu* 'account'. Additionally, the dependencies between these components and *hartu* 'take' are the same. Recall from Section 2.3 that the status of a candidate as an IO/LO/CO is based on comparing its CSS with the canonical structure of an IO. CSSes contain lemmas of the lexicalized components, which means that (suffix-like) adpositions in Basque are ignored in this comparison. This is why Example (42) counts as an LO of (41), despite the different adpositions *-n* 'LOC' and *-tik* 'ABL'.

(41) **Kontu-a-n** **hartu** du lagun-a-ren iritzi-a. (EU)
account-ART-LOC take AUX friend-ART-GEN opinion-ART.ABS
Took into account the opinion of his/her friend. 'He/She took his/her friend's opinion into account.'

(42) Diru-a hartu du kontu-tik. (EU)
money-ART.ABS take AUX account-ABL
Took money from the account. 'He/She withdraw money from the account.'

This behavior and modeling of adpositions is in sharp contrast with languages using prepositions on the one hand, and those using adverbial prefixes on the other. Prepositions are standalone words and can constitute independent lexicalized components of VMWEs. For instance, given the VID (EN) ***take*** *money* ***into account***, the occurrence (EN) *take money from my account* cannot be an LO/CO candidate because one lexicalized component (*into*) is missing. Conversely, adverbial prefixes, pervasive in Slavic languages, are inherent parts of the verb's lemma, i.e., they do not vanish in the process of lemmatization.[29] Therefore, given an IRV (PL) ***wy-nosić się*** 'out-carry oneself'⇒'to go away', an occurrence with a different prefix, like *pod-nosić się* 'lift oneself'⇒'stand up', can never be considered an LO/CO candidate.

7.3.2. German-specific phenomena

VIDs give raise to 27% of LOs in German (Table 5). Few of those (unlike in Basque, Greek and Portuguese) fall into the LITERAL-MORPH class (Table 2). The main reason is

[29]They resemble German VPCs as (DE) *auf-nehmen* 'up-take'⇒'to take up', but they are not separable.

that most of them stem from VIDs containing, along with the head verb, a functional word like an expletive pronoun or an adverb. The morphological range for the IO-LO distinction is therefore drastically reduced. Example (43) shows a VMWE with an expletive pronoun, and (44) a corresponding LO.

(43) **Es gilt** Hemmungen zu überwinden und zu lernen mit dem Lampenfieber
it holds inhibitions to overcome and to learn with the stage-fright
umzugehen. (DE)
to.deal

'You have to overcome inhibitions and learn how to deal with stage fright.'

(44) Es gilt der Grundsatz der Gleichbehandlung, erklärt die Sprecherin. (DE)
it holds the principle of equal-treatment says the speaker

'The principle of equal treatment applies, says the speaker.'

Besides the clear semantic contrast (the VMWE in (43) does not imply a legal provision), the two uses of *es gilt* 'it applies'⇒'one should' also differ with respect to their syntax: the VMWE in (43) governs a *zu*-infinitive, whereas the LO instance in (44) governs a noun phrase. Since the governed category is essential for the different readings to emerge, we have annotated the LO as LITERAL-SYNT.

In our German corpus, there is no common lemmatization for personal pronouns. *Es* 'it' is lemmatized as *es*, *er* 'he' as *er*, etc. Therefore, Example (45) cannot be suggested as an LO of (43) by the heuristics, even though this would be perfectly justified.

(45) Er gilt als russischer Mark Zuckerberg: [...] (DE)
he holds as Russian Mark Zuckerberg

'He is considered a Russian Mark Zuckerberg.'

7.3.3. Greek-specific phenomena

Like in German, many LOs of VIDs in Greek contain functional words, mainly pronouns, but in contrast to German, these LOs could be classified as LITERAL-MORPH. This is due to the diversity in how pronouns are modeled in both languages. In German, as just mentioned, each personal pronoun has its own lemma, e.g., *es* 'it' and *sie* 'they' are different lexemes. In Greek, pronouns are seen as exhibiting inflection for person, gender, number and case. Thus, e.g., *το* 'it' and *αυτούς* 'they' are inflected forms of the same lemma *εγώ* 'I'. This yields a large number of LOs. For instance, the VID in (46) comprises a clitic (i.e., a weak form of the personal pronoun) followed by a verb. The clitic *τα* 'them' is fixed with respect to the gender, number and case and does not co-refer with another nominal phrase.

(46) O Γιάννης **τα πήρε** με τα παιδιά. (EL)
O Gianis ta pire me ta pedia.
the John them took with the kids
John took them with the kids. 'John was very angry at the kids.'

The same clitic-verb combinations can occur in an LO, yet the morphosyntactic features of the clitic are not fixed, as in (47), which makes the LO fall into the LITERAL-MORPH category. It may also happen that the clitic in the LO has precisely the same morphology as in the VMWE, in which case the occurrence is labeled LITERAL-OTHER. Further ambiguity stems from clitic doubling (i.e., a construction in which a clitic co-occurs with a full noun phrase in argument position forming a discontinuous constituent with it), as illustrated in (48).

(47) O Γιάννης την πήρε με το αυτοκίνητο. (EL)
O Gianis tin pire me to aftokinito.
the John took her with the car
John took her in his car. 'John gave her a lift'

(48) Η κοπέλα τα πήρε τα έγγραφα (EL)
i kopela ta pire ta egrafa
the girl them took the documents
'The girl took the documents.'

As shown in Table 2, the LITERAL-MORPH class is the most frequent among Greek LOs. The rate of LITERAL-SYNT cases is lower, probably because when syntactic constraints can help solve the IO vs. LO ambiguity, morphosyntactic constraints also apply. In most LITERAL-SYNT cases, IOs either allow only for restricted modification of their elements, or no modification at all, as shown in (49), where the noun *χέρι* 'hand' allows no modifier.

(49) ο δημοσιογράφος τον **κρατάει στο χέρι** (EL)
o dimosiografos ton kratai sto cheri
the journalist him holds in-the hand
The journalist holds him in the hand. 'The journalist has power over him.'

Conversely, LOs allow for modification, and can be identified on the grounds of syntactic features, as shown in (50), where the two modifiers of the noun are underlined.

(50) Στο δεξί του χέρι κρατάει το κουτί (EL)
sto dexi tu cheri kratai to kuti
in-the right his hand holds the box
'He holds the box in his right hand.'

Borderline cases between metaphors and VIDs were also identified, as shown in (51). Their corresponding LOs, like in (52), were marked as LITERAL-OTHER.

(51) Κάλεσε τους πολίτες να **βγουν** **στους** **δρόμους**. (EL)
kalese tus polites na vjun stus dromus
asked,03.SG the citizens to get-out.3PL to-the streets.
He asked citizens to get out to the streets. 'He asked the citizens to protest'

(52) Οι ποντικοί βγήκαν στους δρόμους του Παρισιού εξαιτίας […] (EL)
i pontiki vjikan stus dromus tu Parisiu eksetias […]
the rats went-out to-the streets of-the Paris because-of […]
'The rats appeared in the streets of Paris because of […]'

7.4. VPCs

Among our five languages of study, VPCs are mainly exhibited in German. LOs of a VPC occur whenever the verb is used literally and the particle is spacial. Thus, Example (53) is an LO of the VPC from Example (21) on page 17.

(53) Dem Michael wurden beide Schneidezähne aus-geschlagen (DE)
the.DAT Michael were both incisors out-knocked
'Michael's both incisors were knocked out.'

Despite their potential for LOs illustrated in Example (53), for many VPCs it is difficult to even imagine an LO. Trivially, this is the case where the verb is only used together with the particle, for example the verb *statten* in *aus-statten* 'equip'. But also VPCs such as *auf-geben* 'give up' are concerned, where it is rather the combination of verb and particle which is idiomatic. In the case of *auf-geben*, one might expect the availability of a literal meaning 'give upward', but this meaning is only available with the particle *hinauf*. Since both cases are particularly common in German VPCs (*aus-statten* and *auf-geben* alone occur 5 and 7 times in the corpus), this positively biases the idiomaticity rate.

Nevertheless, the few LOs which do occur in German are still dominated by VPCs 70%), probably due to their dominance also in the IOs (Table 5). Recall also from Table 2 that the majority of LITERAL annotations in the VPC category are classified as LITERAL-OTHER. The justification is similar to the one proposed in Section 7.3.2: since the particle has no inflection at all, VPCs and their LOs can hardly be distinguished in German based on the morphology of their components.

8. Characteristics of coincidental occurrences

Since LOs are contrasted in this work with IOs on the one hand and with COs on the other hand, it is interesting to also understand generic and language-specific

reasons for COs to arise. Recall that the heuristics described in Section 4 include WindowGap, which looks for a co-occurrence of the lexicalized components of a known VMWE within a window containing at most 2 gaps (external words). This leaves room for a large potential of COs and, indeed, those extracted only by the WindowGap method are 1.2 to 2.3 times more numerous than those yielded by BagOfDeps. Such candidates, e.g., (55) which is a CO of (54), in which the words in focus are not linked by direct syntactic dependencies, are of little general interest, except when language-specific studies cause their proliferation (see below).

(54) Es **kommt** auf die Qualität insgesamt **an**. (DE)
It comes on the quality totally on.
'It depends totally on the quality.'

(55) Union rannte an, kam zum Ausgleich … (DE)
Union ran on, came to deuce …
'Union attacked, came to a deuce …'

In the COs extracted with BagOfDeps, the syntactic dependencies are usually different from those occurring in the corresponding IOs. For instance, in (56) the dependency between the verb and the noun is of type *nmod*, while it is *obj* in the corresponding LVC in Example (28). Similarly, in (57), the verb *δίνω* 'give' is linked to the noun *απάντηση* 'answer' with the *subj* relation, while the *obj* relation occurs in the LVC ***δίνω απάντηση*** 'give an answer'.

(56) Teraz nie mam nikogo innego na jego miejsce. (PL)
now not have.1st.SING no-one else on his place
'Now, I have no one else to replace him.'

(57) Η απάντησή του μου δίνει αφορμή για […] (EL)
I apantisi tu mu dini aformi jia […]
the answer his me gives chance for […]
'His answer triggers […].'

Recall, however, from Figure 2 and Section 2.3 that sharing the same dependencies with an IO does not necessarily give an occurrence the status of an LO. It is, instead, the canonical structure of an IO's type which counts for evaluating the equivalence of syntactic relations.

8.1. Basque-specific phenomena

Basque has, by far, the highest number of COs, as attested in Table 2. It also has the highest extended coincidentality rate, especially in VIDs, as seen in Table 3. Many of the COs in Basque include nouns with adpositions, which vanish in the process of

lemmatization, as discussed in Section 7.3.1. For instance, in the VID from Example (58) the noun *aurre* 'front' is bare, and it is the direct object of the verb *egin* 'do'. Occurrences (59) and (60) contain the same noun but with adpositions, which is why their dependency to the verb is of different nature and they are COs rather than LOs.

(58) Arazo-e-i **aurre** **egin** zien. (EU)
problems-ART-DAT front.BARE do AUX
Did front to the problems. 'He/She faced the problems.'

(59) Irakasle-a-ren aurre-a-n egin zuen ariketa. (EU)
teacher-ART-GEN front-ART-LOC do AUX exercise.ART.ABS
'He/She did the exercise in front of the teacher.'

(60) Joan aurre-tik egin zuen ariketa. (EU)
leave front-ART.ABL did AUX exercise.ART.ABS
Did the exercise from front leaving. 'He/She did the exercise before leaving.'

Note that this example is quite analogous to (56) vs. (28), where the preposition does not vanish but is dependent on the noun, and therefore does not intervene in the comparison of the CSSes. It is therefore unclear why precisely the COs of this type are so much more frequent in Basque than in other languages exhibiting prepositions. Possible reasons are lemmatization errors in some corpora, or the fact that verbs in VMWE often govern functional words rather than nouns (e.g. in German VPCs, in German and Greek VIDs, and in Polish IRVs), which mostly excludes the use of prepositions.

8.2. Portuguese-specific phenomena

Portuguese has the second highest number of COs and ICR (Tables 2 and 3), especially in VIDs, like Basque, but also in IRVs. This is notably due to complex attachment mechanisms in reflexive clitics. They are adjacent to verbs in Portuguese, occurring immediately before (e.g., *me lavei* 'RCLI.1SG washed'⇒'I washed myself'), immediately after (e.g., *lavei-me* 'washed-RCLI.1SG') or, in some rare cases, in the middle of the verb, between its root and its suffix (e.g., *lavar-me-ei* 'wash-RCLI.1SG-FUT.1SG'⇒'I will wash myself'). A set of (more or less deterministic) rules allow choosing one of the three alternatives (e.g., a sentence cannot start with a reflexive clitic).

While the attachment of the clitic to its directly adjacent verb is mostly unambiguous, the interaction between reflexive clitics and verbal chains (e.g., auxiliary, modal, and controlled verbs) can be complex.[30] For instance, consider the verb *dever* 'to owe',

[30] In Brazilian Portuguese, a reflexive clitic is always adjacent to its verb (e.g., *vai se lavar* 'will RCLI wash'). European Portuguese has different rules, however, with auxiliary and modal verbs interposed between the clitic and the main verb (e.g., *se vai lavar* 'RCLI will wash'). We focus on Brazilian Portuguese only.

which is also used as a modal verb to express obligatoriness ('must'). In Example (61), the verb is combined with a reflexive clitic forming an IRV ***se deve*** *a* 'RCLI owe to'⇒'results from'. Examples (62) and (63), however, are not IOs of this VMWE, but candidates that must be annotated as a CO and an LO respectively.

(61) A demora **se** **deve** à burocracia. (PT)
the delay RCLI owe to.the bureaucracy
'The delay is due to the bureaucracy.'

(62) Os interessados devem se inscrever. (PT)
the interested.PL must RCLI register
'Those who are interested must register.'

(63) Deve se utilizar roupa ventilada. (PT)
must RCLI use clothes ventilated
'One must use ventilated clothes.'

The choice here depends on whether the clitic is attached to the main verb (CO) or to the modal verb (LO). In (63), the clitic marks an impersonal/middle reading of the whole verbal chain, hence the candidate is annotated as an LO (LITERAL-SYNT). Example (62), however, does not have this interpretation, as the clitic marks the reflexive object of the main verb *inscrever* 'register'. Therefore, it is annotated as a CO.

This distinction is tricky, but negation can be used as a test. One of the rules used to choose the clitic's position with respect to the verb is that negation "attracts" the clitic. The negation of Example (63) becomes *Não se deve utilizar* 'Not RCLI must use', indicating that the clitic is attached to the modal verb *dever* 'must'. In Example (62), negation does not change word order and fails to "attract" the clitic: *não devem se inscrever* 'not must RCLI register', indicating that the clitic attaches to the main verb.

8.3. Polish-specific phenomena

A similar ambiguity in the attachment of reflexive clitics occurs in Polish. It is less frequent but sometimes harder to solve, since *się* 'RCLI' benefits from the relatively free word order in this language and can often be separated from its governing verb. For instance the IRV in (64) triggers a CO in (65), where the reflexive clitic appears closer to the modal *ma* 'should' than to the infinitive *zmienić* 'change' which it depends on. One must therefore be extremely careful while annotating such cases. A possible test is to skip the modal and check if the clitic remains with the main verb as in *wszystko się zmieni* 'everything RCLI change.FUT'⇒'everything will change'.

(64) **Miał się** dobrze. (PL)
had RCLI well.
He had himself well. 'He was fine.'

(65) Teraz ma się wszystko zmienić. (PL)
Now has.to/should RCLI everything change.
'Now everything should change.'

9. Characteristics of erroneous occurrences

In this section, we are interested in the candidates labeled WRONG-LEXEMES, i.e., those which were extracted by the heuristics but do not respect Condition 1 from page 7. In other words, they have either different lemmas or different POS than the lexicalized components of an attested VMWE. Recall from Section 4 that the heuristics check the lemma but not the POS, so as to maximize recall even in presence of errors in morphosyntactic annotation.

As shown in Table 2, WRONG-LEXEMES are very frequent in German, Basque and Portuguese. In each case, this is due to the existence of homographs (understood here as words with the same lemma but different POS). One common case is the ambiguity of some common verbs between a main verb and an auxiliary. For instance, in (66), the auxiliary *tem* 'has' is ambiguous with the light verb appearing in the LVC ***tem força*** 'has strength'.

(66) O time tem mostrado força para reverter resultados. (PT)
the team has shown strength to revert results.
'The team has shown the strength to turn the results around.'

Other dominating classes of homographs are language-specific.

9.1. Basque-specific phenomena

Some Basque nouns (like some Hindi nouns[31]), such as the one in the LVC in Example (67), look identical to adjectives. This happens in (68), which triggers a candidate with a wrong lexeme.

(67) Plan-a-ren **berri** **eman** ziguten. (EU)
plan-ART-GEN news.BARE give AUX
Gave us news of the plan. 'They informed us about the plan.'

(68) Plan berri-a eman ziguten. (EU)
plan new-ART give AUX
'They gave us the new plan.'

Correct lemmatization can also be hindered by adpositions. Namely, several adverbs, such as *berriz* 'again' in Example (69), were formed by adding a postposition

[31] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=lvc

(here: *-z* 'INST') to a noun or an adjective (here: *berri* 'new'). Lemmatization of such adverbs is error-prone, therefore the occurrence in (69) was extracted on the basis of the LVC from Example (67).

(69) Plan-a berriz eman ziguten. (EU)
plan-ART again gave AUX
'They gave us the plan again.'

9.2. German-specific phenomena

Cases labeled WRONG-LEXEMES in German can be attributed to a large extent to particles in VPCs, which often have homographs with a different POS tag such as prepositions (e.g. *an* 'on'), the indefinite article *ein* 'a' and the infinitive marker *zu* (similar to *to* in English). For instance, in Example (70), the preposition *an* 'on' is wrongly confused with the particle appearing in the VPC from Example (54) in page 38.

(70) Beide Teams kamen an die free-throw-line. (DE)
both teams came on the free-throw-line.
'Both teams came up to the penalty line.'

9.3. Portuguese-specific phenomena

In Portuguese, one of the most frequent types of WRONG-LEXEMES stems from the fact that the conjunction *if* and the 3rd-person reflexive pronoun are homographs: *se*. Thus, a conditional sentence such as (71) is extracted on the basis of the IRV *perguntar-se* 'ask-RCLI'⇒'wonder'.

(71) Pergunta se sua mulher poderá vir. (PT)
asks if his wife can-3S-FUT come-INF.
'He asks if his wife will be able to come.'

Another common ambiguity is due to the fact that the subjunctive form *desse* of the verb *dar* 'to give' is a homograph of the contraction *desse* = *d-esse* 'of.this'. While, in this case, the lemmatized forms should have been different, errors in the underlying morphological annotation led to candidates such as the one in (72), extracted on the basis of the VID ***dar jeito*** 'give way'⇒'to find a workaround' .

(72) Foi bom porque vencemos e desse jeito. (PT)
was good because won-1PL and of.this way.
'It was a good thing, because we won, and in such manner.'

Other spurious candidates were proposed due to errors in lemmatization. For example, the verbs *ser* 'to.be' and *ir* 'to.go' have identical surface forms in some tenses

(e.g., *ele foi* 'he was / he went'). In the set of annotated expressions, there are cases in which ***foi bem*** 'went well'⇒'succeeded' and ***se foi*** 'RCLI went'⇒'left' had the word *foi* lemmatized as *ser*. This gave rise to the proposition of the spurious candidates *ser bem* 'be well' and *se ser* 'RCLI be'.

10. Related Work

Literal interpretation of utterances has been an important topic of debate in the philosophy of language. For instance, Recanati (1995) addresses the "standard model" by Grice (1989), which stipulates that "the interpretation of non-literal utterances proceeds in two stages: [a] the hearer computes the proposition literally expressed by the utterance; [b] on the basis of this proposition and general conversational principles, he or she infers what the speaker really means". Recanati (1995) further refutes the Gricean model by showing that, while non-literal interpretations presuppose literal ones, the latter are not necessarily processed before the former. This work does not explicitly address MWEs (i.e. expressions in which non-literal interpretations are conventionalized) but the proposed models of utterance interpretation (the *accessibility-based serial model*, in which only the most accessible interpretation is processed, and the *parallel model*, in which several sufficiently accessible interpretations are processed in parallel) seem applicable to MWEs, too.

Literal occurrences of MWEs, often called their literal readings or literal meanings, have also received a considerable attention from both linguistic and computational communities. From the psycholinguistic viewpoint, Cacciari and Corradini (2015) put special interest on the interplay between literal and idiomatic readings, as well as their distributional and statistical properties, when discovering how idioms are stored and processed in the human mind. Popiel and McRae (1988) collect ratings of frequency and familiarity for literal and figurative interpretations of 30 different idiomatic expressions in English. They find out that figurative interpretations obtain higher rankings in both aspects than literal interpretations. These results are further corroborated by Geeraert et al. (2018), who study the acceptability of lexical variation in VMWEs through rating and eye-tracking experiments. Judges are presented with sentences containing LOs and IOs of a VMWE with more or less variation. They judge the acceptability of the sentences, and at the same time the fixation duration is measured by eye tracking. The results show, in particular, that sentences with LOs are less acceptable than those with IOs, although the fixation duration for the former is shorter than for the latter. Overall, speakers do not feel comfortable with LOs. These results seem consistent with our quantitative analysis showing that LO are rare in our corpora across typologically different languages.

As to linguistic modelling, links between LOs and IOs are used by Sheinfux et al. (2019) to propose a novel typology of verbal idioms. It relies on figuration (the degree to which the idiom can be assigned a literal meaning) and transparency (the relationship between the literal and idiomatic reading). In *transparent figurative* idioms, the

relationship between the literal and the idiomatic reading is easy to recover (*to saw logs* 'snore'). In *opaque figurative* idioms, the literal picture is easy to imagine but its relationship to the idiomatic reading is unclear (*to shoot the breeze* 'chat'). Finally, in *opaque non-figurative* idioms, no comprehensible literal meaning is available, notably due to cranberry words which have no status as individual lexical units (*to take umbrage* 'to feel offended'). Their study also argues that the links between LOs and IOs can indicate which morphosyntactic variations are allowed or prohibited for some idioms.[32] Namely, transparent figurative idioms exhibit more flexibility than opaque figurative ones, because, in the former, the speakers can more easily relate to individual components and transpose their literal properties to the metaphoric level.

LOs and IOs were also addressed in the context of syntactic modelling by formal grammars. The challenge is to account for the difference between LOs and IOs when their syntax is identical. Abeillé and Schabes (1989) show how this problem can be elegantly solved by Lexicalized Tree-Adjoining Grammars containing a finite set of elementary (initial or auxiliary) trees, each of which has at least one lexicalized element. MWEs are represented as special kinds of elementary trees in which heads are made out of several lexical items that need not be contiguous. During parsing, a sentence can be derived by combining elementary trees via substitution (inserting an elementary tree at a non-terminal leaf) or adjunction (inserting an elementary tree at a non-terminal internal node), which yields a derived tree (the syntactic structure of the sentence) and a derivation tree (showing which elementary trees have been combined and how). While parsing ambiguous expressions (e.g., *he* ***kicked the bucket***), the idiomatic and the literal occurrences obtain the same derived trees, but the derivation trees differ. Accordingly, the idiomatic semantics stems from direct attachment of lexical items in the elementary trees, while the literal compositional semantics is a product of substitution (of non-terminal nodes with lexicon items). Lichte and Kallmeyer (2016) go even further and show how LTAGs combined with frame semantics can be used to model the LO-IO ambiguity only in the semantics. Here, derived trees and derivation trees remain identical across readings.

The LO-IO ambiguity is also considered a major challenge in computational processing of MWEs (Constant et al., 2017). This survey notably offers a state of the art in MWE identification, which is modelled by some approaches as a word sense disambiguation (WSD) problem: candidate expressions are extracted beforehand and then they are to be classified as literal or idiomatic. For example, Hashimoto and Kawahara (2008) deal with the ambiguity between literal and idiomatic interpretations of Japanese MWEs in a supervised WSD framework. The features, fed to a binary SVM classifier, account mainly for the morphosyntactic properties of the candidate MWEs, as well as for the lemmas, POS and domains of the words surrounding the them.

Fazly et al. (2009) use unsupervised MWE identification based on statistical measures of lexical and syntactic flexibility of MWEs. They draw upon the assumption

[32]Similar conclusions are drawn by Pausé (2017) from a corpus study of French VMWEs.

that usages in the canonical forms for a potential idiom are more likely to be IOs, and those in other forms are more likely to be LOs. There, the notion of an LO 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 is restricted to the 28 most frequent verb-object pairs and their manually validated IOs and LOs, i.e., COs are excluded from performance measures (unlike in our approach). Their precision and recall in LO identification range from 0.18 to 0.86 and from 0.11 to 0.61, respectively. These results are hard to compare to ours (Table 6), due to the very different understanding of the task and its experimental settings.

Peng et al. (2014) propose another approach to automatically classify LOs and IOs based on bag-of-words topic representations for 1–3 paragraphs containing the candidate phrase. Peng and Feldman (2016) further show how the same problem can be addressed via distributional semantics, where the semantics of a candidate expression, and of its component words, can be represented by their context vectors. In the same vein, Köper and Schulte im Walde (2016) automatically classify German particle verbs into literal or idiomatic by relying, notably, on distributional vectors (e.g. *aus-klingen* 'out-sound'⇒'end') and of their base verbs (e.g. *klingen* 'sound'). Other features, like abstractness of the context words, draw upon the hypothesis that idiomatic particle verbs are more likely to occur with abstract subjects or complements.

Distributional semantics also proves useful in the related task of predicting the semantic compositionality of an expression. Note that subtle links exist between idiomaticity and semantic non-compositionality. On the one hand, the LO-IO opposition is a dychotomy, and as such it did not seem problematic to apply in our corpus annotation experiments. On the other hand, idiomaticity usually stems from non-compositional semantics but this non-compositionality is known to be a matter of scale rather than a binary phenomenon. Estimating the *degree of (non-)compositionality* in MWEs is a convincing showcase for distributional semantics, where it is modelled via the degree of (non-)compositionality of the context vectors of their component words (see e.g., Katz and Giesbrecht 2006).

We are aware of only two previous works, our own, where the LO phenomenon was assessed in quantitative terms. In Waszczuk et al. (2016), we estimate the idiomaticity rate of Polish verbal, nominal, adjectival, and adverbial MWEs at 0.95, which confirms our current results also with respect to non-verbal VMWE categories. More importantly, this work also shows that the high idiomaticity rate can speed up parsing, if appropriately taken into account by a parser's architecture. Further, in Savary and Cordeiro (2018) we pave the way towards this article, by making the first attempt towards defining the notion of LO, and by estimating the idiomaticity rate of Polish VMWEs (at 0.98) on a smaller corpus.

Several datasets containing IO/LO annotations of MWEs were developed in the past. The dataset of Polish IOs and LOs created by us for the Savary and Cordeiro

(2018) publication, is openly available[33] and contains over 3,000 IOs, 72 LOs and 344 COs. The dataset of Tu and Roth (2011) consists of 2,162 sentences from the British National Corpus in which verb-object pairs formed with *do*, *get*, *give*, *have*, *make*, and *take* are marked as positive and negative examples of LVCs. Tu and Roth (2012) built a crowdsourced corpus in which VPCs are manually distinguished from compositional verb-preposition combinations, again for six selected verbs. Cook et al. (2008) present the VNC Tokens dataset, containing almost 3,000 occurrences of 53 Verb+Noun combinations in direct object relation, annotated as literal or idiomatic. In all, only 18% of all combinations were annotated as literal, which is roughly consistent with our study. Hashimoto and Kawahara (2008) offer a Japanese counterpart of these resources, with 146 idioms and over 102,000 example sentences. Sentences were automatically preselected in a corpus if they contained occurrences of the components of a reference MWE, and if the dependencies between those components were "canonical". This probably means that syntactic variability in LOs is underrepresented in this dataset. The authors mention that "some idioms are short of examples", which is corroborates our high idiomaticity rate results in another, typologically different, language. Our resource, described in this article, has a larger scope than these previous datasets: we address 5 languages from 5 language genera, and we cover VMWEs of unrestricted syntactic structures and lexical choices. The corpus is available under open licenses.

Let us finally mention datasets which provide human annotation of IO/LO candidates in a finer framework where semantic compositionality is estimated on a multivalued scale. Bott et al. (2016) offer such a resource for German VPCs, and Ramisch et al. (2016) for English, French and Portuguese Noun-Noun and Adjective-Noun compounds. A review of such datasets can be found in Cordeiro et al. (2019).

11. Conclusions and future work

This article offers an in-depth study of the phenomenon of literal occurrences of verbal multiword expressions, as well as of their interactions with two closely related phenomena: idiomatic occurrences on the one hand, and coincidental occurrences on the other. We firstly propose formal definitions of these three bordering notions, which were missing in the literature so far. The definitions stipulate that LOs, and consequently also COs, should be understood not only in semantic but also in syntactic terms, which motivates their study in treebanks. We then propose a thorough methodology to quantitatively and qualitatively estimate the importance of LOs. It consists in: (i) heuristics for automatic extraction of LOs tuned towards high recall with reasonable precision, (ii) a VMWE-annotated reference corpus in 5 typologically different languages, and (iii) manual annotation based on detailed annotation guidelines designed as decision trees. The results of this annotation are openly available.[34]

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

[34] http://hdl.handle.net/11372/LRT-2966

They constitute a novel resource, given that previous datasets with IO-and-LO annotation were mostly dedicated to a selected language and MWE category.

We claim to have shown that LOs are ***rare birds*** 'exceptional individuals' in our corpus, both among VMWE tokens and types, in all five languages under study. When syntactic conditions necessary for an idiomatic reading are fulfilled, this reading occurs in 96%–98% of the cases, as formalized via the IdRate. These results are only slightly less consistent across VMWE types, and range from 90% in Basque VIDs to 100% in Greek LVCs. This is an important finding from the linguistic viewpoint, because most VMWE could potentially be used literally, but they are rarely so in our corpus. This fact is somehow surprising since local ambiguity is inherent to natural language and humans generally deal with it very efficiently. For instance, numerous single words exhibit both rich polysemy and high frequency, and listeners easily disambiguate them based on context. IO-LO ambiguity can also be easily solved by context in most cases, and yet LOs occur surprisingly infrequently. We put forward the explanation of this fact as an interesting research question.

Given the instances of LOs found in the corpus, we also perform their qualitative analysis. Namely, we explain the conditions under which LOs occur in various VMWE categories, whether cross-lingually or in a language-specific manner. We show examples of morphosyntactic constraints which VMWE impose and which, if known in advance, e.g., from VMWE lexicons, might help automatically distinguish IOs from LOs. These observation might help tune various MWE processing tools (e.g., via fine-grained feature engineering). We additionally point at correlations that exist between the syntactic structure of VMWEs and their capacity to exhibit LOs. For example, many LOs are triggered by those VMWEs in which a head verb governs a functional word only (IRVs, VPCs and VID with expletive pronouns or adverbs). As future work, we wish to further examine these interactions.

We also provide quantitative analyses of LOs from the viewpoint of NLP, where automatic MWE identification is a major challenge for semantically-oriented downstream applications. There, IOs are to be opposed not only to LOs but also to COs (in which the lexemes in focus do occur, but not in the right syntactic configuration). We show that the predominance of IOs in this case is strong for German, Greek and Polish, but weaker for Basque and Portuguese. We show examples of language-specific phenomena which contribute to this fact. We also briefly account for some types of lexical ambiguity which challenge automatic IO/LO/CO extraction methods, and make them highly dependent on the quality of the underlying morphosyntactic annotation.

To conclude, in spite of being rare birds, LOs do *cause a stir* 'incite trouble or excitement'. Firstly, the IO-LO opposition provides a stimulating background for psycholinguistics and language-modeling considerations, which yields interesting insights into human language. Second, the IO-LO ambiguity is considered one of the major challenges in the NLP and has attracted much attention from the community, given that it relates to tasks such as MWE identification. Thirdly, even if we have shown that the LO phenomenon is quantitatively much more modest than expected,

it is still important due to both cross-lingually valid and language-specific phenomena, which are both interesting and not trivial to capture.

Let us finally stress that this is one of the first and few attempts to approach the naturally occurring IO-LO ambiguity on a larger scale in a cross-linguistic setting. We hope that this will inspire subsequent work in a variety of topics, be it in theoretical linguistics, psycholinguistics or computational linguistics.

Acknowledgements

This work was supported by the IC1207 PARSEME (PARSing and Multi-word Expressions) COST action[35], by the French PARSEME-FR project (ANR-14-CERA-0001)[36] and by German CRC 991 grant from Deutsche Forschungsgemeinschaft (DFG)[37].

Bibliography

Abeillé, Anne and Yves Schabes. Parsing Idioms in Lexicalized TAGs. In Somers, Harold L. and Mary McGee Wood, editors, *Proceedings of the 4th Conference of the European Chapter of the ACL, EACL'89, Manchester*, pages 1–9. The Association for Computer Linguistics, 1989. URL http://dblp.uni-trier.de/db/conf/eacl/eacl1989.html#AbeilleS89.

Baldwin, Timothy and Su Nam Kim. Multiword Expressions. In Indurkhya, Nitin and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition, 2010. ISBN 978-1-4200-8592-1.

Bott, Stefan, Nana Khvtisavrishvili, Max Kisselew, and Sabine Schulte im Walde. G$_{h}$ost-PV: A Representative Gold Standard of German Particle Verbs. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon*, Osaka, Japan, 2016.

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

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. Multiword Expression Processing: A Survey. *Computational Linguistics*, to appear, 2017.

Cook, Paul, Afsaneh Fazly, and Suzanne Stevenson. The VNC-Tokens Dataset. In *Proceedings of the Workshop on Multiword Expressions*, 2008.

Cordeiro, Silvio, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 2019. doi: 10.1162/COLI_a_00341. (to appear).

[35] http://www.parseme.eu

[36] http://parsemefr.lif.univ-mrs.fr/

[37] https://frames.phil.uni-duesseldorf.de/

Dryer, Matthew S. and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL https://wals.info/.

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

Fazly, Afsaneh, Paul Cook, and Suzanne Stevenson. Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1):61–103, 2009. doi: 10.1162/coli.08-010-R1-07-048. URL https://doi.org/10.1162/coli.08-010-R1-07-048.

Geeraert, Kristina, R. Harald Baayen, and John Newman. "Spilling the bag" on idiomatic variation. In Markantonatou, Stella, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 1–33. Language Science Press., Berlin, 2018. doi: 10.5281/zenodo.1469551.

Grice, Herbert Paul. *Studies in the Way of Words*. Harvard University Press, Cambridge, Mass., 1989.

Hashimoto, Chikara and Daisuke Kawahara. Construction of an Idiom Corpus and its Application to Idiom Identification based on WSD Incorporating Idiom-Specific Features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 992–1001. Association for Computational Linguistics, 2008. URL http://aclweb.org/anthology/D08-1104.

Inurrieta, Uxoa, Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez-Dios, Antton Gurrutxaga, Ruben Urizar, and Inaki Alegria. Verbal Multiword Expressions in Basque Corpora. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 86–95, 2018.

Katz, Graham and Eugenie Giesbrecht. Automatic Identification of Non-Compositional Multi-Word Expressions using Latent Semantic Analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia, July 2006. URL http://www.aclweb.org/anthology/W/W06/W06-1203.

Köper, Maximilian and Sabine Schulte im Walde. Distinguishing Literal and Non-Literal Usage of German Particle Verbs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California, 2016. URL http://www.aclweb.org/anthology/N16-1039.

Lichte, Timm and Laura Kallmeyer. Same syntax, different semantics: A compositional approach to idiomaticity in multi-word expressions. In Piñón, Christopher, editor, *Empirical Issues in Syntax and Semantics 11*, pages 111–140, 2016. URL http://www.cssp.cnrs.fr/eiss11/.

Markantonatou, Stella, Carlos Ramisch, Agata Savary, and Veronika Vincze. Preface. In Markantonatou, Stella, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press, Berlin, 2018. ISBN 978-3-96110-123-8. doi: 10.5281/zenodo.1469527.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A Multilingual Treebank Collection.

In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation* , LREC 2016, pages 1659–1666. European Language Resources Association (ELRA), 2016. ISBN 978-2-9517408-9-1. 23-28 May, 2016.

Patejuk, Agnieszka and Adam Przepiórkowski. *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, 2018. (263 pages).

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

Peng, Jing and Anna Feldman. Automatic Idiom Recognition with Word Embeddings. In *SIMBig (Revised Selected Papers)*, volume 656 of *Communications in Computer and Information Science*, pages 17–29. Springer, 2016.

Peng, Jing, Anna Feldman, and Ekaterina Vylomova. 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)*, pages 2019–2027, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1216`.

Popiel, Stephen J. and Ken McRae. The figurative and literal senses of idioms, or all idioms are not used equally. *Journal of Psycholinguistic Research*, 17(6):475–487, Nov 1988. ISSN 1573-6555. doi: 10.1007/BF01067912. URL `https://doi.org/10.1007/BF01067912`.

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

Ramisch, Carlos, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, Aline Villavicencio, and Rodrigo Wilkens. How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany, 2016. ACL. doi: 10.18653/v1/P16-2026. CORE2018 rank: A*. `https://aclweb.org/anthology/P16-2026`.

Ramisch, Carlos, 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. 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. URL `http://aclweb.org/anthology/W18-4925`.

Recanati, François. The alleged priority of literal interpretation. *Cognitive Science*, 19:207–232, 1995. URL `https://jeannicod.ccsd.cnrs.fr/ijn_00000181`.

Savary, Agata and Silvio Cordeiro. Literal readings of multiword expressions: as scarce as hen's teeth. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT 16), Jan 2018, Prague, Czech Republic*, pages 64 – 72, Prague, Czech Republic, Jan. 2018.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the EACL'17 Workshop on Multiword Expressions*, 2017.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Sla vomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Lie bes kind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Fe derico Sangati, Ivelina Stoyanova, and Veronika Vincze. PARSEME multilingual corpus of verbal multiword expressions. In Markantonatou, Stella, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth. Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press, Berlin, 2018. ISBN 978-3-96110-123-8. doi: 10.5281/zenodo.1469527.

Sheinfux, Livnat Herzig, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner. Verbal MWEs: Idiomaticity and flexibility. In Parmentier, Yannick and Jakub Waszczuk, editors, *Representation and Parsing of Multiword Expressions*, pages 5–38. Language Science Press, Berlin, 2019.

Tu, Yuancheng and Dan Roth. Learning English Light Verb Constructions: Contextual or Statistical. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, MWE '11, pages 31–39. Association for Computational Linguistics, June 2011. URL `http://www.aclweb.org/anthology/W11-0807`.

Tu, Yuancheng and Dan Roth. Sorting out the Most Confusing English Phrasal Verbs. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation*, SemEval '12, pages 65–69. Association for Computational Linguistics, 2012. URL `http://dl.acm.org/citation.cfm?id=2387636.2387648`.

Waszczuk, Jakub, Agata Savary, and Yannick Parmentier. 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, 2016. URL `http://aclweb.org/anthology/C/C16/C16-1042.pdf`.

Appendix: VMWEs with the highest extended literality rate and frequency of literal occurrences

VMWE	ELR	VMWE	Freq.
ausbauen 'dismount'⇒'enlarge'	0.8	*abgeben* 'give away'⇒'loose'	5
abwehren 'repel'⇒'repel'	0.67	*der heissen* 'its name is'⇒'it means that'	4
ansteigen 'increase'⇒'increase'	0.67	*ausbauen* 'dismount'⇒'enlarge'	4
einleiten 'lead in'⇒'initiate'	0.67	*umstellen* 'surround'⇒'rearrange'	3
sehen an 'watch'⇒'consider'	0.67	*gewachsen sein* 'be grown'⇒'withstand'	3
abgeben 'give away'⇒'loose'	0.625	*gehen weiter* 'go further'⇒'continue'	3
abgegeben (part.) 'give away'⇒'loose'	0.6	*abgegeben (part.)* 'give away'⇒'loose'	3
gewachsen sein 'be grown'⇒'withstand'	0.6	*sehen an* 'watch'⇒'consider'	2
umstellen 'surround'⇒'rearrange'	0.6	*recht haben* 'have the right'⇒'be right'	2
abgestellen (part.) 'park'⇒'switch off'	0.5	*nehmen ab* 'take off'⇒'decrease'	2

Table 7. VMWEs with the highest ELR and LO frequency in German

VMWE	ELR	VMWE	Freq.
τα βάζω 'them put'⇒'to be against'	0.83	*τα ρίχνω* 'them pour'⇒'to blame'	5
εκδίδω ανακοίνωση 'issue announcement' ⇒ 'to announce'	0.83	*εκδίδω ανακοίνωση* 'issue announcement' ⇒ 'to announce'	5
τα ρίχνω 'them throw'⇒'to blame'	0.83	*τα ρίχνω* 'them throw'⇒'to blame'	5
έχω στο χέρι 'have in the hand' ⇒ 'to have control over'	0.75	*τα παίρνω* 'them take'⇒'to become furious'	4
ανοίγω την πόρτα 'open the door'⇒'to allow'	0.67	*το ίδιο κάνει* 'does the same'⇒'never mind'	4
βρίσκομαι σε θέση 'be in position'⇒'to be able to'	0.6	*έχω στο χέρι* 'have in the hand'⇒'to have control over'	3
το ίδιο κάνει 'does the same'⇒'never mind'	0.57	*βρίσκομαι σε θέση* 'be in position'⇒'to be able to'	3
τα παίρνω 'them take'⇒'become furious'	0.5	*ανοίγω την πόρτα* 'open the door'⇒'to allow'	2
δίνω δύναμη 'give power'⇒'to empower'	0.5	*έχω υποχρέωση* 'have obligation'⇒'to be obliged'	2
κρατώ στο χέρι μου 'keep in the hand' ⇒ 'to have control over'	0.5	*παίρνω θέση* 'take seat'⇒'to express my opinion'	2

Table 8. VMWEs with the highest ELR and LO frequency in Greek

VMWE	ELR	VMWE	Freq.
ate ireki 'open door'⇒'to open sth up to sth'	0.75	*berdin izan* 'be equal'⇒'not to mind'	11
atzetik ibili 'walk behind'⇒'to be behind'	0.67	*alde izan* 'be side'⇒'to be in favour'	7
forma hartu 'take form'⇒'to take shape'	0.67	*gauza izan* 'be thing'⇒'to be able'	7
berdin izan 'be equal'⇒'not to mind'	0.55	*balio izan* 'have value'⇒'to be useful'	5
adar jo 'play horn'⇒'to be kidding'	0.5	*jokoan izan* 'be in game'⇒'to be at stake'	5
ate zabaldu 'open door'⇒'to open sth up to sth'	0.5	*laguntza eman* 'give help'⇒'to help'	4
hitz hartu 'take word'⇒'to take sb at sb's word'	0.5	*nabari izan* 'be evident'⇒'to show'	4
kantu egin 'do song'⇒'to sing'	0.5	*ate ireki* 'open door'⇒'to open st up to st'	3
nabari izan 'be evident'⇒'to show'	0.5	*behar izan* 'have need'⇒'to need'	3
pisu ukan 'have weight'⇒'to have an influence'	0.5	*buru ukan* 'have head'⇒'to be intelligent'	3

Table 9. VMWEs with the highest ELR and LO frequency in Basque

VMWE	ELR	VMWE	Freq.
mieć we krwi 'to have in blood'	0.8	*być w stanie* 'be in state'⇒'be able'	11
zerwać się 'break RCLI'⇒'get up abruptly' ⇒ 'have sth as an innate capacity'	0.8	*mieścić się* 'hold RFLI'⇒'fit'	7
dzielić się 'divide RCLI'⇒'share'	0.78	*znaleźć się* 'find RCLI'⇒'be'	5
oprzeć się 'lean RCLI'⇒'resist'	0.71	*oprzeć się* 'lean RCLI'⇒'resist'	5
dopuszczać się 'allow RCLI'⇒'perpetrate'	0.67	*zerwać się* 'break RCLI'⇒'get up abruptly'	4
prosić się 'ask RCLI'⇒'call for'	0.67	*mieć we krwi* 'have in blood' ⇒ 'have sth as an innate capacity'	4
doprowadzić do zatrzymania 'lead to arresting' ⇒ 'cause arresting'	0.5	*przedstawiać się* 'present RCLI'⇒'look'	3
mieć pewność 'have certainly'⇒'be sure'	0.5	*mieć udział* 'have share'⇒'take part'	3
mieć udział 'have share'⇒'take part'	0.5	*mieć się* 'have RCLI'⇒'be'	3
mieć wynik 'have result'	0.5	*znać się* 'know RCLI'⇒'be an expert'	2

Table 10. VMWEs with the highest ELR and LO frequency in Polish

VMWE	ELR	VMWE	Freq.
formar se 'form RCLI'⇒'graduate'	0.8	*já era* 'already was.3SG.IPRF'⇒'it is over'	68
ver se 'see RCLI'⇒'find oneself (in a situation)'	0.79	*dever se* 'owe RCLI'⇒'be due to'	18
posicionar se 'position RCLI'⇒'express an opinion'	0.67	*ter filho* 'have child'⇒'give birth'	15
quero ver 'want.1SG.PRS to.see'⇒'I doubt / I dare'	0.64	*ser a vez* 'be the time'⇒'be someone's turn'	14
ter filho 'have son'⇒'to have a son'	0.62	*ver se* 'see RCLI'⇒'find oneself (in a situation)'	11
fazer cobertura 'make news.coverage'⇒'cover (news)'	0.5	*dizer se* 'say RCLI'⇒'claim to be'	11
fazer placar 'make scoreboard'⇒'score goals'	0.5	*querer.1PS.PRS ver* 'I.want to.see'⇒'I doubt'	9
ganhar números 'gain numbers'⇒'increase in numbers'	0.5	*ir.IMP lá* 'go there'⇒'come on!'	6
morrer em a praia 'die on the beach'⇒'fail at the last stage'	0.5	*querer dizer* 'want to.say'⇒'mean'	4

Table 11. VMWEs with the highest ELR and LO frequency in Portuguese

Address for correspondence:
Agata Savary
agata.savary@univ-tours.fr
University of Tours, IUT of Blois, 3 place Jean-Jaurès, 41000 Blois, France