PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 112 APRIL 2019

EDITORIAL BOARD

Editor-in-Chief

Jan Hajič

Editorial staff

Martin Popel

Editorial Assistant

Jana Hamrlová

Editorial board

Nicoletta Calzolari, Pisa
Walther von Hahn, Hamburg
Jan Hajič, Prague
Eva Hajičová, Prague
Erhard Hinrichs, Tübingen
Philipp Koehn, Edinburgh
Jaroslav Peregrin, Prague
Patrice Pognan, Paris
Alexandr Rosen, Prague
Petr Sgall, Prague
Hans Uszkoreit, Saarbrücken

Published twice a year by Charles University (Prague, Czech Republic)

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: `pbml@ufal.mff.cuni.cz`

ISSN 0032-6585

© 2019 PBML. Distributed under CC BY-NC-ND.

PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 112 APRIL 2019

CONTENTS

Articles

© 2019 PBML. Distributed under CC BY-NC-ND.

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*

© 2019 PBML. Distributed under CC BY-NC-ND. 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 *coincidental 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 $s = \{(1, \mathrm{My}), (2, \mathrm{husband}), (3, \mathrm{says}), \ldots, (9, \mathrm{him}), (10, .)\}$. 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 $\mathrm{surface}(s_i)$, $\mathrm{lemma}(s_i)$ and $\mathrm{pos}(s_i)$ 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), $\mathrm{surface}(s_6) =$ *strings* and $\mathrm{lemma}(s_6) =$ *string*.

A *dependency graph* for a sentence s is a tuple $\langle V_s, E_s \rangle$, where $V_s = \{\langle 1, \mathrm{surface}(s_1), \mathrm{lemma}(s_1), \mathrm{pos}(s_1)\rangle, \ldots, \langle |s|, \mathrm{surface}(s_{|s|}), \mathrm{lemma}(s_{|s|}), \mathrm{pos}(s_{|s|})\rangle\}$ 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 $\mathrm{parent}(s_i)$, through a syntactic label, denoted as $\mathrm{label}(s_i)$. Some tokens may have parent nil (and label root). In Figure 1(a), $\mathrm{label}(s_2) = \mathrm{nsubj}$, $\mathrm{parent}(s_2) = s_4$, $\mathrm{label}(s_4) = \mathrm{root}$, and $\mathrm{parent}(s_4) = \mathrm{nil}$.

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

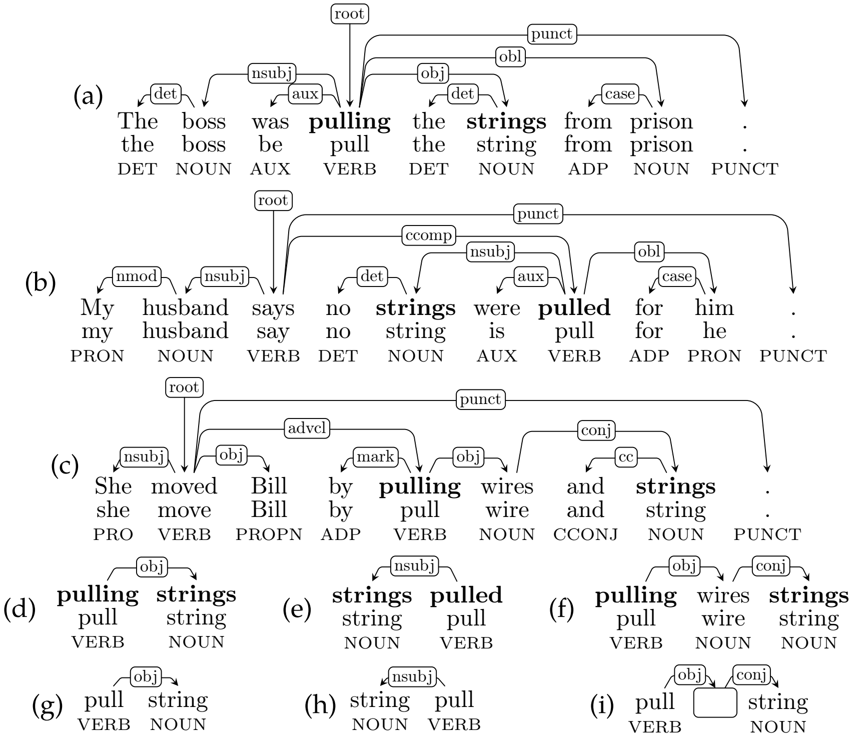

(a)
The boss was pulling the strings from prison .
the boss be pull the string from prison .
DET NOUN AUX VERB DET NOUN ADP NOUN PUNCT
root
det
nsubj
aux
obj
det
obl
case
punct
(b)
My husband says no strings were pulled for him .
my husband say no string is pull for he .
PRON NOUN VERB DET NOUN AUX VERB ADP PRON PUNCT
root
nmod
nsubj
ccomp
det
nsubj
aux
obl
case
punct
(c)
She moved Bill by pulling wires and strings .
she move Bill by pull wire and string .
PRO VERB PROPN ADP VERB NOUN CCONJ NOUN PUNCT
root
nsubj
obj
advcl
mark
obj
conj
cc
punct
(d)
pulling strings
pull string
VERB NOUN
obj
(e)
strings pulled
string pull
NOUN VERB
nsubj
(f)
pulling wires strings
pull wire string
VERB NOUN NOUN
obj
conj
(g)
pull string
VERB NOUN
obj
(h)
string pull
NOUN VERB
nsubj
(i)
pull string
VERB NOUN
obj
conj

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 $V_{\mathrm{css}(p)} = \{\langle \_, \_, \mathrm{lemma}(p_1), \mathrm{pos}(p_1) \rangle, \ldots, \langle \_, \_, \mathrm{lemma}(p_{|p|})), \mathrm{pos}(p_p))\rangle\}_{\mathrm{ms}} \cup \{dummy_1, \ldots, dummy_k\}$, ms denotes a multiset, and $dummy_i$ 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 $k = \mathrm{sub}_p^s(i)$, if there exists $1 \leqslant j \leqslant |p|$ and $l = \mathrm{sub}_p^s(j)$ such that $\mathrm{parent}(s_k) = s_l$, then $\mathrm{parent}_p^s(p_i) := p_j$. Otherwise $\mathrm{parent}_p^s(p_i) := \mathrm{nil}$. For instance, in Figure 1(a), if we take $p = \{p_1, p_2\} = \{(1, \mathrm{pulling}), (2, \mathrm{strings})\}$ and $\mathrm{sub}_p^s(1) = 4, \mathrm{sub}_p^s(2) = 6$, then $\mathrm{parent}_p^s(p_1) = \mathrm{nil}$ and $\mathrm{parent}_p^s(p_2) = p_1$.

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 $\mathrm{sub}_p^s(1) = 5$ and $\mathrm{sub}_p^s(2) = 8$, we have $\mathrm{parent}_p^s(p_1) = \mathrm{parent}_p^s(p_2) = \mathrm{nil}$, 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 $\mathrm{css}(e) = \langle V_{\mathrm{css}(e)}, E_{\mathrm{css}(e)} \rangle$, 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 $\sigma_{\mathrm{ID}}(e)$ be the idiomatic meaning contributed by the VMWE token e in sentence s. Then, the VMWE variant associated with e is defined as $v(e) := \{ e' \mid \mathrm{css}(e') = \mathrm{css}(e), \sigma_{\mathrm{ID}}(e') = \sigma_{\mathrm{ID}}(e)\}$. 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, $t = \{\langle \mathrm{css}_1, \sigma_{\mathrm{ID}}\rangle, \langle \mathrm{css}_2, \sigma_{\mathrm{ID}}\rangle, \ldots, \langle \mathrm{css}_{|t|}, \sigma_{\mathrm{ID}}\rangle\}$, and $\mathrm{css}_i = \langle V, E_i\rangle$, where $V = \{\langle \_, \_, \mathrm{lemma}_1, \mathrm{pos}_1\rangle, \ldots, \langle \_, \_, \mathrm{lemma}_n, \mathrm{pos}_n\rangle\}_{\mathrm{ms}}$. 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 $\{\langle \_, \_, \mathrm{lemma}(p_1), \mathrm{pos}(p_1)\rangle, \ldots, \langle \_, \_, \mathrm{lemma}(p_n), \mathrm{pos}(p_n)\rangle\}_{\mathrm{ms}} = V$.

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 $\mathrm{css}(e)$ 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.

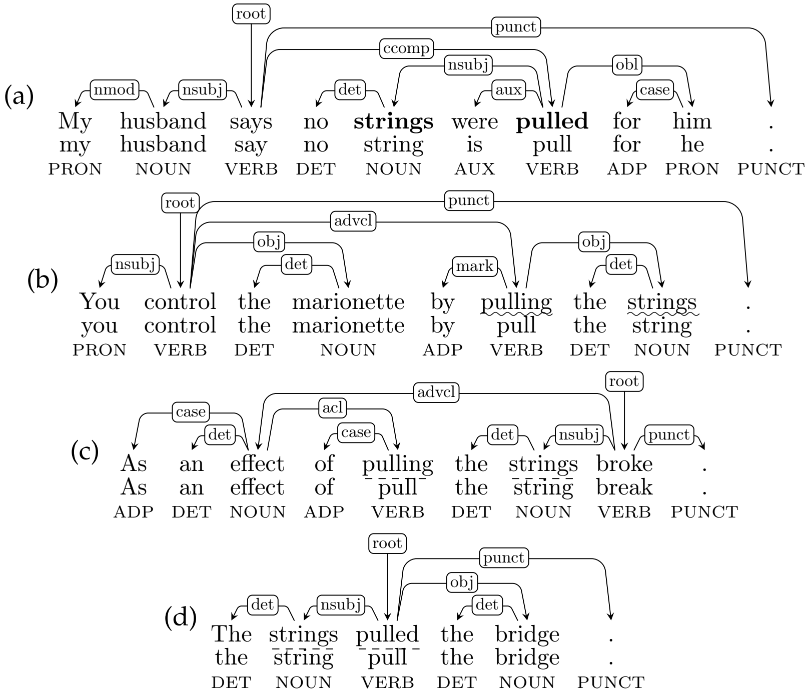

(a)
root
punct
ccomp
nsubj
obl
nmod
nsubj
det
aux
case
My husband says no strings were pulled for him .
my husband say no string is pull for he .
PRON NOUN VERB DET NOUN AUX VERB ADP PRON PUNCT
(b)
root
punct
advcl
obj
obj
nsubj
det
mark
det
You control the marionette by pulling the strings .
you control the marionette by pull the string .
PRON VERB DET NOUN ADP VERB DET NOUN PUNCT
(c)
advcl
root
case
acl
det
case
det
nsubj
punct
As an effect of pulling the strings broke .
As an effect of pull the string break .
ADP DET NOUN ADP VERB DET NOUN VERB PUNCT
(d)
root
punct
obj
det
nsubj
det
The strings pulled the bridge .
the string pull the bridge .
DET NOUN VERB DET NOUN PUNCT

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**.
an offer of Dinamo Zagreb has Kovac already knocked-out
(DE)

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

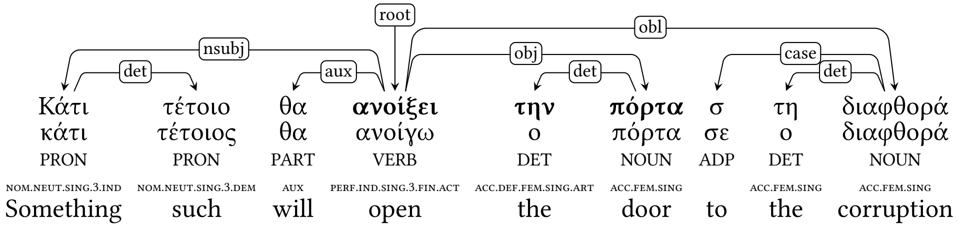


Figure 3. Morphosyntactic annotation for an occurrence context of the VMWE (EL) ***ανοίξει την πόρτα*** *(anixi tin porta) 'open the door'⇒'enable'.*

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 lemmasurface(s_i) as lemma(s_i), if available, and as surface(s_i) otherwise. Additionally, for any string x, cf(x) denotes its casefolded version. For instance, in Figure 1(a), cf(surface(s_1)) $=$ 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 $\text{rno}^r_e : \{1, 2, \ldots, |e|\} \to \{1, 2, \ldots, |e|\}$, such that: (i) for $i \in \{1, 2, \ldots, |e|\}$ and $j = \text{rno}^r_e(i)$, we have $cf(\text{lemmasurface}(e_i)) \in \{cf(\text{lemma}(r_j)),\ cf(\text{surface}(r_j))\}$; 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 $\text{rno}^r_e(1) = 2$ and $\text{rno}^r_e(2) = 1$. 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.

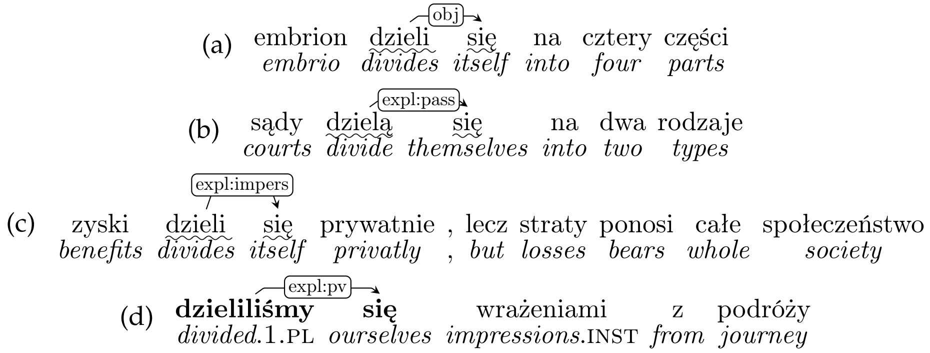

(a) embrion dzieli się na cztery części
embrio divides itself into four parts
obj
(b) sądy dzielą się na dwa rodzaje
courts divide themselves into two types
expl:pass
(c) zyski dzieli się prywatnie , lecz straty ponosi całe społeczeństwo
benefits divides itself privatly , but losses bears whole society
expl:impers
(d) dzieliliśmy się wrażeniami z podróży
divided.1.PL ourselves impressions.INST from journey
expl:pv

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.

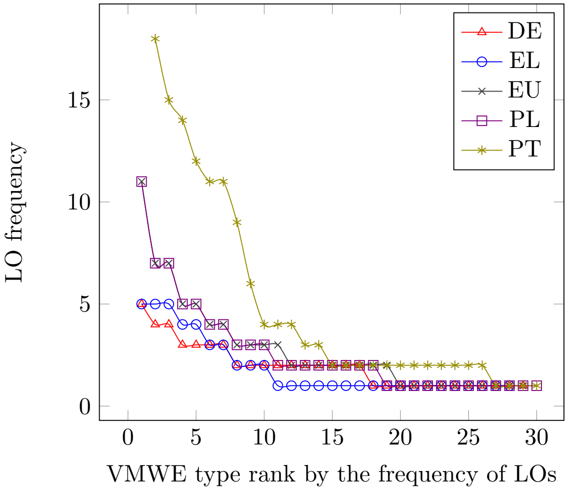


Figure 5. Frequency of LOs of the top-30 VMWE types per language. The VID (PT) ***já era*** *'already was.3SG.IPRF'⇒'it is over' (68 LOs) exceeds the vertical axis and is not shown.*

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

(36) Eles **tem** [essa atividade]$_{obj}$ [**como** uma **opção**]$_{iobj}$. (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) Ο 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) Ο 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 *απάντησή* '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

The Prague Bulletin of Mathematical Linguistics
NUMBER 112 APRIL 2019 55–82

Graph Theory Teaches Us Something About Grammaticality

Koji Arikawa

Department of English and Intercultural Studies, St. Andrew's (Momoyama Gakuin) University, Osaka, Japan

Abstract

Graph theory, which quantitatively measures the precise structure and complexity of any network, uncovers an optimal force balance in sentential graphs generated by the computational procedures of human natural language (C_{HL}). It provides an alternative way to evaluate grammaticality by calculating 'feature potential' of nodes and 'feature current' along edges. An optimal force balance becomes visible by expressing 'feature current' through different point sizes of lines. Graph theory provides insights into syntax and contradicts Chomsky's current proposal to discard tree notations. We propose an error minimization hypothesis for C_{HL}: a good sentential network possesses an error-free self-organized force balance. C_{HL} minimizes errors by (a) converting bottom-up flow (structure building) to top-down flow (parsing), (b) removing head projection edges, (c) preserving edges related to feature checking, (d) deleting DP-movement trajectories headed by an intermediate copy, (e) ensuring that covert wh-movement trajectories have infinitesimally small currents and conserving flow directions, and (f) robustly remedying a gap in wh-loop by using infinitesimally inexpensive wh-internally-merged (wh-IM) edge with the original flow direction.

The C_{HL} compels the sensorimotor (SM) interface to ground nodes so that Kirchhoff's current law (a fundamental balance law) is satisfied. Internal merges are built-in grounding operations at the C_{HL}–SM interface that generate loops and optimal force balance in sentential networks.

1. Introduction – Should we abandon tree notations?

For more than half a century, generative grammar, a plausible candidate for the theoretical base of biolinguistics (Chomsky, 2015), has been using tree-notation as a simple geometrical assistance to express language structures. However, Chomsky (2014) recently stated that tree notations should be abandoned because they are mis-

© 2019 PBML. Distributed under CC BY-NC-ND.
Corresponding author: karikawa@andrew.ac.jp
Cite as: Koji Arikawa. Graph Theory Teaches Us Something About Grammaticality. The Prague Bulletin of Mathematical Linguistics No. 112, 2019, pp. 55–82. doi: 10.2478/pralin-2019-0002.

leading and a branching node in a tree incorrectly indicates that the node is created as a new category (ibid: at approximately 31:58).

(1) "POP [Chomsky (2013)] argues further that projection introduces no new category. That's contrary to phrase-structure grammar and all of its variants and descendants. It also follows from that that the tree notations that are commonly used are highly misleading, and probably they should be abandoned, because the reason is that there is no label for the root that branches. That's just not there. You can't avoid this in the tree notation. But it's not there. In the system, if there is no new category that is introduced by projection, it shouldn't be."

For example, when a verb V and a determiner phrase DP merge, a new set {V, DP} is created. "In its simplest terms, the Merge operation is just set formation" (Berwick and Chomsky, 2016, p. 10).

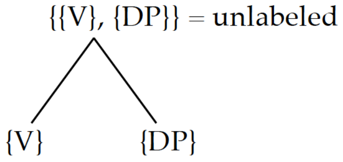


Figure 1. Before the labeling algorithm operation: Non-directed edges

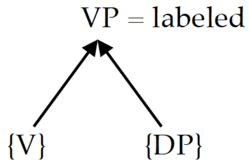


Figure 2. After labeling algorithm operation: Directed edges

Although a new "set" is created, a new "category" is not yet created, i.e., at this point, {{V}, {DP}} is unlabeled (Figure 1). The labeling algorithm (LA) given by Chomsky (2013) later identifies the nature of the set {{V}, {DP}} as the verb phrase (VP) category. Chomsky argued that a tree fails to distinguish between the pre-LA and post-LA structures. However, Chomsky's conclusions were hasty because the unlabeled merge structure before LA becomes a directed tree after LA (Figure 2).

V exists as a set that comprises subsets having phonetic features {Fphon}, semantic features {Fsem}, and formal features {Fform}, i.e., {V} = {{Fphon}, {Fsem}, {Fform}}. Similarly, a DP exists as the set {DP} = {{Fphon}, {Fsem}, {Fform}}. We refer to such a feature set as the "potential" or the "voltage" of the nodes {V} and {DP}, respectively. The merging of V and DP creates an unordered set {{V}, {DP}}, which is an unlabeled exocentric binary branching amalgam, in which the nodes {V} and {DP} are connected to the node {{V}, {DP}}. At this point, the edges are not directed, i.e., there is no feature interaction. LA identifies {{V}, {DP}} as a VP. Here, a less unified amalgam becomes a more unified compound, i.e., neither a DP nor a V. LA reduces (i.e., eliminates) a head feature [X^0] and a categorical feature [D] from {Fform} of {V} and {DP}, respectively, in the amalgam, which creates a more unified compound VP.

The feature reduction is guaranteed by the No Tampering Condition (NTC), which was deduced from the third factor principle of minimal computation (MC) (Chomsky, 2013, p. 40). Let us assume that X and Y merge, and this merger forms a new object Z. NTC specifies that neither X nor Y is modified by the Merge operation. MC requires that X and Y appear in unordered in Z (ibid). At this point, Z is an unlabeled exocentric less-unified amalgam $Z^{unlabeled} = \big\{\{X\}, \{Y\}\big\}$. When LA labels Z, a feature reduction occurs in {X} and {Y} of Z, which yields a labeled endocentric more-unified compound $Z^{labeled} = \Big\{\big\{\{X\} - [f_1]\big\}, \big\{\{Y\} - [f_2]\big\}\Big\}$, where $\big\{\{X\} - [f_1]\big\}$ and $\big\{\{Y\} - [f_2]\big\}$ indicate that formal features $[f_1]$ and $[f_2]$ are reduced from {X} and {Y}, respectively.

Such a feature reduction corresponds to a graph-theoretical "potential drop" or "voltage drop," and the potential drop drives the current flow. After Z is labeled, the linguistic features of X and Y interact with Z. We refer to such feature interactions as the "current" or the "flow." An upward feature interaction is a structure building, and a downward interaction is parsing.

In nature, currents tend to flow in the direction of energy drop, i.e., from a higher to a lower potential energy point. Thus, things fall from the points having high gravitational potentials to the points having low gravitational potentials. Similarly, steam rises from places having high energy densities (i.e., hot places) to places having low energy densities (i.e., cool places). Electric current flows from high-voltage points to low-voltage points, and air flows from areas of high atmospheric pressures to areas of low atmospheric pressures. Similarly, linguistic "current" flows (i.e., feature interaction diffuses) from the nodes V and DP having high "potential" (i.e., full set of features) to a labeled VP bearing less "potential" (i.e., having reduced or a partial set of features).

Another reason for the flow of "current" is "feature inheritance" from a strong phase head to a weak phase head, i.e., from a light verb v to a main verb V, and from a complementizer C to a tense T (Chomsky, 2008). Such a feature transportation causes a "potential drop" (i.e., feature reduction) in v and C. A flow occurs from a place of high potential to a place of low potential; therefore, feature inheritance induces a bottom-up flow. We have assumed the following properties of the structure building.

(2) *Properties of structure building*

a. *Formation of less unified exocentric set amalgam*
A syntactic object is a set comprised of a set of phonetic, semantic, and formal features. When two syntactic objects $\{\alpha\}$ and $\{\beta\}$ merge, a new set $\big\{\{\alpha\}, \{\beta\}\big\}$ is created, which is a less unified exocentric amalgam.

b. *Formation of a more unified endocentric compound*
LA makes $\{\{\alpha\},\{\beta\}\}$ an endocentric category γ, in which a formal feature [f] is reduced from $\{\alpha\}$ and $\{\beta\}$ in $\{\{\alpha\},\{\beta\}\}$, i.e., $\gamma = \{\{\alpha\}-[f_1],\{\beta\}-[f_2]\}$. This is a feature reduction.

c. *Bottom-up interaction of features*
A feature reduction caused by LA induces the upward interaction of features.

d. *Network formation*
A sequential merge followed up by LA creates a sentential network.

An LA changes an unlabeled-undirected exocentric graph to a labeled-directed endocentric network, as shown in Figure 3.

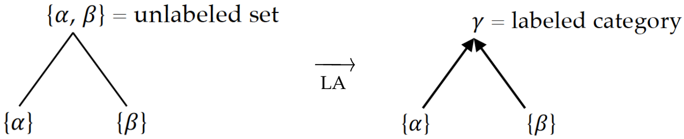


Figure 3. LA converts an unlabeled undirected exocentric graph into a labeled directed endocentric graph

The graph theory distinguishes the pre-LA and post-LA states. Contrary to Chomsky's claim, tree notions are useful for expressing sentential structures.

The remainder of this paper is organized as follows. In Section 2, we introduce the graph theory, i.e., a simple three-step version (Strang, 2016). In Section 3, we demonstrate how the graph theory can reveal a hidden force balance in simple grammatical and ungrammatical sentences. Section 4 shows that the graph theory teaches us something about the island effect. Conclusions are presented in Section 5. The Supplementary materials contain the calculation results.

2. Kirchhoff's current law governs force balance in a sentential network

We take seriously the following important tendency in nature (Strang, 2009, p. 428).

(3) Nature distributes the currents to minimize heat loss (i.e., error).

A difficult problem is what the error is relative to C_{HL}. We also assume the following general property of a network.

(4) *Properties of structure building*
A network possesses a self-organizing ability to balance the internal force in a manner such that error is minimized.

We propose the following hypothesis.

(5) *Error minimization hypothesis for C_{HL}*
A sentential network has a self-organizing ability to balance the internal force in a manner such that error is minimized.

The goal of this paper is to undertake preliminary analysis to compute the self-organizing ability of a sentential network hidden in a phrase structure and investigate whether graph-theoretical factors affect grammaticality.

The simple three-step approach is depicted as follows (Strang, 2016, p. 467). A network with nodes and edges corresponds to a network of masses and springs.

(6) *Simple three steps to uncover optimal force balance of a network*

$$\begin{tikzcd}
\boxed{u} \arrow[d, "A"'] & \boxed{f} \\
\boxed{e} \arrow[r, "C"] & \boxed{y} \arrow[u, "A^T"']
\end{tikzcd}$$

$$\begin{array}{ll}
e = Au & A \text{ is } m \text{ by } n \\
y = Ce & C \text{ is } m \text{ by } m \\
f = A^T y & A^T \text{ is } n \text{ by } m
\end{array}$$

u = Movements [potential] of n masses [nodes] $= (u_1, \cdots, u_n)$
e = Elongations [potential drop] of m springs [edges] $= (e_1, \cdots, e_n)$
y = Internal forces [current; Ohm's Law: $y = ce$] in m springs [edges] $= (y_1, \cdots, y_n)$
f = External forces [mass × gravity; KCL: $f = A^T y$] on n masses [nodes] $= (f_1, \cdots, f_n)$

Step 1 (i.e. $u \to e$) forms an incidence matrix A that expresses the geometry of a graph. Step 2 (i.e. $e \to y$) creates a conductance matrix C that measures how easily flow gets through. Ohm's Law $y = ce$ (current equals conductance times potential difference) determines a physical property c of each edge. We assign low conductance $c = 0.1$ (i.e. feature current is not easy to flow) to an XP-adjoined edge, which causes an island effect. Step 3 (i.e. $y \to f$) uses A^T (A transpose) to reveal optimal force balance hidden in the entire network, where Kirchhoff's Current Law (KCL) $A^T y = f$ (Kirchhoff (1845)) is relevant. Refer Strang (2008), Strang (2011), Strang (2016, p. 452-467) to complement this introductory section.

It is critical to point out that graph theory with KCL not only deals with the structure of an artificial object, such as an electrical circuit, it also deals with the structure of a purely mathematical and abstract geometrical graph where points are connected in various ways. We contend that graph-theoretic analysis of sentential structures is not appreciated sufficiently.

3. What does graph theory teach us about simple-sentence grammaticality?

We consider the following examples that appear to have a similar degree of structural complexity.

(7) a. He likes her.
b. * He likes she. (The intended meaning: *he* = agent, *she* = patient)

As a preliminary extension of the approach to more complex sentences, we calculate the optimal force balance hidden in island phenomena in Section 4.

3.1. Force balance hidden in a grammatical phrase structure

We demonstrate step by step how we reveal hidden force balance in a grammatical phrase structure: sample (7a). See Figure 4. The squared parts are pronounced. A viral formal feature (lower-case letters) is eliminated by its matching virus buster (uppercase letters) in a head (Piattelli-Palmarini and Uriagereka, 2004).

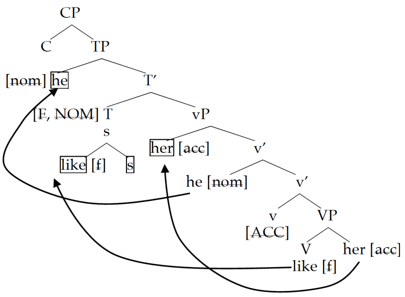


Figure 4. Grammatical phrase structure

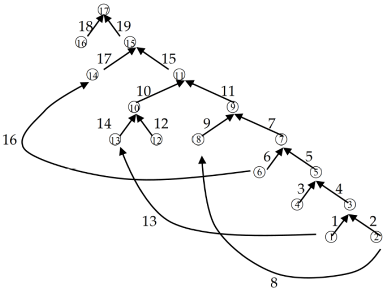


Figure 5. Graph-theoretical translation of grammatical phrase structure

We assume a set of minimal phrase-structure-building guidelines as follows.

(8) *Minimal phrase-structure-building guidelines*
a. The structure is built bottom-up.
b. The sentential heads are V, v, T, and C.
c. A set of external merge (EM; merging two terms from the structure-external Lexicon) builds a vP that contains arguments.
d. Morphological checking occurs with an internal merge (IM; structure-internal merge).
e. The sensorimotor (SM) interface externalizes one copy.

Next, we translate the single-dominance structure into a graph. See Figure 5. Assume the minimal guidelines for translating a phrase structure into a graph.

(9) *Minimal guidelines for phrase-structure-to-graph translation*

a. An IM creates a loop.
b. Nodes and edges are numbered bottom-up.
c. Matrix-clause V is numbered first.
d. Head-related nodes are numbered earlier than non-head-related nodes.
e. Head-related edges are numbered earlier than non-head-related edges.

Guideline (9a) is crucial. When node α undergoes IM, all copies of α are one and the same entity α, i.e., α appears in different places. All copies of α are related and identical. Consequently, all copies of α are connected. A loop closed by internal merge of a copy is a two-dimensional area. A graph without loops is a tree. Guideline (9b) adopts a hypothesis that structure building proceeds bottom-up. Guideline (9c) assumes that a matrix-clause predicate is the starting point of structure building. Guidelines (9d) and (9e) presuppose that a search by a virus-buster in a head is what drives IM, i.e., structural growth. The sentential graph in Figure 5 can be drawn in a plane without graph edge crossing if the graph behaves as a mobile object. KCL applies to a sentential graph because the graph is planar. A dominance relation holds in this graph-theoretic translation, and a species of the linear correspondence axiom (LCA; informally, pronounce top-down; Kayne (1994)) performs linearization in SM. Now, we translate a graph into an incidence matrix A. See Table 1.

A

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰
1	-1		1														
2		-1	1														
3				-1	1												
4			-1		1												
5					-1		1										
6						-1	1										
7							-1		1								
8		-1						1									
9								-1	1								
10										-1	1						
11									-1		1						
12										1		-1					
13	-1												1				
14										1			-1				
15											-1				1		
16						-1								1			
17														-1	1		
18																-1	1
19															-1		1

Table 1. Incidence matrix A

We use Reshish matrix calculator (RMC; matrix.reshish.com) for calculating the rank r (true size) of a matrix and for performing Gaussian elimination. For this A, $r = 16$ with computation time of $0.211\,\mathrm{s}$. The three rows are dependent, i.e., redundant. Rows are dependent when edges form a loop ((Strang, 2016, p. 453)) and independent when edges form a tree. C_{HL} inevitably form loops, i.e., C_{HL} leaves redundancy. Now, we transpose A to obtain a transpose matrix A^T. See Table 2.

A^T

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
①	-1												-1						
②		-1						-1											
③	1	1		-1															
④			-1																
⑤			1	1	-1														
⑥						-1										-1			
⑦					1	1	-1												
⑧								1	-1										
⑨							1		1		-1								
⑩										-1		1		1					
⑪										1	1				-1				
⑫												-1							
⑬													1	-1					
⑭																1	-1		
⑮															1		1		-1
⑯																		-1	
⑰																		1	1

Table 2. Transpose matrix A^T

We create a graph Laplacian matrix A^TA (A^T times A). See Table 3.

$A^TAx = 0$ is not invertible, i.e., not solvable. To solve an apparently unsolvable problem, typically we ground a node, i.e., make the node potential zero (Strang (2008), Strang (2011)). Grounding node ⓝ resembles hanging a spring-mass system at mass ⓝ from a ceiling. The following method is crucial to our analysis.

(10) *Ground-silent-IM-copy method for C_{HL}*
Ground a copy of IM that is not externalized at SM.

We ground IM-related nodes that are not externalized at SM, i.e., kinetic energy used for pronunciation is zero. Thus, we ground nodes ①, ②, and ⑥. The reaction force $S = s_1 + s_2 + s_3$ leaves grounded nodes and enters the root node. We obtain the network shown in Figure 6.

What are the linguistic and cognitive reasons for S? We speculate that SM contains a built-in "grounding" operation that makes at least one of IM-related copies phonetically zero. C_{HL} attempts to solve an apparently unsolvable problem by compelling SM to ground nodes, thereby calculating and creating an optimally force balanced struc-

A^TA

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰
①	2		-1										-1				
②		2	-1					-1									
③	1	1	3		1												
④				1	-1												
⑤			-1	-1	3		-1										
⑥						2	-1							-1			
⑦					-1	-1	3		-1								
⑧		-1						2	-1								
⑨							-1	-1	3		-1						
⑩										3	-1	-1	-1				
⑪									-1	-1	3				-1		
⑫										-1		1					
⑬	1									1			2				
⑭						1								2	1		
⑮											-1			-1	3		-1
⑯																1	-1
⑰															-1	-1	2

Table 3. Graph Laplacian matrix A^TA

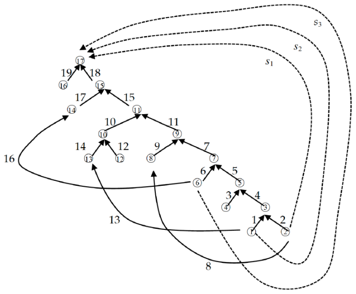


Figure 6. Reaction forces leaving grounded nodes and entering the root node

ture. SM sends it back to C_{HL}, which confirms the structural optimality and dispatches the structure with semantic features to the Conceptual-Intentional (CI) interface. C_{HL} and SM work for CI. Note that their semantic features are not zero. A remaining question is why a failure of phonetic realization in SM is sufficient to trigger grounding in C_{HL}.

Thus, nodes ①, ②, and ⑥ are reduced, i.e., they disappear from A^TA. We obtain a reduced A^TA, which we denote as A^TA_{reduced}. See Table 4.

Now, $A^TA_{\text{reduced}}x = S$ is solvable because we removed infinitely many solutions from $N(A^TA)$. RMC performs elimination and yields the following result. See Table 5.

$A^T A_{\text{reduced}}$

	③	④	⑤	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰
③	3		-1											
④		1	-1											
⑤	-1	-1	3	-1										
⑦			-1	3		-1								
⑧					2	-1								
⑨				-1	-1	3		-1						
⑩							3	-1	-1	-1				
⑪						-1	-1	3				-1		
⑫							-1		1					
⑬							-1			2				
⑭											2	-1		
⑮								-1			-1	3		-1
⑯													1	-1
⑰												-1	-1	2

Table 4. Reduced graph Laplacian matrix $A^T A_{reduced}$

	③	④	⑤	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰
③	3		-1											
④		1	-1											
⑤			5/3	-1										
⑦				12/5		-1								
⑧					2	-1								
⑨						25/12		-1						
⑩							3	-1	-1	-1				
⑪								164/75	-1/3	-1/3		-1		
⑫									101/164	-63/164		-25/164		
⑬										139/101		-25/101		
⑭											2	-1		
⑮												545/278		-1
⑯													1	-1
⑰														267/545

Table 5. Upper triangular matrix U of $A^T A_{reduced}$ *after Gaussian elimination*

The rank is $r = 14$. The computation time was $0.371s$. Finally, we solve the system and obtain the following result. See Table 6.

S consists of a set of syntactic features {{Fphon}, {Fsem}, {Fform}}, the potential of which is approximately equal to the total amount of node potential in TP. The result is consistent with a hypothesis that parsing is incremental (Hale, 2014). These accumulative features flow through those silent copies and return to the root node. Table 6 shows that potential is greatest in the root node ⑰ and the head C ⑯, i.e., $2.041S$, which is approximately twice that of TP ⑮, i.e., $1.041S$. A calculation reveals that the actual current of *S* is $S = s_1 + s_2 + s_3 = -0.999S$, which indicates that a higher node bears the cumulative potential of that of every lower node. C_{HL} recycles

Node potential	Edge current
$x_1 = 0$ (grounded)	$y_1 = -(x_3 - x_1) = -(0.022S - 0) = -0.022S$
$x_2 = 0$ (grounded)	$y_2 = -(x_3 - x_2) = -(0.022S - 0) = -0.022S$
$x_3 = 0.022S$	$y_3 = -(x_5 - x_4) = -(0.067S - 0.067S) = 0$
$x_4 = 0.067S$	$y_4 = -(x_5 - x_3) = -(0.067S - 0.022S) = -0.045S$
$x_5 = 0.067S$	$y_5 = -(x_7 - x_5) = -(0.112S - 0.067S) = -0.045S$
$x_6 = 0$ (grounded)	$y_6 = -(x_7 - x_6) = -(0.112S - 0) = -0.112S$
$x_7 = 0.112S$	$y_7 = -(x_9 - x_7) = -(0.269S - 0.112S) = -0.157S$
$x_8 = 0.135S$	$y_8 = -(x_8 - x_2) = -(0.135S - 0) = -0.135S$
$x_9 = 0.269S$	$y_9 = -(x_9 - x_8) = -(0.269S - 0.135S) = -0.134S$
$x_{10} = 0.374S$	$y_{10} = -(x_{11} - x_{10}) = -(0.561S - 0.374S) = -0.187S$
$x_{11} = 0.561S$	$y_{11} = -(x_{11} - x_9) = -(0.561S - 0.269S) = -0.292S$
$x_{12} = 0.375S$	$y_{12} = -(x_{10} - x_{12}) = -(0.374S - 0.375S) = 0.001S$
$x_{13} = 0.187S$	$y_{13} = -(x_{13} - x_1) = -(0.187S - 0) = -0.187S$
$x_{14} = 0.521S$	$y_{14} = -(x_{10} - x_{13}) = -(0.374S - 0.187S) = -0.187S$
$x_{15} = 1.041S$	$y_{15} = -(x_{15} - x_{11}) = -(1.041S - 0.561S) = -0.48S$
$x_{16} = 2.041S$	$y_{16} = -(x_{14} - x_6) = -(0.521S - 0) = -0.521S$
$x_{17} = 2.041S$	$y_{17} = -(x_{15} - x_{14}) = -(1.041S - 0.521S) = -0.52S$
	$y_{18} = -(x_{17} - x_{15}) = -(2.041S - 1.041S) = -S$
	$y_{19} = -(x_{17} - x_{16}) = -(2.041S - 2.041S) = 0$

Table 6. Node potential and edge current in the best possible force balance

potential energy (i.e., features) by compelling SM to ground silent copies. The recycled features S exit grounded nodes and enter the root node, which C_{HL} reuses for a top-down computation, i.e., parsing. An optimal force balance contains a top-down feature current.

Now we have revealed the force balance hidden in the phrase structure of *He likes her*. See Figure 7. We indicate current strength by arrow points (enlarged by a factor of 10 to make the difference among edge currents easier to see).

The sample is grammatical; therefore, C_{HL} must compute the above force balance as optimal. It is significant that current directions reverse in an optimal force balance. We speculate that structure building (the original graph) occurs bottom-up, while parsing (optimal information flow) occurs top-down. The latter corresponds to "a top down minimalist parser" that "explores a search space defined by inverting the operations of merge and move (i.e., *unmerge* and *unmove*)" (Kobele et al., 2013, p. 35). It is consistent with the statement that "grammatical categories are complex feature structures, actually calculated by the parser itself" (Hale 2014: 17). Fukui and Takano (1998) proposed a similar inverse flow, which they refer to as *demerge*, that linearizes syntactic objects top down at the SM side. We claim that such top down flows reflect a hidden self-organizing optimal force balance. C_{HL} generates an optimal force balance in which the error is minimized by eliminating two edges (i.e., edge 3, which is a head

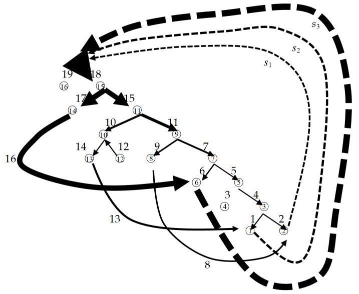


Figure 7. Hidden force balance of the grammatical sentence (7a)

projection of light verb v, and edge 19, which is a head projection of complementizer C). The optimal force balance preserves the original three independent loops. It is also significant that the current direction of edge 12 (an edge connecting two segments of V-adjoined T) is preserved.

The reaction forces $s_1(-0.521S-0.022S = -0.157S)+s_2(-0.187S-0.022S = -0.209S)+$ $s_3(-0.135S-0.112S = -0.633S)$ sum to $-0.999S$, which means that“gravitational force” $0.999S$ pulls the network down. Among the three IM-edges (8, 13, and 16), edge 16 ([nom]-IM edge; subject-raising trajectory) has greater resilience force ($0.521S$) than [acc]-IM-edge 8 ($0.135S$; object-raising trajectory) and [f]-IM-edge 13 (0.187S; V-raising trajectory). Edge 16 has approximately four times stronger current than that of edge 8 and roughly three times stronger current than that of edge 13. To use a spring-mass analogue, the entire network balances largely at ⑥, where the subject DP merges externally. Among the IMedges, edge 16 is analogous to a spring with the largest resilience, i.e., edge 16 works harder to adjust the balance of internal forces. In contrast, edge 8 (object-raising trajectory) is more symmetrical in that it is relatively optimal in the original graph. Node ⑥, where the subject DP merges externally, is a principal balance point of the entire network.

3.2. Force balance hidden in an ungrammatical phrase structure

For the ungrammatical sample (7b), we assume a phrase structure as in Figure 8.

Here, the [nom]-virus-checking fails. Consequently, the internal merge of *she* does not occur. We translate the phrase structure into a graph. See Figure 9.

The hidden force balance in the ungrammatical sample is as in Figure 10. Refer Supplementary 1 for the calculation.

Since the relevant sample is ungrammatical, C_{HL} must exclude the above self-organized force balance as not optimal for C_{HL}, i.e., the error is not minimized. Note that this force balance is optimal mathematically, i.e., it realizes its best possible equi-

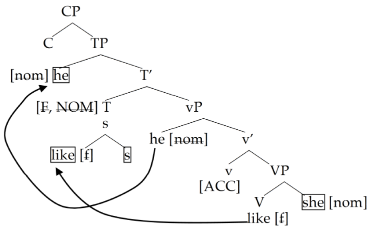


Figure 8. Phrase structure of the ungrammatical sample (7b)

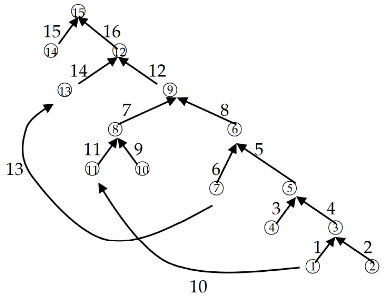


Figure 9. Graph-theoretical translation of the ungrammatical structure

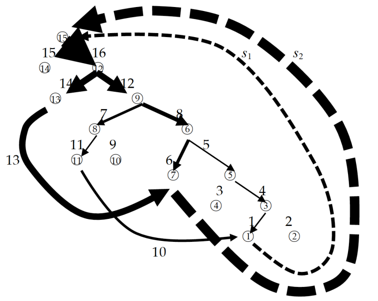


Figure 10. Hidden force balance (i.e., self-organizing ability) of the ungrammatical sample (7b)

librium and obeys KCL. However, it must contain errors that C_{HL} cannot tolerate. We consider the following as a significant observation. Unlike grammatical structure, this ungrammatical structure loses edge 2 (i.e., a complement projection of object pronoun *she*) and edge 9 (i.e., an edge connecting two segments of V-adjoined T). C_{HL} cannot tolerate the disappearance of edges 2 and 9. C_{HL} cannot delete any edge to minimize the error. We will discuss how edge disappearance contributes to grammaticality in the next section. Edge 16 (i.e., TP-to-CP projection) has the greatest resilience (–S) that pulls up the root node ⑮ to compete the "gravity." Among the two IM-edges 10 and 13, [nom]-IM edge 13 (subject-raising trajectory) has greater current force (–0.513*S*), which is approximately 3 times stronger than the other [f]-IM edge 10 (–0.18S; V-adjunction trajectory). The entire network balances principally at

	(7a) Grammatical	(7b) Ungrammatical
Number of nodes in I	17	15
Number of edges in I	19	16
Gross potential in II (S)	7.813	7.721
Absolute gross current in II (S)	4.047	3.822
Number of edge disappeared in II	2	4
— of that of I	11%	25%
Number of independent loops in I	3	2
Number of loops disappeared in II	0	0
Number of loops in II	6	4
Rank of A	16	14
Time to obtain U of A (s)	0.211	0.135
Rank of $A^{T}A_{\text{reduced}}$	14	13
Time to obtain U of $A^{T}A_{\text{reduced}}$ (s)	0.371	0.063
Absolute gross current of IM edges in II (S)	0.843	0.693
Flow direction of IM edges reversed?	Yes	Yes

(7a) *He likes her.* (grammatical) and (7b) * *He likes she.* (ungrammatical)

Table 7. Graph-theoretical properties of grammatical and ungrammatical network

⑦, where the subject DP merges externally. Here, edge 13 is likened to a spring with larger resilience.

3.3. Discussion—How are force balance and grammaticality related?

Here, we denote the original graph as I and the post-grounding-self-organized force balance as II. See Table 7.

A noteworthy difference between grammatical sample (7a) and ungrammatical sample (7b) is that edge 2 (complement projection) and 9 (an edge connecting two segments of V-adjoined T) submerge in sample (7b). An edge disappears when the two connecting nodes have no potential difference (i.e., potential drop), thereby no current flows along that edge. Both ends (nodes) of such an edge become disconnected. If a network loses an edge, it loses a structure and becomes more symmetrical. C_{HL} requires information flow from the complement DP for immunization of viral [acc] in the object pronoun *she*. Similarly, C_{HL} cannot tolerate loss of edge 9. A C_{HL} computation breaks down if no information flows between the two segments of V-adjoined T for immunization of viral [f] in V. Such a symmetry (no change) in the adjunction structure in its mathematically optimal balance must be an intolerable error for C_{HL}. Thus, C_{HL} must require a virus-checking operate through information flow. Both grammatical (7a) and ungrammatical (7b) lose strong-phase head (v, C). A descriptive generalization is as follows.

(11) *Descriptive generalization of force balance in simple structures*

a. C_{HL} generates an optimally-force-balanced network in which the error is minimized by disconnecting heads.

b. C_{HL} generates an optimally force balanced network in which the error is minimized by preserving edges that are related to viral formal feature checking.

Kayne (1984) was essentially correct in that a certain disconnection causes ungrammaticality. Why must heads disconnect in an optimal force balance in C_{HL}? We propose two possible answers for the puzzle.

(12) *Answer A*

When v and C merge with VP and TP, respectively, all features are transferred to V and T, respectively (feature inheritance; Chomsky (2008)). If feature inheritance precedes self-organization of force balance, no information flows from v and C when the force balance is optimal. The strong-phase-head projections from v and C must disappear to make the force balance optimal. The feature-inheritance hypothesis guarantees v = V and C = T. If v = V and C = T, V and T also submerge.

Answer B

Heads are highly symmetrical: they are in the best possible force balance in the first place. Heads are so stable and symmetric that they do not need to adjust the resilience to balance internal forces. C_{HL} uses heads as steady pivots of computation.

Putting aside which answer is preferable, observations seems to support the error minimization hypothesis for C_{HL}, i.e., a good sentential network hides a linguistically optimal force balance pattern. Our approach provides empirical evidence of the importance of current balances to grammaticality.

4. Does graph theory teach us anything about the island effect?

Here, we apply our analysis to more complex structures. We calculate force balance hidden in island-related structures (Ross, 1967), (Chomsky, 1973).

(13) *Island-effect-related examples*

a. Who$_1$ did John read [a story about t$_1$]?

b. * Who$_1$ did John read [a story that amused t$_1$]?

c. * Who$_1$ did [a story about t$_1$] amuse John?

d. John-wa [$_{\text{DP}}$ [$_{\text{NP}}$ [$_{\text{CP}}$ dare-o yorokob-ase-ta] kiji-o]] yon-da-no?
John-TOP who-ACC please-CAUSATVE-PAST article-ACC read-PAST-Q
'What is x, x a person, such that John read an article that pleased x?'

Sample (13a) indicates an overt wh-extraction from a complement DP, where no island effect is observed. Sample (13b) shows an overt wh-extraction from a complex DP that contains a relative clause CP, where an island effect is detected. Sample (13c) contains an overt wh-extraction from subject DP, where an island effect is observed. Sample (13d) is from Japanese, where the wh-phrase *dare* "who" is covertly extracted from a complex DP, as in (13b). Significantly, (13b) shows an island effect whereas (13d) does not. The wh-phrase is pronounced at the IMed position in (13b) while it is pronounced at the EMed position in (13d). For simplicity, we disregard an IM of an object with a projection of v at intermediate steps. Refer Supplementary 2 for the calculation.

4.1. Balance in overt wh-extraction from a complement DP (no island effect)

We assume the following structure for sample (13a), which is reproduced. See Figure 11.

(13) a. Who$_1$ did John read [a story about t$_1$]?

Unlike a pronominal complement that undergoes IM (Section 3.1), we assume that an indefinite complement DP does not undergo IM. Parentheses indicate that the term is not pronounced.

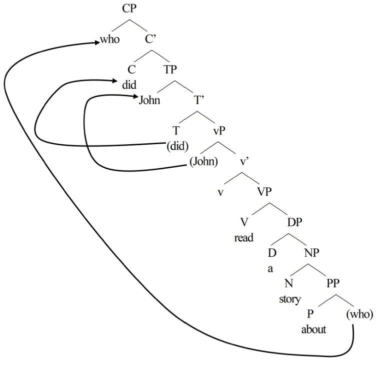


Figure 11. Overt wh-extraction from a complement DP (no island effect)

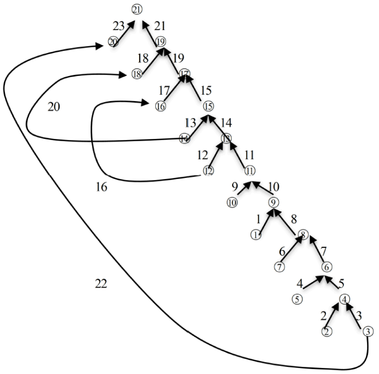


Figure 12. Graph of overt wh-extraction from a complement DP

Next, we translate the above phrase structure into the corresponding graph. See Figure 12.

The calculation reveals the following self-organizing force balance hidden in the above graph (refer Supplementary 2.1. for the calculation). See Figure 13.

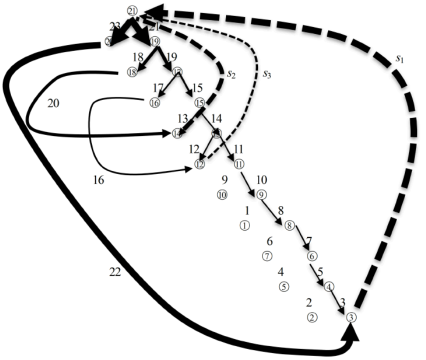


Figure 13. Balance hidden in overt wh-extraction from a complement DP (no island effect)

The optimal force balance shows top-down flow. Head edges disappear. Feature-checking-relevant edges are preserved.

4.2. Balance in overt wh-extraction from a complex DP (island effect)

We assume the following structure for sample (13b), which is reproduced. See Figure 14.

(13) b.* Who$_1$ did John read [a story that amused t$_1$]?

Why does pro not form a loop? We adopt a standard view that the feature checking of viral formal features contained in an externalized (i.e., pronounced) nominal term requires IM, which is the driving force of structural growth. We do not adopt a view in which a base-generated (i.e., externally merged) pro, which is silent, bears [nom] checked off by T by IM. Such an IM of a silent term does not contribute to the substantial structural growth. C$_{HL}$ cannot tolerate such an unsubstantial operation; thus, pro remains at the externally-merged position, where it receives a semantic feature from v. Now we translate the phrase structure into a graph. See Figure 15.

A crucial difference between (13a) and (13b) is that the latter contains an XP-adjunction structure created by edge 13, i.e. the relative-clause CP is adjoined to the DP. Unlike head-adjunction (i.e. V-to-T head adjunction in Section 3), an XP-adjunction

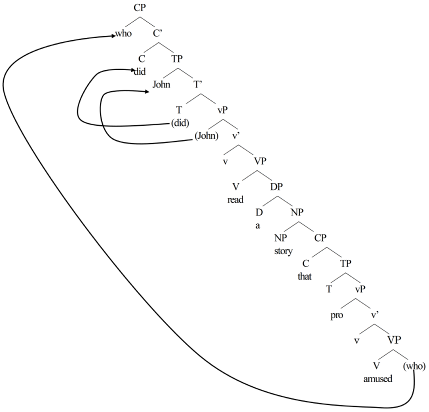


Figure 14. Overt wh-extraction from a complex DP (island effect)

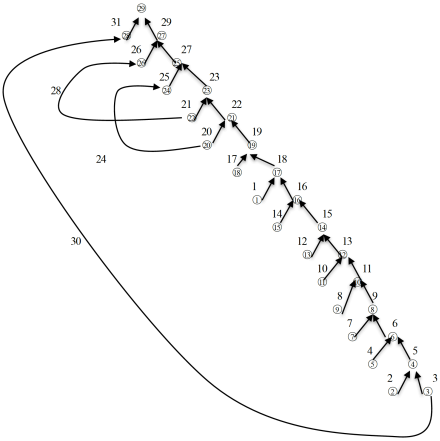


Figure 15. Graph of overt wh-extraction from a complex DP

creates an island. In particular, we assume that the conductance c of an adjoined edge is low. Let us assume that $c_{13} = 0.1$ instead of $c = 1$, which we assume for other edges. When V and NP merge, LA changes {V, NP} to VP (refer Section 1). In contrast, when CP adjoins to NP, NP embeds CP, i.e. NP contains CP. Adjoin is not Merge. If c measures edge cost as in economics (Strang, 2016, p. 458), the cost of NP-CP edge must be low because NP already contains CP. The same condition is used for calculating force balance in a Japanese example (13d) that corresponds to (13b). Refer Table 8 in Supplementary 2.2. for the reduced graph Laplacian matrix $A^T A_{\text{reduced}}$ with $c_{13} = 0.1$. The calculation (see Supplementary 2.2.) reveals the hidden force balance as in Figure 16.

Notably, a gap appears in the wh-loop, i.e. edges 15, 16, 18, and 19 that are necessary to form the wh-loop disappear in its mathematically optimal force balance. It indicates that the defective wh-loop cannot support the costly current of the wh-IM-edge 30. [1]

[1] I would like to thank an anonymous reviewer who urged me to provide a solution to an opened problem in the earlier draft, i.e. 'What causes the difference between (13a) and (13b)?', a long-standing conundrum since Ross (1967). The reviewer's request made us use a lower conductance $c = 0.1$ for an adjoined edge, which unexpectedly brought us a significant result.

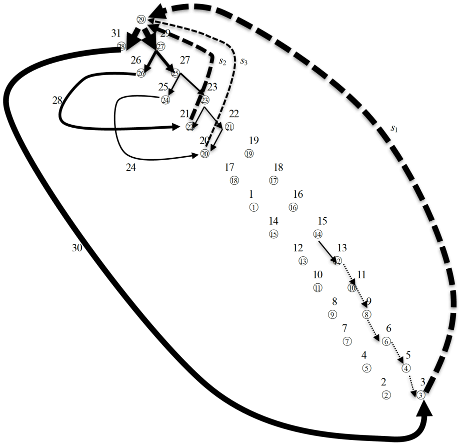
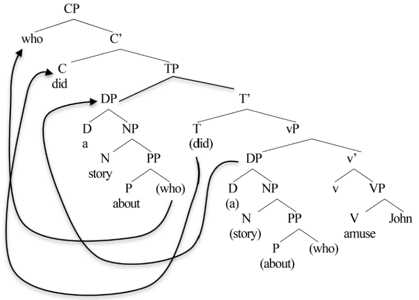


Figure 16. Balance hidden in overt wh-extraction from a complex DP (island effect)

4.3. Balance in overt wh-extraction from a subject DP (island effect)

We assume the following structure for sample (13c), which is reproduced. See Figure 17.

(13) c.* Who$_1$ did [a story about t$_1$] amuse John?



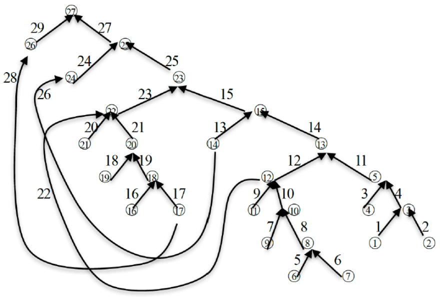


Figure 17. Overt wh-extraction from a subject DP (island effect)

Figure 18. Graph of overt wh-extraction from a subject DP

We translate this phrase structure into a graph. See Figure 18.

The calculation (refer Supplementary 2.3.) uncovers the force balance as in Figure 19.

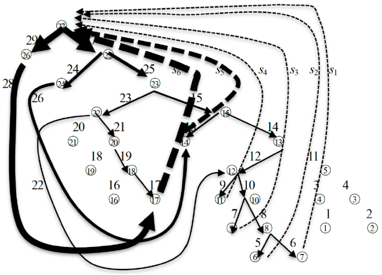


Figure 19. Balance hidden in overt wh-extraction from a subject DP (island effect)

Here, the optimal force balance shows a top-down flow. The head projection edges in the matrix clause disappear. Feature-checking-relevant edges are preserved. It is significant that head projection edges in the silent original copy of the subject island are preserved.

4.4. Balance in covert wh-extraction from a complex DP (no island effect)

We assume the following structure for a Japanese sample (13d), which is reproduced. See Figure 20.

(13) d.	John-wa [$_{\text{DP}}$ [$_{\text{NP}}$ [$_{\text{CP}}$	dare-o yorokob-ase-ta]	kiji-o]]	yon-da-no?
	John-TOP	who-ACC please-CAUSATVE-PAST	article-ACC	read-PAST-Q
	‘What is x, x a person, such that John read an article that pleased x?’			

We translate this into a graph. See Figure 21.

A calculation (refer Supplementary 2.4.) uncovers the hidden force balance as in Figure 22.

Remarkably, current on edge 30 (wh-IM edge) is $y_{30} = 0.0004S$, which is about 1240 times less than that of the corresponding overt wh-IM in English example (13b). Herein, the wh-IM edge preserves the original direction. Thus, it seems that such an infinitesimally small wh-IM current preserving the original flow direction does not require a complete wh-loop. Moreover, head projection edges disappear, including those in the complex-DP island. Feature-checking-relevant edges are preserved, except for edge 22, which is a DP-movement trajectory led by an intermediate copy of the topic phrase.

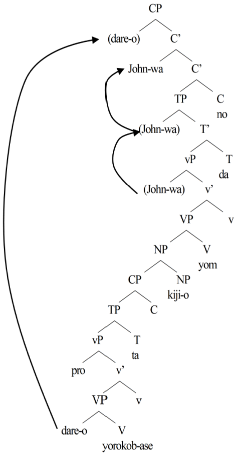


Figure 20. Covert wh-extraction from a complex DP (no island effect)

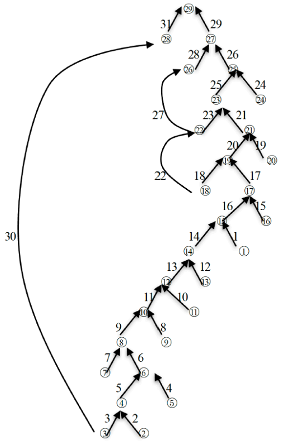


Figure 21. Graph of covert wh-extraction from a complex DP

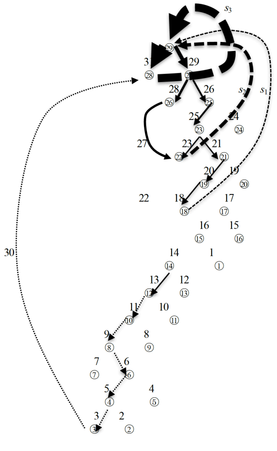


Figure 22. Balance hidden in covert wh-extraction from complex DP

4.5. Discussion—How are force balance and island effect related?

In Table 8, we highlight the properties of networks with and without the island effect. Here, the original graph and the network with self-organized force balance are abbreviated as I and II, respectively.

	(13a)	(13b)*	(13c)*	(13d)
Number of nodes in I	21	29	27	29
Number of edges in I	23	31	29	31
Gross potential in II (S)	2.933	3.157	3.181	2.141
Absolute gross current in II (S)	2.847	2.854	2.875	1.868
Number of edges disappeared in II	5	13	8	14
— of that of I	22%	42%	28%	45%
Number of independent loops in I	3	3	3	3
Number of loops disappeared in II	0	1	0	2
Number of loops in II	6	5	9	3
Rank of A	20	28	26	28
Time to obtain U of A (s)	0.419	1.029	0.36	0.138
Rank of $A^T A_{\text{reduced}}$	18	26	21	26
Time to obtain U of $A^T A_{\text{reduced}}$ (s)	0.1	0.383	0.301	0.345
Absolute current of wh-IMed edges in II (S)	0.497	0.497	0.495	0.0004
Flow direction of wh-IMed edge in II reversed?	Yes	Yes	Yes	No
Does wh-loop contain adjunction structure?	No	Yes	No	Yes

(13a): grammatical overt wh-extraction from complement DP; (13b)*: ungrammatical overt wh-extraction from complex DP; (13c)*: ungrammatical overt wh-extraction from subject DP; (13d): grammatical covert wh-extraction from complex DP

Table 8. Graph theoretical properties of island-effect-related force balance

4.5.1. Grammatical (13a) versus ungrammatical (13b)*

As an anonymous reviewer pointed out, the fact that the absolute current of wh-IMed edge in II for grammatical (13a; Figure 13) and ungrammatical (13b; Figure 16)* is identical seems to indicate that our analysis fails here. However, there is a fundamental difference between the two, i.e. (13b)* lacks a wh-loop. A crucial difference between (13a) and (13b)* is that the latter contains an XP-adjunction structure (i.e. the relative-clause CP adjoins to the DP) in the wh-loop. An important condition is that an adjoined edge bears low conductance, i.e. $c_{13} = 0.1$. Therefore, the ungrammatical structure (13b; Figure 16)* has an incomplete wh-loop with a gap. No edge means no potential difference and no current flow. An incomplete wh-loop cannot support the costly wh-IM edge bearing relatively high current ($0.497S$) that reversely flows into the original wh-copy, a position to which a semantic feature is assigned. Note that C_{HL} allows an adjunction structure itself. A non-wh-sentence containing an adjoined edge is grammatical (e.g. *'John read a story that amused Mary.'*). A calculation reveals that all EM (externally-merged)-edges unrelated to loops disappear in a tree (structure without loops), i.e. they are optimal in the first place. However, C_{HL} disallows a sentential structure constructed exclusively by EM, i.e. IM must operate in C_{HL}.

4.5.2. Ungrammatical (13b)* versus grammatical (13d)

The current difference regarding the wh-IMed edge (wh-movement trajectory) between (13b)* and (13d) is remarkable. The wh-IM current of edge 30 (wh-movement trajectory) of (13b; Figure 16)* is ~1240 times greater than that of edge 30 of (13d; Figure 22). The resilience of edge 30 in (13d; Figure 22) is extremely small ($0.0004S$; relatively close to zero) and preserves the original flow direction that guarantees wh-interpretation. The same graph-theoretical result must be realized in other "wh-in-situ" languages, such as Chinese and Korean, where a similar immunity to island effect has been observed since Huang (1982). Significantly, the wh-IM edge 30 bearing infinitesimally small current and the original direction robustly remedies a gap in a wh-loop.

For C_{HL}, (13d) is grammatical because the error (i.e., "heat loss" in wh-movement trajectory) is minimized, while (13b)* is ungrammatical because the error is not minimized. A zero-current edge is likened to an inelastic wire and is symmetrical in that it is optimal in the original graph in the first place. A similar property is found in zero-current edges growing from heads (refer Section 3). It is significant that a movement trajectory of a wh-phrase that is externalized at the original position in II behaves as a head projection edge. An extremely low cost of a wh-IMed edge with the original direction is sufficient to self-balance the entire network in wh-in-situ languages. In such languages, the cost of wh-IM (wh-movement trajectory) must be very small, which Huang (1982) predicted and observed.

Huang hypothesized that the wh-IM in wh-in-situ languages takes place after spell-out (SO, i.e. a derivational point where information is sent to SM and CI). IM after SO does not affect pronunciation, thereby ensuring zero externalization cost. However, such a hypothesis faces a problem relative to why wh-IM takes place before SO in some languages (e.g. English) and after in others (e.g. Japanese). We argue against such a wh-movement parameter. In contrast to Huang's take, we assume that wh-IM (wh-movement) takes place before SO in all languages, i.e., the structure building is the same for C_{HL} of "Homosapiensese," i.e., human natural language. It is a mathematical (linear algebraic/graph theoretical) distinction of hidden force balance that causes the contrast (13b)* vs (13d). If a current is fundamentally an error, thereby causing a heat, the relevant error is minimized to a greater degree in the network of (13d; Figure 22). More specifically, the gross potential of (13b; Figure 16)* is approximately 1.4 times greater than that of (13d; Figure 22), and the absolute gross current of (13b; Figure 16)* is roughly 1.5 times stronger than that of (13d; Figure 22).

Furthermore, the current direction of the wh-IMed edge is preserved when the wh-phrase is externalized in the original position in (13d; Figure 22), unlike (13a; Figure 13), (13b; Figure 16)*, and (13c; Figure 19)*, where the wh-phrase is externalized at a higher IMed position. Thus, wh-IM in (13d; Figure 22) is also more symmetrical relative to the direction of the information flow. It is also significant that feature-checking-relevant edges are preserved, with the exception of edge 22, which is a DP-movement trajectory that is led by an intermediate copy of the topic phrase. This comprises empirical evidence that a movement trajectory between an original copy and an intermediate copy is optimal throughout the derivation, i.e., it does not need to adjust the resilience. The principal balance point of the network in (13d; Figure 22) is ㉘, which is the target of [wh]-checking and is the closest to the root node CP. The above observations comprise evidence for the error minimization hypothesis for C_{HL}, i.e., the force balance and current (error) minimization within the entire network affects grammaticality.

4.5.3. Grammatical (13a) versus ungrammatical (13c)*

It is significant that head projection edges in the silent original copy of the subject island are preserved in (13c; Figure 19)*. C_{HL} cannot tolerate such a head-projection-edge preservation and computes that the error is not minimized. Unlike grammatical force balance in (13a; Figure 13) and (13d; Figure 22), where the balance point is either the bottom or top of the entire network, ungrammatical (13c; Figure 16)* has their balance point at an intermediate wh-copy that is neither assigned a semantic role nor is its viral formal feature checked off. Such an ontologically weak status disqualifies an intermediate copy as an optimal balance point of the entire network. These constitute additional factors that control the error minimization hypothesis for C_{HL}. Furthermore, ungrammatical (13c; Figure 16)* hides a force balance that resembles that of the ungrammatical simple sentence **He likes she* (Section 3), where the complement *she* is disconnected from the entire structure. In other words, the terms in matrix-clause v′ are disconnected from the entire structure in (13c; Figure 16)*. A certain disconnection causes grammaticality (Kayne, 1984).

4.5.4. Simple sentence versus complex sentence

One may predict that the gross potential and absolute gross current in II of island-effect-related samples must be greater than those of simple samples because the former appear to require more energy to compute more complex structures. However, this prediction fails. As Tables 7 and 8 indicate, the net potential and absolute net current in II of simple examples are greater than those of island-effect-related examples. For C_{HL}, a simple sentence is not so simple, and a complex sentence is not so complex.

4.5.5. Why does C_{HL} contain IM?

Given the above results, we see a hint relative to answering a difficult problem, i.e., why does C_{HL} contain IM? Chomsky states that we should allow ourselves to be more puzzled as to why this is so.

(14) "Displacement [IM] had always seemed—to me in particular—a curious imperfection of language. … Pursuit of SMT [strong minimalist thesis] reveals that displacement with this property of multiple interpretation ("the copy theory of movement") is the simplest case. … This is a significant discovery, I think—too long in coming, and insufficiently appreciated, as are its consequences" (Chomsky, 2015, p. x).

SMT states that the faculty of language (FL = C_{HL}) is a perfect solution to the legibility problems that the two external interfaces (i.e., the conceptualintentional (CI) and sensorimotor (SM)) impose on C_{HL}. Consider the following example with the two copies, where the lower copy is silent.

(15) Which book did John read (which book)?
'For which x, x a book, such that John read x?'

At the initial step, verb V assigns a semantic role [patient] to the original copy of *which book* (i.e., the lower variable x) when the copy EMs with V. At a later step, C IMs with *which book* (i.e., the higher wh-phrase working as the operator binding the variable x) and the sentence is interpreted as a direct wh-question in CI. Here, MC requires one copy to be externalized. The higher copy is externalized in English-type languages, whereas the lower copy is externalized in Chinese-type languages. SMT reveals that IM is the simplest possible solution to the legibility conditions that CI and SM impose on C_{HL}. Thus, Chomsky's answer is as follow.

(16) *Why did nature create IM in C_{HL}?*
Nature created IM in C_{HL} because IM was the simplest way to balance multiple interpretation in a sentence. (Chomsky's answer)

In this paper, we add a graph-theoretic reason as to why C_{HL} contains IM, noting that IM creates loops. A crucial question to ask at this point is as follows.

(17) Do we require loops for interpretability of *any* syntactic structure? If we do, there must be loops in a sentential structure. This has thick implications for syntax.

Suppose that the following assumptions hold.

(18) a. A sentential-structure building uses an IM.

b. An IM creates a loop.
c. A sentential structure is a graph generated by C_{HL}.
d. A graph possesses balance that obey KCL to equilibrate the internal force.
e. Loops are solutions to KCL (Strang (2009), Strang (2011), Strang (2016)).

Graph theory, which is an application of linear algebra, standardly maintains assumptions (18d) and (18e). The minimalist program assumes (18a). If (18b) and (18c) hold, which is a perspective that is contra-Chomsky (2014), a sentential structure must contain loops to balance the internal force. However, more loops do not mean more optimal force balance. If C_{HL} tolerates and interprets within a certain threshold of a force-balance state, and loops are solutions to KCL, C_{HL} must require a certain pattern of force balance containing loops for interpretability in any syntactic structure. Specifically, an unpronounced IM-copy, whose phonetic externalization is determined to zero by SM to answer the legibility problems posed by C_{HL}, corresponds to a grounded node in a graph necessary for solving an apparently unsolvable problem. IM may be a built-in grounding operation that nature has created in the C_{HL}-SM interface. Information (i.e., linguistic features) flows around in a sentential network. C_{HL} needs IM to optimally self-balance the internal force in a sentential network. "What are the actual solutions to [KCL] $A^Ty = 0$? The currents must balance themselves. The easiest way is to flow around a loop" (Strang, 2016, p. 456). IM may have emerged in C_{HL} because IM was the easiest way to balance currents in a sentential network. We answer Chomsky's puzzle as follows.

(19) *Why did nature create IM in C_{HL}?*
Nature created IM in C_{HL} because IM was the easiest way to balance currents and minimize errors in a sentential network. (Our answer)

5. Conclusions

In structure building, when a union set is labelled by LA (Chomsky (2013)), edges become directed, i.e., features flow upward. Contra Chomsky (2014), who claims that we should abandon graph notations in C_{HL} research, we claim that we must maintain graph notations. A graph theory equipped with KCL provides insight into grammaticality.

A significant concept that we adopt is "nature distributes the currents to minimize the heat loss (i.e., error)" (Strang, 2009). A sentential network generated by a natural object C_{HL} minimizes the error, which corresponds to what SMT refers to as a perfect solution to the legibility problems. Thus, we propose the error minimization hypothesis for C_{HL}: a good sentential network IM creates possesses a self-organizing ability to balance the internal force in a manner such that error is minimized.

We adopt Strang's simple-three-step approach of graph theory to uncover a hidden force balance in any network. Step 1 is a "geometry" step, where we translate a sentential graph (translated from a phrase structure) into an incidence matrix A. Step

2 is a "physics" step, where we investigate edge conductance matrix C. We assume that C is the identity matrix unless an edge involves XP-adjunction structure, in which case we assume $c = 0.1$. Step 3 is a "balance" step, where we use KCL $A^T y = f$ to uncover a hidden force balance in a sentential network. Here, the relevant matrix is $A^T A$ (a graph Laplacian matrix), which appears in various areas of mathematics relative to error minimization.

We calculated the hidden force balance in simple and island-effect-related sentences that are both grammatical and ungrammatical. C_{HL} minimizes errors by (a) converting bottom-up flow (structure building) to top-down flow (parsing), (b) removing head projection edges, (c) preserving edges related to feature checking, (d) deleting DP-movement trajectories headed by an intermediate copy, (e) ensuring that covert wh-movement trajectories have infinitesimally small currents and conserving flow directions, and (f) robustly remedying a gap in wh-loop by using infinitesimally inexpensive wh-internally-merged (wh-IM) edge with the original flow direction. The C_{HL} compels the sensorimotor (SM) interface to ground nodes such that Kirchhoff's current law (a fundamental balance law) is satisfied. Internal merges are built-in grounding operations at the C_{HL}-SM interface that generate loops and optimal force balance in sentential networks.

Acknowledgements

I would like to thank two anonymous reviewers for their valuable comments. The first reviewer urged me to provide a solution to an opened problem in the earlier draft, i.e., What are the graph-theoretical factors that distinguish a non-island property of a complement (non-complex) DP and island property of a complex DP containing a relative-clause CP? As hinted in the previous version, I used a lower conductance $c = 0.1$ for an adjoined edge in the latter and unexpectedly obtained a significant result. The first reviewer led us to this crucial factor. The second reviewer anticipated that the Chomskian graph-theoretical techniques are difficult to understand for people working with syntactico-semantic techniques based on lexicalized and dependency tools. However, Chomsky and his colleagues after (Chomsky, 2013) proposed to banish graph theoretical approach (e.g. (Kayne, 1984)) from the research field of language system (C_{HL}) within the minimalist (biolinguistic) program. I believe I can convince the readers that graph theory is a valuable tool to uncover the precise syntactic structure of a sentence. I would like to thank Enago (www.enago.jp) for the English language review. Finally, I would like to express my gratitude to Martin Popel, Hiroaki Sano, and Tomokazu Sano for spending their valuable time to convert my Word files to the LaTeXformat, which I am a novice at.

Bibliography

Berwick, C. Robert and Noam Chomsky. *Why Only Us: Language and Evolution*. MIT Press, 2016.

Chomsky, Noam. Conditions on Transformations. *A Festschrift for Morris Halle*, 1973.

Chomsky, Noam. On phases. *Current Studies in Linguistics Series*, 45:133, 2008.

Chomsky, Noam. Problems of projection. *Lingua*, 130:33–49, 2013.

Chomsky, Noam. Problems of Projection: Extensions, 2014. URL `https://www.youtube.com/watch?v=Icv_sCsIu6A`. A lecture at Olomouc Linguistics Colloquium.

Chomsky, Noam. *The minimalist program*. MIT press, 2015. 20th anniversary edition.

Fukui, Naoki and Yuji Takano. symmetry in syntax:merege and demerge. *Journal of East Asian Linguistics*, 7:27–86, 1998.

Hale, T. John. *Automaton Theories of Human Sentence Comprehension*. CSLI Publications, 2014.

Huang, C.-T. James. *Logical relations in Chinese and the theory of grammar*. PhD thesis, MIT, 1982.

Kayne, S. Richard. *Connectedness and Binary Branching*. Foris, 1984.

Kayne, S. Richard. *The Antisymmetry of Syntax*. MIT Press, 1994.

Kirchhoff, (von Studiosus. Ueber den Durchgang eines elektrischen Stromes durch eine Ebene, insbesondere durch eine Kreisförmige [On the transit of an electric current through a plane, in particular through a circular]. *Annalen der Physik und Chemie*, 64:487–514, 1845.

Kobele, M. Gregory, Sabrina Gerth, and John Hale. Memory resource allocation in top-down minimalist parsing. *Formal Grammar: 17th and 18th international conferences*, pages 32–51, 2013.

Piattelli-Palmarini, Massimo and Juan Uriagereka. The immune syntax: the evolution of the language virus. *Variation and Universals in Biolinguistics*, pages 341–377, 2004.

Ross, John Robert. *Constraints on variables in syntax*. PhD thesis, MIT, 1967.

Strang, Gilbert. Computational Science and Engineering I. Lecture 12: Graphs and Networks, Lecture 13: Kirchhoff's Current Law, 2008. URL `https://ocw.mit.edu/courses/mathematics/18-085-computational-science-and-engineering-i-fall-2008/index.htm`. Opencourseware. Cambridge, Massachusetts: MIT.

Strang, Gilbert. *Introduction to Linear Algebra, Fourth Edition*. Wellesley-Cambridge Press, 2009.

Strang, Gilbert. Linear Algebra. Lecture #12: Graphs, Networks, Incidence Matrices, 2011. URL `https://ocw.mit.edu/courses/mathematics/18-06sc-linear-algebra-fall-2011/`. Opencourseware. Cambridge, Massachusetts: MIT.

Strang, Gilbert. *Introduction to Linear Algebra, Fifth Edition*. Wellesley-Cambridge Press, 2016.

Address for correspondence:
Koji Arikawa
karikawa@andrew.ac.jp
Department of English and Intercultural Studies,
St. Andrew's University (Momoyama Gakuin)
1-1, Manabino, Izumi, Osaka, 594-1198

PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 112 APRIL 2019 83–101

Design of a Multiword Expressions Database

Pavel Vondřička

Charles University, Faculty of Arts, Institute of the Czech National Corpus, Praha, Czechia

Abstract

The paper proposes design of a generic database for multiword expressions (MWE), based on the requirements for implementation of the lexicon of Czech MWEs. The lexicon is aimed at different goals concerning lexicography, teaching Czech as a foreign language, and theoretical issues of MWEs as entities standing between lexicon and grammar, as well as for NLP tasks such as tagging and parsing, identification and search of MWEs, or word sense and semantic disambiguation. The database is designed to account for flexibility in morphology and word order, syntactic and lexical variants and even creatively used fragments. Current state of implementation is presented together with some emerging issues, problems and solutions.

1. Introduction

Multiword expressions (MWEs) have been long in the focus of the theoretical linguistics as well as NLP, especially during recent years. Breaking the seemingly clear borderline between lexicon and grammar, they obstruct each successful NLP task based on this traditional dichotomy.[1] As much as their actual definition differs, in its widest meaning reaching from proper names or fixed idiomatic expressions with non-compositional meaning to light verb constructions or seemingly free, but actually statistically idiosyncratic collocations,[2] so differ also the applications and implementations dealing with them. A particular language and the more or less limited scope of view also determine the complexity of their description and identification.

[1] As presented already by Sag et al. (2002).

[2] Sag et al. (2002) speak generally about *institutionalized phrases*, but also about simple *statistical affinity* – the phenomenon seems to be broader.

© 2019 PBML. Distributed under CC BY-NC-ND. Corresponding author: pavel.vondricka@ff.cuni.cz

Cite as: Pavel Vondřička. Design of a Multiword Expressions Database. The Prague Bulletin of Mathematical Linguistics No. 112, 2019, pp. 83–101. doi: 10.2478/pralin-2019-0003.

The PARSEME survey on MWE resources (Losnegaard et al., 2016) shows that many MWE lists and lexicons are still limited to contiguous sequences of words or lemmas, and that has also been the case of some approaches to Czech MWEs.[3] However, the limitations of this method become quickly apparent especially in Slavic languages with relatively free word order. Treatment of non-contiguous MWEs has therefore been lately addressed more intensively, although it has been called a "challenge" by Savary (2008).

In the current project *Between Lexicon and Grammar*[4] we aim at a complex description of Czech MWEs, targeting both various NLP applications and human users at the same time. The resulting database must therefore cover many diverse, incompatible or even contradictory requirements. Our goal is also to cover not only the explicit and exact use of established MWEs, but also as many as possible of their variants and unusual modifications appearing in various texts, where the creativity of language users[5] seems rather unlimited.

This goal does not only require some treatment of the non-contiguity, but also treatment of variable word order as well as variable lexical members. Components of MWEs may be more or less freely inflected or modified, omitted or even replaced by some rather unusual lexeme, while the core meaning of the MWE is still kept. Restrictions on the morphology and its various irregularities within MWEs have already been addressed by many projects,[6] but treatment of lexical variability seems to be a rather rare case. It has been addressed (at least to some degree) for example by Villavicencio et al. (2004) or Grégoire (2010) in the project DuELME, which obviously also takes into account difference in occurrence (frequency) of the particular variable components. The approach of Al-Haj et al. (2013) offers a very simple and elegant solution both to the lexical variability and the variability of word order at the same time. Nevertheless, our project still aims at even greater flexibility in the identification of MWEs or even their fragments.

2. Flexibility of MWEs and their core

As shown by Hnátková et al. (2017) and Jelínek et al. (2018), even seemingly fixed MWEs may appear in different variants or modifications. Some of their components may be optional and many components may vary or be modified. Often it is enough to keep just a core or fragment of the original MWE in order to recall the original meaning and construct a metaphor or other form of word play, where the rest of the

[3] See e.g. Pala et al. (2008).

[4] For more details see the articles by Hnátková et al. (2017), Hnátková et al. (2018) and Jelínek et al. (2018).

[5] Sometimes obviously caused also by their lack of knowledge of the established form or its correct language use.

[6] See Savary (2008); Oflazer et al. (2004); Al-Haj et al. (2013); Czerepowicka and Savary (2018), etc.

MWE may be twisted or replaced with some other construction in any way.[7] It is thus important to record, how every component of the MWE varies in the common use and how it may vary potentially.[8] In addition, it is useful to identify the minimal core components (fragments) that are able to identify the original meaning even if the rest of the components is missing or replaced. We have also found examples of multiple minimal cores which can convey the meaning; any of them may be used by the language user.[9]

All these possibilities of MWEs, which go far beyond the possibilities of simple words, raise the question of the identity of a MWE. As long as the meaning of the original MWE contributes to the message of the text, it should be considered its part and detected in the process of parsing. That should probably apply also in cases, where the identification of the MWE and its original meaning is not a necessary condition for the recipient to understand the message of the text; the use of a modified MWE may as well be just a part of the art form (e.g. to express irony, humour) or the authors desire to make an impression of wittiness and creativity, and not necessarily part of the message itself.

On the other hand, there is often a possibility of overlap with the literal use of the same combination of words. Such a risk varies a lot among different MWEs and depends mostly on the amount of their anomalies. The identification and disambiguation of MWEs thus remains a very difficult task in many cases.

3. MWE entry: identity

The question of identity of a MWE opens the question of the identity of one single entry in the database. We consider the meaning as the main criterion for the distinction of single entries. The expression *jít přes čáru* 'cross the line' can have three different meanings, depending on the meanings of the noun *čára*: 1) the literal meaning commonly appearing in sports, where lines often demarcate a play field; 2) the colloquial meaning of borderline between countries, referring commonly to the phenomenon of illegal emigration to the Western Europe during the communist rule in

[7] This has been illustrated by the example of the biblical quote 'it is easier for a camel to go through the eye of a needle than for a rich man to enter into the kingdom of God' encountered in various forms and allusions in the data of the Czech National Corpus, such as: *'a camel would rather enter into a kingdom of heaven than* I would pass a thread *through the eye of the needle*', 'the bypass should be threaded through the area like *the camel through the eye of a needle*', '*it is easier to go through the eye of a needle than* to get access to EU funds' or 'Klaus forces two elephants to be pulled *through the eye of a needle*'. (Hnátková et al., 2017)

[8] Of course, the potential variability is actually unlimited, but observed variants may be presented according to real occurrence in corpus data.

[9] E.g. the expression *hodit flintu do žita* (lit. 'throw the rifle into the rye (field)', meaning 'throw in the towel', 'give up'), can either appear in forms such as 'throw [something] into the rye' (e.g. 'throw the camera into the rye', i.e. 'abandon the career of a photographer') or 'throw the rifle [somewhere]'. (Hnátková et al., 2017)

Czechoslovakia; 3) the most abstract meaning of any negative phenomenon exceeding some acceptable threshold. The literal meaning should not be covered by the database, while the other two meanings should probably have two independent entries.

However, the form of the MWE remains an important factor as well. In combination with the high variability of many MWEs, it may become difficult in many cases to decide whether some expression is still a variant of another expression with the same or very close meaning, or whether it should be considered an independent MWE. The pragmatic factor can therefore motivate the distinction of two or more partly identical MWEs with the same meaning[10] if one single (merged) description of the variability would be too complex. This concerns especially complicated dependencies among several components of a MWE which must or must not appear together. Some basic examples motivated by syntactic alternations are presented in Section 10 and those are still kept within a single entry, but more complex cases can be encountered.

Another question concerns possible derivations of MWEs – e.g. passivization or nominalization of a verbal MWE. We do not want to create separate entries for such derivations, unless their meaning, usage or behaviour differs significantly. Instead, the possibilities or restrictions imposed on the common types of derivation should be defined for every MWE entry, where such a possibility can be expected by the grammar.

4. MWE entry: structure

4.1. Requirements

A MWE consists of two or more components, understood – by definition and the name itself – as *words*. These "words" may be more or less fixed: some components may be realized by one particular word (lexeme) only, some by a choice of several different words (lexemes); some can be formed just by any lexeme or even a whole phrase of a particular type, just like any standard valency element.[11] Some components may be freely inflected or modified, while others are subject to various restrictions. The MWE therefore needs a definition by means of its components and their various possible realizations. It is important for us to be able to describe the features

[10] The meaning of MWEs is also often considerably more complex than meaning of simple words. The meaning of each individual component (word) still contributes to the meaning of the whole expression even for relatively fixed expressions, and the variations may always modify it to some degree.

[11] Like single word units, MWEs take valency elements as well. Some MWEs are just verbal phrases, where some of the valency elements are filled by fixed expressions while others remain open. Some MWEs can also take valency elements which none of the components would require or allow by itself. This has already been presented by Hnátková et al. (2017) on the example of the expression *dát na srozuměnou, že…* 'to make clear that…' (lit. 'to give on understanding that…'), where the valency slot for *that*-clause is not specified by any of the component parts.

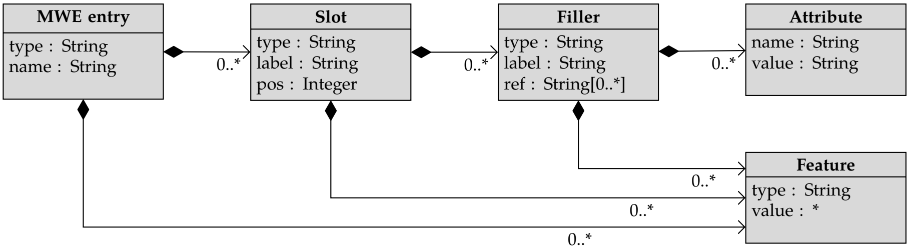


Figure 1. MWE entry structure (basic model)

and restrictions of the MWE, its components and their respective possible realizations, independently.

Since we want to describe the entries and their components both for the purpose of NLP parsing and for human users, there will be a need to record very different types of features. Different features also concern different levels of description. While valency, usage type or syntactic type are features of the whole MWE, internal modifiability concerns rather its components, and lexical idiomaticity is a feature specific to the particular lexeme used as a component. Some features may apply to several levels: style or register usually depend on the use of a particular lexeme, if there is a choice of several varaints. However, it may also be the feature of the whole MWE: the expression *mít něco na háku* ('couldn't care less', lit. 'to have st. on the hook') is rather colloquial despite of the fact that it consists solely of standard words – the metaphorical use is not standard anymore.

4.2. Entry structure

For full flexibility, we define the entry pattern by means of *slots* and *fillers*, common terms used for this type of description in computational linguistics.[12] The entry itself consists of *slots* and *features* referring to the MWE as a whole. Slots represent the single components of the MWE (pattern), which is the syntagmatic dimension of the MWE. Slots consist of *fillers* and the slot-specific *features*. Fillers represent the paradigmatic dimension of the components: the possible variants which may be used to realize the particular component. The primary role of fillers is to represent actual (terminal) tokens to be matched in the data.

In the process of annotation of the data of the Czech National Corpus, the experimental identification of MWEs has been applied as the last step, on top of the tokenized, lemmatized, tagged and desambiguated texts. The current parser FRANTA

[12] The basic principles follow (in a simplified form) the proposal for a structured lexical description as described by Vondřička (2014).

`lemma="ryba"`
`tag="NNF[SP]1"`
	noun *ryba* ‘fish’
	in nominative singular or plural

`lemma="stát"`
`tag="V"`
	verb *stát* ‘stand / cost / happen’
	in any form

`tag="AA"`	any common adjective in any form

Table 1. Example definitions of positional attributes for different types of fillers

uses a combination of the positional attributes *lemma* and morphological *tag* to match token patterns in the data. The patterns have been defined in a special MWE list called FRANTALEX. This list has been used as the primary source of initial data for the MWE database. Therefore we define the fillers by means of a combination of positional attributes (*lemma* and *tag*) and their values that must be matched in the text in order to identify the MWE. However, the fillers may actually declare just any arbitrary positional attributes used to identify the matching tokens. Other restrictions, such as possible word order, modifications or transformations, can be defined by means of additional features. Figure 1 shows the scheme of the whole entry structure.

The attributes to be matched may also be underspecified: the *tag* value may contain just a prefix referring to the part-of-speech or a regular expression to match a custom choice of acceptable morphological forms. Specification of the *lemma* may be completely avoided in cases when just any lexeme of some particular part of speech or morphological category may fill the position, but its presence is still necessary (or typical) for the identification of the MWE (see Table 1 for examples). Of course, the filler may provide its own additional *features* as well.

For strictly fixed expressions, a slot will mostly contain only one possible filler defining the particular type of token to be matched. More flexible expressions may contain a list of several synonymous or otherwise alternative fillers. Since the fillers may also have their own features, it is possible to document their actual relative usage (e.g. by terms of corpus frequency) or further individual effects on the other slots or on the MWE as a whole. Such slots can be classified as *fixed* or “closed”. In case of relatively *open* slots, the fillers may be underspecified as mentioned above. They may also represent only the most typical representatives of a relatively open semantic class. That is relevant in cases where such a group of acceptable fillers cannot be fully defined formally in an explicit way. We call this third kind of slots *semi-open*. Of course, such incomplete description can currently only be of limited use for a NLP parser, but it will still remain a useful hint for human users of the lexicon.

If we want a slot or filler to represent a whole phrase of some type (e.g. in the case of valency elements), we cannot use a combination of positional attribute values to match one single token anymore. We need to use specific features to define the phrase type (restriction) instead. Such description probably cannot be directly used by a simple low-level parser such as FRANTA, but it can be useful for human users and later also for possible higher-level parsers operating also at the syntactic level.

4.3. Classification of features

Features are generic pairs of *type* (name) and *value*.[13] For easier organization and systematization of various types of features, we use a hierarchical system of specification of the *type* by means of a path in an arbitrary hierarchy of features, using colon as the separator. At the top level, features are classified as morphological, syntactic, semantic, statistical, related to the form, purely user- or editor-oriented notes, etc. Further levels are divided as needed: as specific groups of features, by particular theory, source of data, etc. This also allows us to store multiple similar features from different sources (or for different purposes) at the same time.

In case we need to include multiple alternative values of some type of feature, custom subspecification may be used. This applies especially to user notes, examples from real texts or statistical values. For example, the basic type of features for absolute frequency `:stats:fq:abs` is expected to be extended by additional custom subspecification of the corpus (and possibly subcorpus) used to acquire the frequency value, e.g. `:stats:fq:abs:BNC:fiction`. This allows the database to be searchable by features both using underspecification of the type (by means of a path prefix) or its full (sub)specification as needed.[14]

5. Multi-level and multi-purpose utilization of the structure

The flexible design allows for multi-purpose utilization of the entry components. Features can easily be used (and classified) both for purely technical purposes of NLP processing tools and to store information aimed at human users of the database, such as definitions, examples or notes.

Information may also be provided at several levels of description, also in parallel if needed: surface restrictions on the form or occurrence of particular components (such as those presented in Section 10) may actually result from regular alternations on higher levels (e.g. syntax), but at the time of the initial import of patterns from the FRANTALEX list, this information is provided just in the form of simple surface rules, restricting the possible occurrence of particular tokens (forms) in the sentence, and must be later reinterpreted manually in order to obtain a more appropriate, higher-level linguistic description. On the other hand, new MWEs created manually in the

[13] For the purpose of effective implementation and searching, a single feature record may actually have several values of different type in the database: string value, numeric value, etc. Currently, we do not want to utilize this technical property of implementation in the data model, but keep the feature as a purely atomic property.

[14] The possibility of custom subspecification makes it possible to add additional information to the value of the feature and it thus makes the atomicity of the feature object rather ostensible.

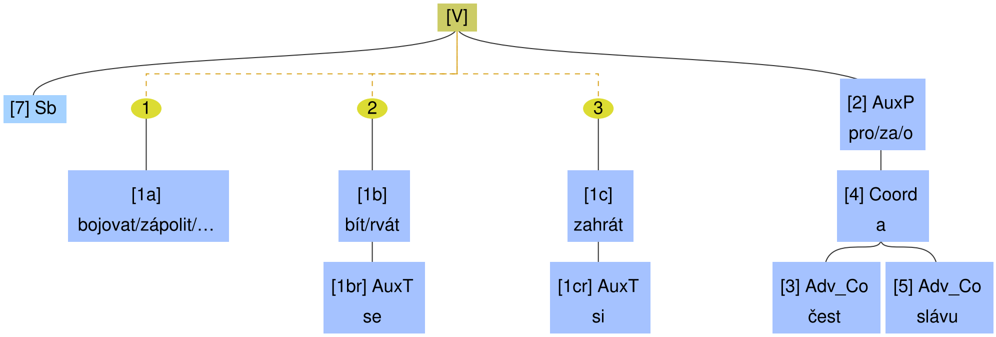


Figure 2. Dependency structure for the expression ***bojovat za čest a slávu***

database or imported from other sources – and already provided with the more abstract linguistic descriptions – must also allow for an automatic reinterpretation of the information back into the form of the (low-level) surface rules the parser is actually able to process and apply to the data lacking any higher-level annotation. This may actually also apply to such basic phenomena as grammatical agreement. Future advanced parsers may possibly utilize the higher-level descriptions (such as syntactic relations and alternations) more directly.

6. Representation of tree structures

In the database, it is desirable to capture tree structures such as dependency and constituency structure of an expression. Dependency relations between the components can easily be recorded in the form of slot features. One single feature is needed as reference to the parent slot and another one to identify the syntactic function of the component. Such relations can easily be projected into the resulting tree structure, as illustrated in the Figure 2 showing a combination of dependency tree with variant subtrees (explained later in Section 8).

However, constituency trees need non-terminal nodes and we need to be able to refer and assign features to them as well. Therefore, they should be represented by standalone objects in the database, equivalent to the slots. That is the reason why just grouping the components by means of features would not be a satisfactory solution.

The flat structure of slots and fillers does not allow for nesting. Previously, we have suggested to use recursive structures for lexical descriptions, where fillers can branch into further sequences of "subslots".[15] However, indexing and querying recursive data structures is still a very demanding task not well supported by the current database and search engines. Therefore the idea was abandoned. Instead, we decided

[15] See Vondřička (2014) for more details.

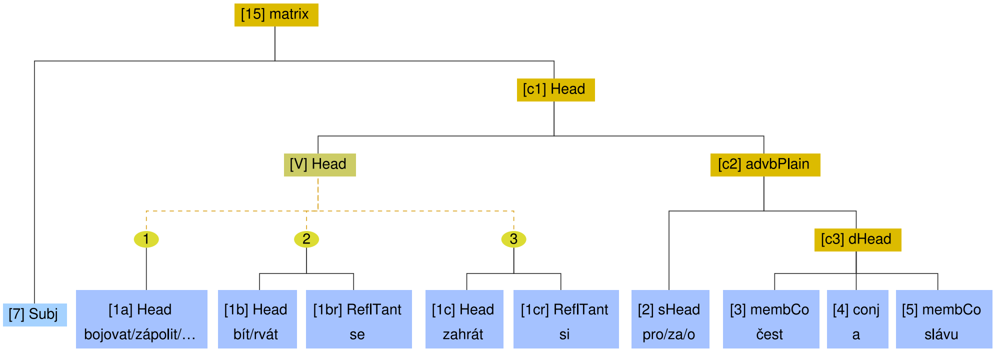


Figure 3. Constituency structure for the expression ***bojovat za čest a slávu***

to keep to the flat structure, but to allow fillers to refer to a sequence of other slots by means of their identifiers (labels). This brings the possibility to add non-terminal fillers (and their respective non-terminal slots). Again, the tree structure can easily be reconstructed by nesting the slots additionally, as seen in Figure 3.

Various advantages and disadvantages emerge from this design: indexing and searching for both terminal and non-terminal nodes is equally simple, but traversing relations between them in a single query is not supported by the search engine. That means that searching for MWEs by their structure – e.g. by syntactic (or other) relations – would be difficult to implement. Currently, we do not expect the need to search the database by tree structures, but in case this would be necessary, the structures can be reconstructed for all entries, encoded into some kind of searchable patterns and indexed separately by the same or a more appropriate engine.[16] Another advantage is the possibility to record several independent tree structures within a single entry, which corresponds well to the requirement of multifunctionality. A partial disadvantage is the potential need for treatment of possible partial trees, overlapping trees and orphan nodes.

7. The treacherous term “word”

The most problematic issue of “multi-word entities” is the fact that the term is based on the linguistically not well defined concept of “word”. Relying solely on the orthographic aspect of using space or punctuation marks as boundaries between “words” is very treacherous even in languages using Latin alphabet.[17] This can be well demonstrated by the English example of the triple acceptable spelling “airstream”,

[16] E.g. a graph database.

[17] This issue has been well discussed e.g. by Savary (2008).

"air stream" and "air-stream". Similar phenomena concern also Czech and become especially urgent when dealing with many colloquial MWEs, where there is no standard established for their spelling.[18]

A tendency to split composed words in Czech seems quite obvious lately, probably by the influence of English spelling which does not make this distinction. On the other hand, there has traditionally been an opposite tendency of merging more or less established prepositional phrases in the standard language: a continuum between already established adverbs (or prepositions) such as *včas* ('in time'), *dohromady* ('together') and less established combinations such as *na příklad/například* ('for example'), where both spellings are still in use, and *do ztracena/doztracena* ('(peter out) to nothing'), where the single-word spelling is still much less common, despite the fact that the noun alone can be extremely rarely encountered in other contexts. The situation becomes especially unstable in case of many colloquial exclamations, such as *pro Boha/proboha!* ('for God('s sake)!'), which usually only appear expressively in direct speech. The borderline between words and MWEs becomes quickly very unclear, and it proves much more as a problem from the practical perspective of low-level NLP parsers, rather than from the theoretical point of view of linguistics.

Such alternatives cannot be simply automatically merged in the source text either, since ambiguity may still exist as seen in the example (1). There is often a tendency to make a distinction between the adverbial, prepositional or particle meaning (such as 'for example') and the original literal meaning ('at example/exercise') by merging the words together as in the variant (1-a), in analogy to other already established adverbs of this type, and also to avoid confusion like in the example (1-b). However, this is not always the rule and the less common or lexicalized the combination is, the more unpredictable the spelling is. In such cases, individual factors such as education, age and conservatism play an important role.

(1) a. Podívejte se například na příklad číslo 7.
Look for example at example number 7.
'For example, look at the example/exercise number 7.'
b. Podívejte se na příklad na příklad číslo 7.

The problem also arises when dealing with standard multiword components of MWEs such as reflexive verbs or other analytic forms. The system of slots and fillers (representing always a single token only) can only deal with single-word alternations, but it cannot deal with alternation of non-reflexive verbs with reflexive verbs requiring an additional reflexive pronoun. One possible solution is to use non-terminal slots for such variants again.[19] In this case, the fillers do not represent the terminal to-

[18] Not to mention the fact, that the CNC project aims at retaining and mapping also non-standard phenomena in the language, including those classified by many people as "mistakes".

[19] Another solution is to define separators as fully-fledged components of the MWE, which may be omitted (Savary, 2008; Czerepowicka and Savary, 2018), or to define MWE components at the level of mor-

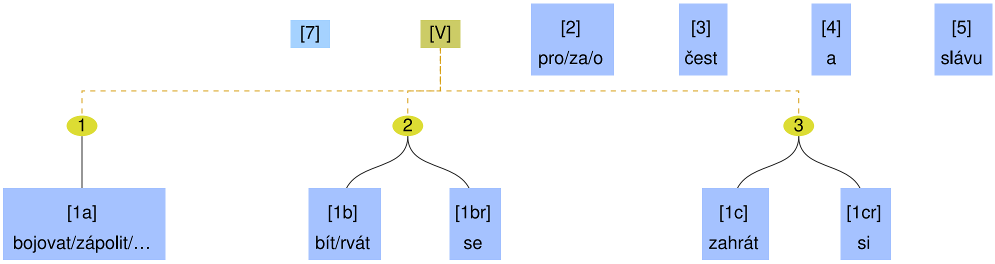


Figure 4. Variants of the expression ***bojovat za čest a slávu***

kens directly, but refer to other (terminal) slots or sequences of several slots that may alternate. Since the order of slots does not necessarily represent a real word order, even more complex alternations or dependencies can be defined by means of these non-terminal slots.

If we extend the possibility of reference beyond the limits of the entry itself, we can also describe MWEs containing (embedding) other MWEs, such as *držet/viset (jen) na čestné slovo* (lit. 'hold/hang (just) on a word of honour', meaning 'fixed/fastened in an unreliable, wonky way') embedding the MWE *čestné slovo* ('word of honour').

8. Alternatives requiring a different number of tokens in practice

The variability beyond the limits of simple words can be demonstrated on the example of the expression *bojovat za čest a slávu* ('fight for honor and glory'). Three different prepositions can be used in this expression (*za*, *pro* and *o*) and these can be listed as fillers of a single slot. Unfortunately, this does not apply to all the verbs which may also alternate here. The variant tree-structure of the expression can be seen in Figure 4.

Beside simple verbs, reflexive verbs can also appear in this expression, and these require a reflexive particle – i.e. an additional token. For this reason, they cannot fit into a simple list of alternative terminal fillers representing a single token anymore. In addition, both types of Czech reflexive verbs can occur here: those using reflexive pronoun in accusative case (e.g. *bít se*) and those using reflexive pronoun in dative case (e.g. *hrát si*).[20] For this purpose we have created a non-terminal "variant-slot" labeled `V`, representing all the various verbs, with three fillers (shown in the figure as

phemes (Al-Haj et al., 2013). However, this would be difficult to implement in our case, where the identification of MWEs is currently applied at the end of the liguistic analysis of data, which have previously been processed by a tokenizer and tagger unaware of the existence of MWEs.

[20] The question may be raised whether the meaning can really be considered identical in all these particular cases, but that does not change the situation in principle.

elliptical nodes with numbers 1, 2 and 3) for the three classes of verbs: the first one refers to the terminal slot listing simple verb fillers only, the second one refers to the sequence of slots listing the reflexive verb(s) requiring accusative and the reflexive pronoun in accusative itself, and the third one refers to the sequence of slots listing the verbs requiring dative and the reflexive pronoun in dative itself. All the other slots are terminal slots and remain orphans in this partial tree-view. Slots are labelled arbitrarily, by default by numbers in the order of addition.[21]

The variants branching into their own subtrees also make the visualization of syntactic trees more complicated. As shown in Figure 3, the alternating subtrees fit quite well into the constituency tree, at least as long as the variants correspond to the syntactic subtrees of the whole expression. However, their visualization within dependency trees is in principle impossible without adding a third dimension to the scheme: the non-terminal node breaks the principle of direct dependency between terminals, since it represents several terminals at the same time and their dependencies cannot point to all of them individually (see Figure 2). Therefore, the verbal dependencies need to point to the non-terminal node as their parent. The non-terminal node branches again into the three different verb groups it represents, but this is not a relation of dependency anymore. Their reflexive particles may then depend on the verbs directly again. In the scheme, we try to visualize the different type of relation again by means of different type of lines between the non-terminal node and the verbs.

9. Non-terminal slots

The database thus currently uses two types of non-terminal slots: slots for complex variants (multi-token alternations) and slots for non-terminal nodes within the constituency structure. Technically, only the fillers can actually represent terminals or non-terminals, but we want to avoid mixing terminal and non-terminal fillers in a single slot, so that the slots can also be clearly classified as the terminal and non-terminal nodes they are supposed to represent.

Slots representing valency elements – i.e. whole phrases of some type as described in Section 4.2 – represent actually a third type of non-terminal slots in the database, even though they do not refer to other components within the entry itself.

10. Internal dependencies

Several types of internal dependencies between the components of a MWE have already been encountered, which make the process of parsing and MWE identification more complex. One of them concerns MWEs using some lexeme repeatedly – these may also have variants or modifications concerning the repeated lexeme itself. The expression *Bůh dal, Bůh vzal* (lit. 'God gave – God took') can be used in various mod-

[21] In the figure, the slot labeled by No. 7 represents a generic subject of the verbal phrase.

ifications such as *život dal, život vzal* ('life gave, life took'), *čas dal, čas vzal* (time), *stát dal, stát vzal* (state/government), etc. All the modifications are based on a repeated lexeme which may vary itself. Therefore we need to define the lemma of the consequent slot (filler) as a reference to the lemma actually used in the first slot. For this purpose we currently define a special placeholder[22] with reference to another slot as the value of the filler's lemma. In case we encounter more complex dependencies of this type, we might need to find another appropriate solution.[23]

Another type of dependencies concerns optional components. These are often projections from a higher level, such as syntactic alternations. However, on the surface level of the parser they need to be specified as well. In the expression *mít NĚCO pro (svou) (vlastní) potřebu* ('to have ST. for (one's) (own/personal) use') both the possessive pronoun and the adjective are optional, but at least one of them must be present to specify the possessor of the 'use'. The expression *naložit NĚCO na NĚČÍ bedra* ('load ST. on SO.'s shoulders') alternates with the form *naložit NĚKOMU NĚCO na bedra*: on the surface level the addressee can either be expressed by an indirect object in dative, or as the possessor (attribute) of the 'shoulders'. If we want to define the MWE in one single entry, we must define both the indirect object and the possessive adjective as optional components, but indicate their mutual exclusivity in some way. The optimal solution to this kind of problems is still in discussion.

The variability of verbs in the example in Section 8 is another example of a low-level projection from a higher level, where the necessity of an additional token – the reflexive particle – would probably be easier to declare as a lexical or syntactic feature of the reflexive verbs.

11. Minimal fragments

It has also been mentioned in Section 2 that MWEs can also take part in some text in the form of creatively used fragments.[24] These fragments may go far beyond the common limits of variability or optionality of the MWEs components. It seems therefore useful to list the minimal combinations of components which have been proved to be sufficient to trigger the meaning of the MWE even if it has been heavily modi-

[22] We currently use the form `${`*target-slot-label*`}`

[23] The expression *hlava nehlava* (lit. 'head non-head', meaning 'without any regards'), where the repeated form is negated, has been discussed as an example of a prototype of a more general pattern applicable in theory to any other word (specifying closer the addressee of the (lack of) 'regards') as well. However, changing the base lexeme would also imply a change of the meaning, so that this construction in the generalized form belongs rather to the domain of grammar or some kind of 'multi-word word-formation', rather than to the lexicon directly.

[24] As demonstrated in detail by Hnátková et al. (2018) and Jelínek et al. (2018).

fied. For this purpose we use a separate feature containing the list of slot identifiers which are capable to represent a minimal core fragment of the MWE.[25]

12. Available sources of data

The primary source of MWEs for the database is its predecessor, FRANTALEX. Rather than a proper database, it is a list of about 36000 patterns (simple rules) for the parser FRANTA. This parser has been used to identify and tag MWEs in the corpora of contemporary Czech within the Czech National Corpus. These patterns have mostly been based on the descriptions included in the traditional Czech Phraseological Dictionary (Čermák et al., 1983–2009), but they have been extended by actual observations of the corpus data: common variants of the MWEs missed by the parser or incorrectly identified combinations of the same tokens having their original literal meaning (false positives).

As mentioned before, FRANTA identifies MWEs as combinations of particular tokens identified by their lemma and morphological tag, with limited possibilities to define restrictions on gaps between them and the acceptable variability of the word order. Because of the simplicity of the parser, which operates at the surface level of a morphologically tagged text only, and the relatively free word order in Czech, many of the patterns actually identify different variants of one and the same MWE, and in some cases even variations with a different word order only. Such patterns must therefore be manually combined into single (but more complex) descriptions, before they can be imported into the new database as base for new MWE entries. As long as the FRANTA parser or a similar surface-level tool is used to identify the MWEs in the corpora, we must also be able to reverse the process and generate all the alternative rules from the merged complex entries in the new database with updated information.

Additional process is used to generate syntactic structures (dependency and constituency) for the existing patterns, both those imported from FRANTALEX and those created manually. They are generated by a syntactic tagger[26] trained on the data of the Prague Dependency Treebank[27] for each MWE, manually checked and added to its entry in the database. The constituency structures are then created by conversion from the manually corrected dependency structures. The dependency relations are added as features to the existing terminal slots, while constituency structures require adding new slots for non-terminal nodes.

[25] This solution is very similar to the more general solution presented by Al-Haj et al. (2013), which is also used to describe variable word order as well as optional and alternative use of different components and their mutual surface dependency.

[26] See Martins et al. (2013).

[27] See Hajič et al. (2018).

Other sources of MWEs, light-verb constructions (LVC) and named entities (NE) based on real data are also available from the development and annotation of the Prague Dependency Treebank and related projects.[28] These offer also a higher-level annotation, but they are to some degree limited to the texts of the PDT. Overlap with the primary source can also be expected and the possibilities of utilizing and merging the different sources will need closer inspection.

13. Practical issues

The desire for a multi-purpose resource uniting various different sources of data brings some unavoidable issues or pitfalls to be resolved in order to keep consistency of the data across the lexicon. The first one is the variability of the data sources based on different approaches and with different goals in mind. While the FRANTALEX database is a set of raw surface patterns based on actual observations of MWE variability in the annotated texts of the Czech National Corpus, the syntactic annotation offers higher-level abstractions of many of these observations and variations. However, the parsing and identification of MWEs will still need to be applied to syntactically unparsed raw data, and the need to project the higher-level abstractions from the database to the surface dependencies in the form of simple rules will remain necessary.[29]

The generic database structure also offers several possible solutions to many phenomena. Again, we can take the difference between a higher level classification and a surface description from FRANTALEX as an example: while a valency dictionary would define a valency slot as a phrase of some type, e.g. a prepositional phrase specifying the preposition and the case of the nominal phrase to be used (i.e. one single open non-terminal slot in the database), FRANTALEX will provide a pair of components (terminal slots): the fixed preposition and an open slot for a noun, possibly marked as 'open for modification'.

The pragmatic approach of the FRANTALEX database may thus be in conflict with the desire for theoretical purity (systematicity) and conceptual consistency. This will need to be resolved in order to make the database a unified resource.

Another issue is the dependency of the database on the current state of tokenization and morphological annotation of the data to be parsed for MWEs. The database must also try to account for possible common mistakes in morphological analysis or disambiguation. In the case of rule-based disambiguation, this may result in a circular dependency: the identification of MWEs depends on the morphological disambiguation, and the disambiguation may depend on the identification of MWEs. This problem concerns especially the issue with the variability of words alternatively split or merged by various language users. The morphological tagger currently cannot be

[28] E.g. Vallex (Lopatková et al., 2016; Kettnerová et al., 2012) or SemLex (Straňák, 2010).

[29] As already mentioned in Section 12.

expected to tag a combination of two words in the same way as one single composed word, especially if there is real danger of ambiguity: the single word *například* can (or must) be unambiguously tagged as a particle, but the phrase *na příklad* can safely[30] be analyzed as a combination of a preposition and a noun only, at least as long as a closer and unambiguous syntactic analysis of the whole sentence is not available.

The dependency of the fillers on a particular tagger or tagset can (to some degree) be reduced by defining multiple positional attributes (even virtual or planned future attributes) prefixed by some kind of "namespace" in a similar manner as the classification of feature types.

14. Implementation and user interface

The database has currently been implemented as a part of a more generic database of corpus annotation units, sharing a common infrastructure and principles. Elasticsearch is used as backend engine for searching and storing the entries in the form of JSON documents. A data model written in Python is used as an intermediate abstraction, providing a generic API.

The API also provides management of metadata about all object types stored in the database. The types of entries, slots and fillers can be classified in the same way as the types of the features. Each object type can also be provided with descriptions and definition of its contents, access restrictions and visualization hints, requirements on subspecifications, methods of editing and presentation for different users and purposes and so on. Definitions of the features may, for example, specify a particular type of value(s), so that basic input validation and appropriate searching criteria may be automatically applied by the interface.

Current frontend user interface is designed using the Angular.js and Bootstrap frameworks. It uses the API and the metadata to create customized and highly configurable user interface on the fly. We expect it to be able to present and visualize the data in different ways suitable for different types of users. Other user interfaces can also be created for more specialized purposes, using the generic API.

15. Conclusion

We have described the generic framework used to encode and manage the database of Czech MWEs and the principles of their encoding for various purposes. The main innovation is its open and flexible structure, aimed at multiple levels of description and multiple purposes, including linguistic description aimed at human users and the effort to cover also creative use and modifications of established MWEs in real lan-

[30] Underestimation of these distinctions leads frequently to wrong analysis in many taggers with excessive interpretative ambitions. Morphological taggers simply cannot be as smart as some linguists would like them to be.

guage use, both theoretically and formally. Several problems of the encoding strategy and their possible solutions have been discussed.

The classification of MWEs applied in this project and the actual contents of the database entry have been previously described in detail by Hnátková et al. (2017). We have not dealt here with the generation of the morphology of MWEs either, since in our project the MWEs are currently being identified in texts analyzed previously. A possible deeper integration of the MWE database into the process of disambiguation and parsing of Czech textual data remains an open question for further research.

Acknowledgements

This paper is part of the project *Between Lexicon and Grammar* (2016–2018), supported by the Grant Agency of the Czech Republic (reg. no. 16-07473S), and the implementation of the *Czech National Corpus* project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

Bibliography

Al-Haj, Hassan, Alon Itai, and Shuly Wintner. Lexical Representation of Multiword Expressions in Morphologically-complex Languages. *International Journal of Lexicography*, 27(2): 130–170, 12 2013. ISSN 0950-3846. doi: 10.1093/ijl/ect036. URL https://doi.org/10.1093/ijl/ect036.

Czerepowicka, Monika and Agata Savary. SEJF – A Grammatical Lexicon of Polish Multiword Expressions. In Vetulani, Zygmunt, Joseph Mariani, and Marek Kubis, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 59–73, Cham, 2018. Springer International Publishing. ISBN 978-3-319-93782-3.

Grégoire, Nicole. DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1):23–39, Apr 2010. ISSN 1574-0218. doi: 10.1007/s10579-009-9094-z. URL https://doi.org/10.1007/s10579-009-9094-z.

Hajič, Jan, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Hajičová, Jiří Havelka, Petr Homola, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jarmila Panevová, Lucie Poláková, Magdaléna Rysová, Petr Sgall, Johanka Spoustová, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. Prague Dependency Treebank 3.5, 2018. URL http://hdl.handle.net/11234/1-2621. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Hnátková, Milena, Tomáš Jelínek, Marie Kopřivová, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová, and Pavel Vondřička. Eye of a Needle in a Haystack. Multiword Expressions in Czech: Typology and Lexicon. In Mitkov, Ruslan, editor, *Computational and Corpus-Based Phraseology: Second International Conference, Europhras 2017, London, UK, November 13–14,* *2017, Proceedings*, volume Lecture Notes in Computer Science, vol. 10596, pages 160–175, Cham, 2017. Springer International Publishing. ISBN 978-3-319-69805-2. doi: 10.1007/978-3-319-69805-2_12. URL `https://doi.org/10.1007/978-3-319-69805-2_12`. ISBN: 978-3-319-69805-2.

Hnátková, Milena, Tomáš Jelínek, Marie Kopřivová, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová, and Pavel Vondřička. Lepší vrabec v hrsti nežli holub na střeše. Víceslovné lexikální jednotky v češtině: typologie a slovník. *Korpus – gramatika – axiologie*, (17/2018): 3–22, 2018. ISSN 1804-137X.

Jelínek, Tomáš, Marie Kopřivová, Vladimír Petkevič, and Hana Skoumalová. Variabilita českých frazémů v úzu. *Časopis pro moderní filologii (Journal for Modern Philology)*, 100(2): 151–175, 2018. ISSN 0008-7386 (Print), ISSN 2336-6591 (On-line).

Kettnerová, Václava, Markéta Lopatková, and Eduard Bejček. The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In Fjeld, Ruth and Julie Torjusen, editors, *Proceedings of the 15th EURALEX International Congress*, pages 434–443, Oslo, 2012. Department of Linguistics and Scandinavian Studies, University of Oslo.

Lopatková, Markéta, Václava Kettnerová, Eduard Bejček, Anna Vernerová, and Zdeněk Žabokrtský. *Valenční slovník českých sloves VALLEX*. Karolinum, Praha, 2016. ISBN 978-80-246-3542-2.

Losnegaard, Gyri Smørdal, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. PARSEME Survey on MWE Resources. In Calzolari, Nicoletta (Conference Chair), 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)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Martins, A., M. Almeida, and N. A. Smith. Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers. In *Annual Meeting of the Association for Computational Linguistics – ACL*, pages 617–622, August 2013.

Oflazer, Kemal, Özlem çetinoğlu, and Bilge Say. Integrating Morphology with Multi-word Expression Processing in Turkish. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, pages 64–71, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1613186.1613195`.

Pala, Karel, Lukáš Svoboda, and Pavel Šmerk. Czech MWE Database. In Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. `http://www.lrec-conf.org/proceedings/lrec2008/`.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword Expressions: A Pain in the Neck for NLP. In Gelbukh, Alexander F., editor, *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, London, 2002. Springer.

Savary, Agata. Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, (1-2):1–53, 2008. URL `https://hal.archives-ouvertes.fr/hal-01023019`.

Straňák, Pavel. *Annotation of Multiword Expressions in The Prague Dependency Treebank*. PhD thesis, Univerzita Karlova v Praze, Prague, Czech Republic, 2010.

Villavicencio, Aline, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. Lexical Encoding of MWEs. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, pages 80–87, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1613186.1613197`.

Vondřička, Pavel. *Formalized contrastive lexical description: a framework for bilingual dictionaries.* LINCOM GmbH, München, 2014. ISBN 978-3-86288-428-5.

Čermák, František et al. *Slovník české frazeologie a idiomatiky (SČFI)*, volume 1–4. Academia/Leda, Prague, 1983–2009.

Address for correspondence:
Pavel Vondřička
`pavel.vondricka@ff.cuni.cz`
Institute of the Czech National Corpus
Faculty of Arts, Charles University
nám. Jana Palacha 1/2
CZ-11638 Praha 1, Czech Republic

PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 112 APRIL 2019

INSTRUCTIONS FOR AUTHORS

Manuscripts are welcome provided that they have not yet been published elsewhere and that they bring some interesting and new insights contributing to the broad field of computational linguistics in any of its aspects, or of linguistic theory. The submitted articles may be:

- long articles with completed, wide-impact research results both theoretical and practical, and/or new formalisms for linguistic analysis and their implementation and application on linguistic data sets, or
- short or long articles that are abstracts or extracts of Master's and PhD thesis, with the most interesting and/or promising results described. Also
- short or long articles looking forward that base their views on proper and deep analysis of the current situation in various subjects within the field are invited, as well as
- short articles about current advanced research of both theoretical and applied nature, with very specific (and perhaps narrow, but well-defined) target goal in all areas of language and speech processing, to give the opportunity to junior researchers to publish as soon as possible;
- short articles that contain contraversing, polemic or otherwise unusual views, supported by some experimental evidence but not necessarily evaluated in the usual sense are also welcome.

The recommended length of long article is 12–30 pages and of short paper is 6–15 pages.

The copyright of papers accepted for publication remains with the author. The editors reserve the right to make editorial revisions but these revisions and changes have to be approved by the author(s). Book reviews and short book notices are also appreciated.

The manuscripts are reviewed by 2 independent reviewers, at least one of them being a member of the international Editorial Board.

Authors receive a printed copy of the relevant issue of the PBML together with the original pdf files.

The guidelines for the technical shape of the contributions are found on the web site `http://ufal.mff.cuni.cz/pbml`. If there are any technical problems, please contact the editorial staff at `pbml@ufal.mff.cuni.cz`.