Literal occurrences of multiword expressions: quantitative and qualitative analyses

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

University of Tours, Paris-Diderot & Aix-Marseille (France), Tübingen (Germany), Basque Country (Spain), Athena Research Center (Greece)

IPIPAN, Warsaw, 14 January 2019

Multiword expressions

- Word combinations, which exhibit lexical, syntactic, semantic, pragmatic and/or statistical **irregularities**.
- Pervasive feature: **non-compositional semantics** - the meaning of an MWE cannot be deduced from the meanings of its components, and from its syntactic structure, in a way deemed regular for the given language.
  - (PL) ***mieć muchy w nosie*** 'to have flies in one's nose'⇒'to be bad-tempered'

Multiword expressions

Heterogeneous nature of MWEs

- compounds 'złożenia'
  - (PL) ***na przykład*** 'for example', ***panna młoda*** 'young maid'⇒'bride'
- complex terms 'terminy wielowyrazowe'
  - (PL) ***układ scalony*** 'integrated circuit'
- multiword named entities 'wielowyrazowe jednostki nazewnicze'
  - (PL) ***Europejski Bank Odbudowy i Rozwoju*** 'European Bank for Reconstruction and Development'
- light-verb constructions 'analityzmy werbo-nominalne'
  - (PL) ***podjąć decyzję*** 'make a decision'
- phrasal verbs 'czasowniki frazowe'
  - (EN) *to* ***make up*** *for sth* 'nadrobić coś'
- idioms 'idioms'
  - (PL) ***mieć muchy w nosie*** 'have flies in one's nose'⇒'to be bad-tempered'
- proverbs 'przysłowia'
  - (PL) ***nie wywołuj wilka z lasu*** 'don't tempt your fate'

Literal occurrences of MWEs

Example

(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)

SOA

- Using the **interplay between literal and idiomatic readings**, and their distributional and statistical properties, to discover how idioms are stored and processed in human mind [Cacciari and Corradini(2015)]?
- Links between literal and idiomatic readings can inform us which **morpho-syntactic variation** is allowed or prohibited by some MWEs [Sheinfux *et al.*(2017), Pausé(2017)]
- Distinguisthing literal and idiomatic readings as one of major **challenges in MWE-related NLP** [Constant *et al.*(2017)]
- Using context to **automatically distinguish literal and idiomatic occurrences** [Peng *et al.*(2014), Peng and Feldman(2016)]

Quiz: What is a literal occurrence?

(1) The boss was **pulling** the **strings** from prison.

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

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

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

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

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

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

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

(9) The strings pulled the bridge.

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

What is a literal occurrence?

- Idiomatic occurrences (IOs)

(1) The boss was **pulling** the **strings** from prison.

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

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

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

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

- Literal occurrences (LOs)

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

- Coincidental occurrences (COs)

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

(5) Determine the maximum force you can $\underline{\text{pull}}_{1,2}$ on the $\underline{\text{string}}_1$ so that the $\underline{\text{string}}_2$ does not break.

(9) The strings pulled the bridge.

- Out of scope (different lexemes)

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

Literal occurrence – a definition

Given a MWE e with components $e_1, \ldots, e_n$, a *literal occurrence* (LO) of e is a co-occurrence e' of words $e'_1, \ldots, e'_n$ fulfilling the following conditions:

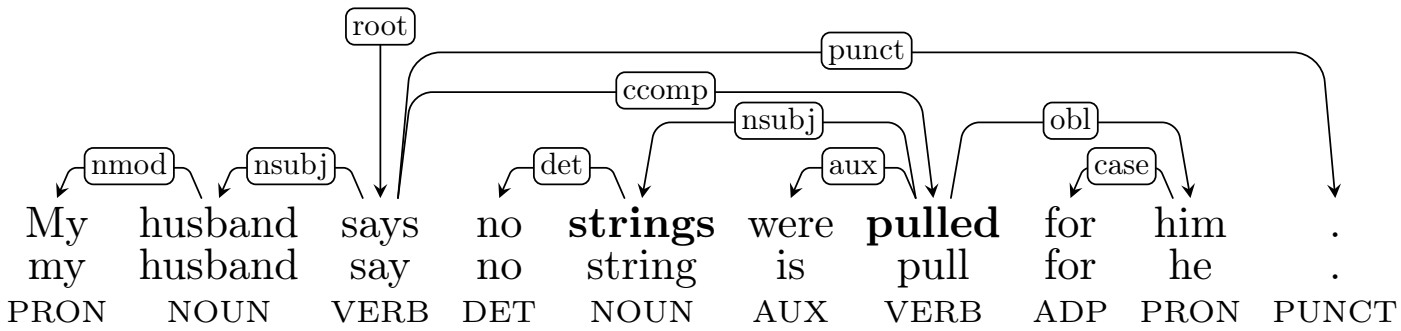
1. $e'_1, \ldots, e'_n$ can be attributed the **same lemmas and parts of speech** as $e_1, \ldots, e_n$.
2. The syntactic **dependencies** between $e'_1, \ldots, e'_n$ are the same or **equivalent** as in the canonical form of e^a
3. e' is **not a MWE occurrence**

---

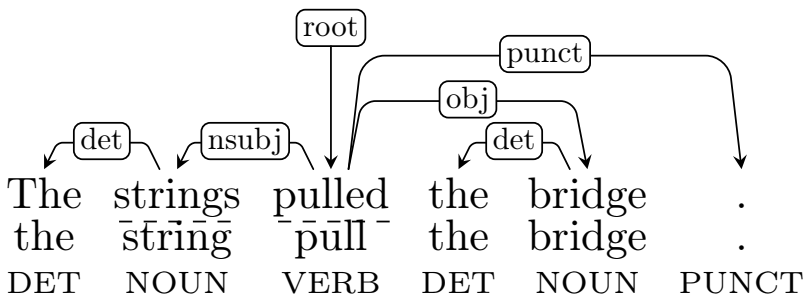
[a] **Canonical form**: the least marked syntactic form preserving the idiomatic meaning, here: *the boss **pulled strings***. A form with a finite verb is less marked than one with an infinitive or a participle, the active voice is less marked than the passive, etc. A CF is expected to capture the semantic roles 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.

Importance of the canonical form

(a)



(b)



Same dependency (nsubj) between *strings* and *pulled* but (b) is not a LO of (a).

Research questions

- Focus on **verbal** MWEs (VMWEs) – frequent discontinuity, ambiguity and flexibility
- **Quantify** the LO phenomenon:
  - relative frequency of LOs with respect to IOs and COs
  - distribution of this distribution across different VMWE types and categories
- Study **cross-lingual** aspects of LOs:
  - cross-lingually valid reasons for LOs to occur
  - language-specific reasons
  - studied languages: Basque, German, Greek, Polish and Portuguese

Data

PARSEME corpus of verbal MWEs

[Savary *et al.*(2018), Ramisch *et al.*(2018)]

- coordinated effort of 20 language teams
- unified terminology, typology and annotation guidelines
- corpus of 20 languages, 6,000,000 words, 80,000 annotated VMWEs

Corpus 1.1

Language	**Sentences**	**Tokens**	**VMWE categories**						**Tagset**
			All	VID	LVC	IRV	VPC	Others	
Basque	11,158	157,807	3,823	20%	80%	0%	0%	0%	UD+BT
German	8,996	173,293	3,823	36%	8%	8%	48%	0%	UD
Greek	8,250	224,762	2,405	27%	71%	0%	5%	0%	UD
Polish	16,121	274,318	5,152	10%	40%	44%	0%	6%	UD
Portuguese	27,904	638,002	5,536	20%	46%	16%	0%	0%	UD

VMWE typology (v. 1.1)

- **Universal** categories (valid for all languages):
  - verbal idioms (**VIDs**)

    wyciągnąć nogi 'stretch legs'⇒'die'
- light verb constructions (**LVCs**)
  - **LVC.full**: ***mieć miejsce*** 'have place'⇒'take place'
  - **LVC.cause**: ***dać prawo*** 'give right'⇒'grant right'
- **Quasi-universal** categories (valid for many languages):
  - inherently reflexive verbs (**IRVs**)

    oglądać się *(na innych)* 'watch oneself (on others)'⇒'count on (the others)'
  - verb-particle constructions (**VPCs**)
    - **VPC.full** *(EN) to* ***do in*** 'to kill'
    - **VPC.semi** *(EN) to* ***eat up*** 'to eat completely'
  - multi-verb constructions (**MVCs**) – mainly in Asian languages

Automatic extraction of LO candidates

Heuristics

For each annotated VMWE extract all (non-annotated) sequences containing the same lexemes (lemmas+POS), under 4 heuristics:

- **WindowGap** – matched tokens are separated by no more than **2 gaps**
- **BagOfDeps** – matched tokens form a **weakly connected graph** (arc directions and labels are ignored)
- **UnlabeledDeps** – matched tokens form a **connected graph** (labels are ignored)
- **LabeledDeps** – matched tokens form a **connected graph** and the dependency **labels are identical** as in the VMWE

Manual annotation of LO candidates

Annotation categories

E = annotated VMWE, C = candidate LO of E,

- ERRORS – E is not a VMWE, C is in fact a VMWE, C is a non-verbal MWE, C has wrong lexemes
- COINCIDENTAL – the dependencies are not preserved
- LITERAL – the dependencies are preserved but the idiomatic meaning is lost
  - LITERAL-MORPH – LO that could be automatically distinguished from an IO by checking morphological constraints
  - LITERAL-SYNT – LO that could be automatically distinguished from an IO by checking syntactic constraints
  - LITERAL-OTHER – LO that could be automatically distinguished from an IO only by checking more elaborate constraints (e.g. semantic, contextual)

Results of the heuristics (task of finding LOs)

Lang.	WindowGap			BagOfDeps			UnlabeledDeps			LabeledDeps			All		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
EU	0.05	0.94	0.05	0.07	0.72	**0.06**	0.06	0.50	0.06	**0.07**	0.18	0.05	0.05	**1.00**	0.05
DE	0.08	0.78	0.07	0.12	0.90	0.11	0.13	0.90	0.11	**0.14**	0.77	**0.12**	0.09	**1.00**	0.08
EL	0.11	0.86	0.10	0.15	0.88	0.13	0.15	0.80	**0.13**	**0.16**	0.51	0.12	0.11	**1.00**	0.10
PL	0.30	0.96	0.23	0.43	0.75	0.27	0.49	0.69	**0.28**	**0.52**	0.22	0.15	0.27	**1.00**	0.21
PT	0.14	0.98	0.13	0.17	0.62	0.14	0.20	0.59	0.15	**0.34**	0.37	**0.18**	0.13	**1.00**	0.11

- The heuristics are skewed towards high recall.
- Automatic identification of LOs, given lemmas and dependencies, is a hard task.
- Sliding window and dependency-based statistics are complementary.

Results of the manual annotation

	DE	EL	EU	PL	PT
Annotated MWEs (IDIOMATIC)	3823	2405	3823	5152	5536
Candidates from at least one heuristic	926	445	2618	384	1997
ERRORS	820	268	1394	65	1058
COINCIDENTAL	24	126	1082	207	668
LITERAL	79	51	131	105	258
↪ LITERAL-MORPH	7	24	66	7	73
↪ LITERAL-SYNT	14	10	40	27	44
↪ LITERAL-OTHER	58	17	25	71	141

Idiomaticity rate

$$CoRate_{CAT} = \frac{|CO_{CAT}|}{|CO_{CAT}|+|LO_{CAT}|+|IO_{CAT}|} \quad IdRate_{CAT} = \frac{|IO_{CAT}|}{|LO_{CAT}|+|IO_{CAT}|}$$

IdRate in Polish

Category	# COs	# LOs	# IOs	CoRate	IdRate
VID	39	19	508	0.07	0.96
IRV	66	58	2285	0.03	0.975
LVC	100	21	2068	0.05	0.99
ALL	207	105	5213	0.04	0.98

IdRate in other languages

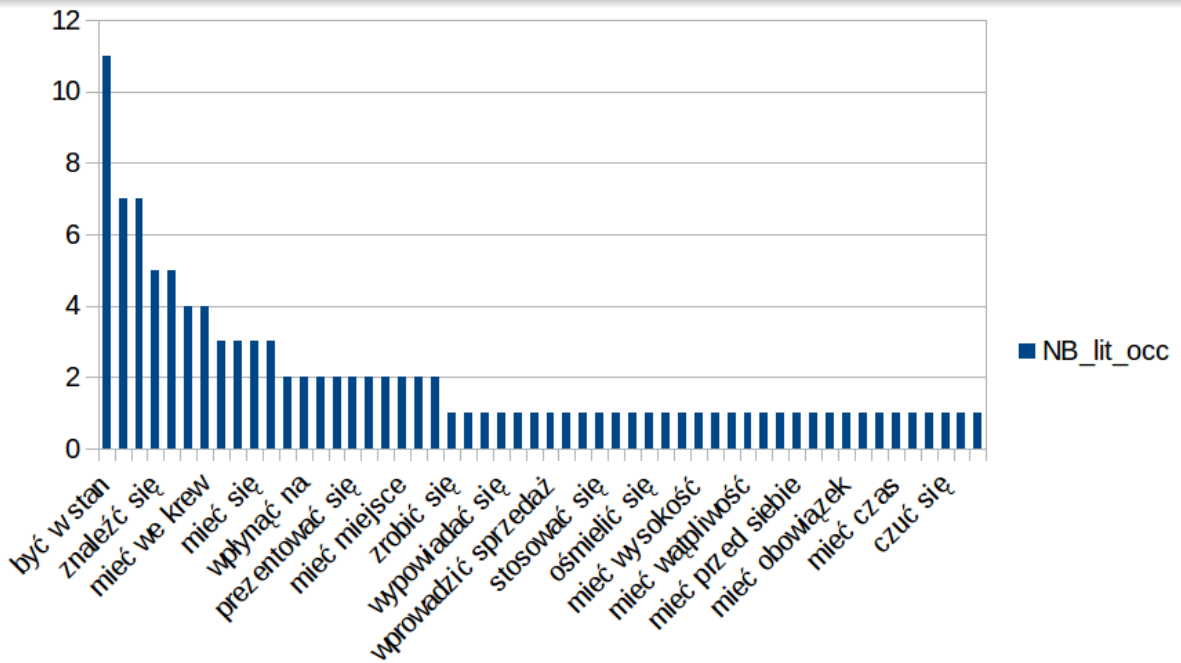
Language	# COs	# LOs	# IOs	CoRate	IdRate
Basque	1082	131	4276	0.2	0.97
German	24	79	4073	0.005	0.98
Greek	126	51	2613	0.05	0.98
Portuguese	668	258	5758	0.1	0.96

Distribution of LOs across types

Zipfian distribution

The 105 literal readings concern 54 VMWEs (types) in total (out of 1703, i.e. 3%).

VMWE	# occ.
być w stanie 'be in state'⇒'be able to'	11
mieścić się 'fit oneself'⇒'be located',	7
dzielić się 'share oneself'⇒'share'	7
18 VMWEs	2-5
33 MWEs	1



Conditions under which literal occurrences take place

VIDs - cross-language conditions

- The VID is figurative (the literal meaning is easy to imagine)

  Dawno juz powinien był ***wyciągnąć nogi*** 'He should have stretched legs long ago.'⇒'He should have died long ago.'
  Położyłem się na trawie i wyciągnąłem nogi. 'I lay down on the grass and stretched my legs.'

- Violated morphological or syntactic constraints lead to disambiguation.

  Papież ***wyniósł****by* ***na ołtarze*** *Jana Pawła II.* 'The pope would bring John Paul the 2nd out on altars.'⇒'The pope would canonize John Paul the 2nd.'
  Za chwilę kardynałowie ***wyniosą*** *obraz* ***na ołtarz*** *przed kościołem.* 'In a while the cardinals will bring the painting out on the altar in front on the church.'
  Nie ***będziemy w stanie*** *nawiązać z nim kontaktu* 'We will not be in the state to make contact with him.'⇒'We will not be able to make contact with him.'
  Komendant był w stanie nietrzeźwości. 'The commandor was is the state of nietrzeźwość'

- The LO is a frequent collocation

  Służenie nam ***mają we krwi*** 'They have serving us in blood.'⇒'Serving us is their innate ability.'
  Miał we krwi ponad 1,5 promila alkoholu. 'He had over 1.5 promil alcohol in his blood.'

Conditions under which literal occurrences take place

LVCs

- Cross-language conditions
  - A predicative noun has a **non-predicative homograph**

    *Rabunki **miały miejsce** na peryferiach stolicy* 'have place'⇒'take place'
    Łódź miała miała miejsce postoju na przystani. 'The boat had its parking place in the dock.'
    *Drabina **miała** 2,5 metra **wysokości*** 'the ladder had 2.5 meters of height'
    Przecież mamy Jego Wysokość Króla IV RP 'But we have His Height King of the 4th Polish Republic.'
- Language-specific conditions (PL)
  - **negation of the copula** *być* 'to be' is expressed by the light verb *nie ma* 'not has'⇒'there is no'

    *Imigranci **mają powody** do niepokoju.* 'Immigrants have reasons to worry.'
    Nie ma powodów do niepokoju. 'Not has reasons to worry'⇒'.'There are no reasons to worry.

Conditions under which literal occurrences take place

IRVs - cross-language conditions

- The verb has a clearly different meaning in the LO

  Mieści się *tu rektorat universytetu.* 'The university rectorate fits ilself here.'⇒'The university rectorate is located here.'
  Urządzenie mieści się w dłoni. 'The device fits itself in a palm.'⇒'the device fit in a palm.'

- True **reflexive** or **reciprocal** uses of the reflexive clitic

  Dzielili się *święconym jajkiem.* 'They shared themselves with a święcone egg.'⇒'They shared as święcone egg.'
  Embrion dzieli się na cztery części. 'The embrio divides itself into four parts.'

- **Impersonal** or **middle passive** alternations

  Mam się *dobrze.* 'I have myself well.'⇒'I'm fine.'

  To się ma, co los przyniesie. 'This has oneself what the fate brings.'⇒'One has what the fate brings.'

- Violated **morphological or syntactic constraints** lead to disambiguation.

  Polityk ***dopuszczał się*** *bezprawia.* 'The politician allows oneself.ACC outlaw acts.GEN.'⇒'The politician dopuszczać się bezprawia.'
  Dopuszcza się taką formę sprzedaży. 'Allows oneself.ACC such form.ACC of sale.'⇒'Such form of sale is allowed'

Reasons for literal occurrences to take place

Hypotheses

- Speakers generally tend to **avoid ambiguity** between literal and idiomatic readings (unless this ambiguity is intended, e.g. in word plays).
- Literal occurrences of VMWEs do occur when:
  - The LO is **hard to rephrase**, e.g. if the VMWE components are functions words (*się* 'oneself', *up*), or the LO is a strong collocation (*mieć 0.2 promila alkoholu we krwi*).
  - The VMWE imposes **morpho-syntactic constraints** which the LO violates. This leads to disambiguation.
  - Other contextual features are strongly disambiguating (topic).

Conclusions

- When a VMWE can occur, it does occur (LO is a rare phenomenon)
- LOs have a Zipfian distribution
- Distinguishing LOs from IOs is not a major challenge, most of it can be handled by methods focused on a few frequent cases.
- The knowledge of morphosyntactic constraints imposed by a VMWE help solve many ambiguities.
- The rate of COs is varies greatly from language to language and is high in Basque (20%) and Portuguese (10%)
- Distinguishing COs from IOs is a major challenge in these languages, even if syntactic dependencies in a VMWE are known.
- The heuristics are efficient in checking corpus annotation consistency.

Future work

- Merging aspectual variants for better identification and more accurate LO definition: *da się* vs. *daje się*.
- Defining a minimal format of a MWE lexicon for efficient MWE identification.

Bibliography I

Cacciari, C. and Corradini, P. (2015).
Literal analysis and idiom retrieval in ambiguous idioms processing: A reading-time study.
Journal of Cognitive Psychology, **27**(7), 797–811.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017).
Multiword expression processing: A survey.
Computational Linguistics, **to appear**.

Pausé, M.-S. (2017).
Structure lexico-sentaxique des locutions du français et incidence sur leur combinatoire.
Ph.D. thesis, Université de Lorraine, Nancy, France.

Peng, J. and Feldman, A. (2016).
Automatic idiom recognition with word embeddings.
In *SIMBig (Revised Selected Papers)*, pp. 17–29. Springer.

Peng, J., Feldman, A., and Vylomova, E. (2014).
Classifying idiomatic and literal expressions using topic models and intensity of emotions.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Bibliography II

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018).
Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions.
In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pp. 222–240. Association for Computational Linguistics.

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., vomír Čéplö, S., Cordeiro, S. R., Eryiğit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., bes kind, C. L., Monti, J., Escartín, C. P., van der Plas, L., QasemiZadeh, B., Ramisch, C., derico Sangati, F., Stoyanova, I., and Vincze, V. (2018).
PARSEME multilingual corpus of verbal multiword expressions.
In S. Markantonatou, C. Ramisch, A. Savary, and V. Vincze, eds., *Multiword expressions at length and in depth. Extended papers from the MWE 2017 workshop*, pp. 87–147. Language Science Press, Berlin.

Sheinfux, L. H., Greshler, T. A., Melnik, N., and Wintner, S. (2017).
Representation and Parsing of Multiword Expressions, chapter Verbal MWEs: Idiomaticity and flexibility, pp. 5–38.
Language Science Press, Berlin.