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

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

in *Prague Bulletin of Mathematical Linguistics*

EUROPHRAS, 9 Sept 2021

Outline

1 Intro
2 SOA
3 Definitions
4 Data
5 Results
6 Interesting cases
7 Wrap-up

Multiword expressions are. . .

Definition

Combinations of words which exhibits lexical, morphosyntactic or semantic idiosyncrasies

Properties

- Discontinuous → *Carlos* ***made*** *an unusual* ***presentation***
- Non compositional → a ***hot dog*** is not a *dog*
- Ambiguous → a ***piece of cake*** is something easy or something to eat
- . . .

Multiword expressions are...

Definition

Combinations of words which exhibits lexical, morphosyntactic or semantic idiosyncrasies

Properties

- Discontinuous → *Carlos* ***made*** *an unusual* ***presentation***
- Non compositional → a ***hot dog*** is not a *dog*
- Ambiguous → a ***piece of cake*** is something easy or something to eat
- ...

This presentation is about *MWE ambiguity*

MWEs can have literal occurrences

- *The boss was **pulling** the **strings** from prison.*
  'The boss was making use of his influence while in prison.'
- *You control the marionette by pulling the strings.*

Outline

1 Intro
2 SOA
3 Definitions
4 Data
5 Results
6 Interesting cases
7 Wrap-up

Literal occurrences of MWEs – state of the art

Philosophy of language: **Computing interpretations** in human mind

- Grice (1989): two-stage interpretation of an utterance by the hearer:
  - computing the **literal** interpretation
  - infering the **true** interpretation (from literal one + the conversational principle)
- Recanati (1995): literal interpretations are not necessarily computed before the non-literal ones
  - **accessibility-based serial model**: only the most accessible interpretation is processed,
  - **parallel model**: several sufficiently accessible interpretations are processed in parallel

Psycholinguistics: **Readers prefer idiomatic readings**

- Cacciari and Corradini (2015): discovering how idioms are **stored and processed in human mind**
- Popiel and McRae (1988): 30 English MWEs rated; figurative interpretations – **more frequent and more familiar** to speakers than literal interpretations
- Geeraert et al. (2018): rating and eye-tracking: figurative interpretations – **more acceptable** than literal ones, althought their **fixation** duration is **longer**

Literal occurrences of MWEs – state of the art

Linguistics: Modelling **links between literal & idiomatic readings**

- Abeillé and Schabes (1989); Lichte and Kallmeyer (2016): encoding literal vs. idiomatic readings, with the same syntax, in **formal grammars** (LTAG)
- Sheinfux et al. (2019), Pausé (2017): links between literal and idiomatic readings indicate **which morphosyntactic variations are allowed** or prohibited

Datasets

- Literal and idiomatic readings (binary scale):
  - Tu and Roth (2011, 2012): English LVCs and VPCs with 6 common verbs
  - Cook et al. (2008): English verb-noun combinations
  - Hashimoto and Kawahara (2008): Japanese idioms and their literal readings
  - Savary and Cordeiro (2018): Polish verbal MWEs and their literal readings
  - Ehren et al. (2020): German vebal idioms and their literal readings
- Degree of compositionality (multi-valued scale):
  - Bott et al. (2016): German VPCs
  - Ramisch et al. (2016): English, French and Portuguese Noun-Noun and Adjective-Noun compounds
  - Cordeiro et al. (2019): A review of such datasets

Literal occurrences of MWEs – state of the art

NLP: Literal-idiomatic distinction considered a major challenge (Constant et al., 2017), and modelled in terms of

- Hashimoto and Kawahara (2008): supervised **WSD**
- Fazly et al. (2009): statistical measures of lexical and syntactic **flexibility**
- Peng et al. (2014): **topic representation** for occurrence paragraphs
- Peng and Feldman (2016); Köper and Schulte im Walde (2016): vector distance in **distributional semantics**
- Ehren et al. (2020): **NN**-based classification

NLP: quantifying literal readings

- Waszczuk et al. (2016): **idiomaticity rate** of Polish MWEs is **0.95**; this can be leveraged for **faster parsing**
- Savary and Cordeiro (2018): **idiomaticity rate** of Polish **verbal** MWEs is **0.98**

Three research questions

- How to **define** literal occurrences of MWEs?
- How **frequent** are literal occurrences of MWEs?
- What are the **cross-lingual** characteristics of literal occurrences?
  - Study them in Basque, German, Greek, Polish and Portuguese

Context

Focus on **verbal multiword expressions** (VMWEs) in the **PARSEME corpora** using **Universal Dependencies** as the underlying syntactic formalism.

Outline

1. Intro
2. SOA
3. Definitions
4. Data
5. Results
6. Interesting cases
7. Wrap-up

PARSEME definitions (Savary et al., 2018)

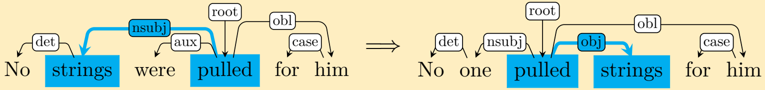
Verbal multiword expression (VMWE)

MWE occurrence = sequence of syntactic words $e = e_1, \ldots, e_n$ $(n > 1)$ where:

- All components $e_1, \ldots, e_n$ of e are **lexicalized** (replacing them by semantically related words results in a meaning shift which goes beyond what is expected from the replacement)
- The **canonical form** of e builds a (weakly) **connected graph** whose **head** is a **verb**.
- e is **semantically non-compositional**, i.e. passes the tests from the **PARSEME guidelines**

Canonical form

Least syntactically marked syntactic variant which preserves the idiomatic reading (active voice is less marked then passive, etc.)



VMWE typology (validated in 25 languages)

Universal categories (all languages)

- verbal idioms (**VID**)
  - (EN) *to* ***call it a day***
- light-verb constructions (**LVCs**)
  - (EN) *to* ***give*** *a* ***lecture*** (LVC.full)
  - (EN) *to* ***grant rights*** (LVC.cause)

Quasi-universal categories (many languages)

- inherently reflexive verbs (**IRVs**)
  - (EN) *to* ***help oneself*** 'to take something freely'
- verb-particle constructions (**VPCs**)
  - (EN) *to* ***do in*** 'to kill' (VPC.full)
  - (EN) *to* ***eat up*** (VPC.semi)
- multi-verb constructions (**MVCs**)
  - (HI) ***kar le-na*** 'do take.INF' 'to do something (for one's own benefit)'

What is a literal occurrence?

- *The boss was **pulling** the **strings** from prison.*
- *You control the marionette by pulling the strings.*
- *He strings paper lanterns on trees without pulling the table.*
- *As an effect of pulling, the strings broke.*
- *Determine the maximum force you can pull on the string so that the string does not break.*
- *The strings pulled the bridge.*
- *My husband says no **strings** were **pulled** for him.*
- *She moved Bill by **pulling** wires and **strings**.*
- *The article addresses the political **strings** which the journalist claimed that the senator **pulled**.*
- *He was there, **pulling** the **strings**, literally and metaphorically.*

Literal occurrence I

For a given MWE $e = e_1, \ldots, e_n$,

nsubj

No	**strings**	were	**pulled**	for	him
no	string	is	pull	for	he
DET	NOUN	AUX	VERB	ADP	PRON

$\Longrightarrow$

obj

No	one	**pulled**	**strings**	for	him
no	one	pull	string	for	he
DET	PRON	VERB	NOUN	ADP	PRON

a *literal occurrence* (LO) of e is a **co-occurrence** e' of words $e'_1, \ldots, e'_n$ in which:

(a) $e'_1, \ldots, e'_n$ have the **same lemmas and POS** as $e_1, \ldots, e_n$.

He	**strings**	paper	lanterns	on	trees	without	**pulling**	the	table
he	string	paper	lanterns	on	tree	without	pull	the	table
PRON	VERB	NOUN	NOUN	ADP	NOUN	ADP	VERB	DET	NOUN

(b) The syntactic **dependencies** between $e'_1, \ldots, e'_n$ are the **same or equivalent** to those between $e_1, \ldots, e_n$ in e's **canonical form**.

obl

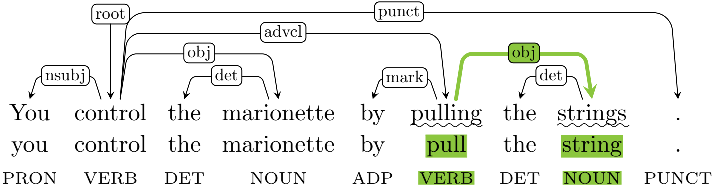
You	can	pull	on	the	strings
you	can	pull	on	the	string
PRON	AUX	VERB	ADP	DET	NOUN

nsubj

The	strings	pulled	the	bridge	.
the	string	pull	the	bridge	.
DET	NOUN	VERB	DET	NOUN	PUNCT

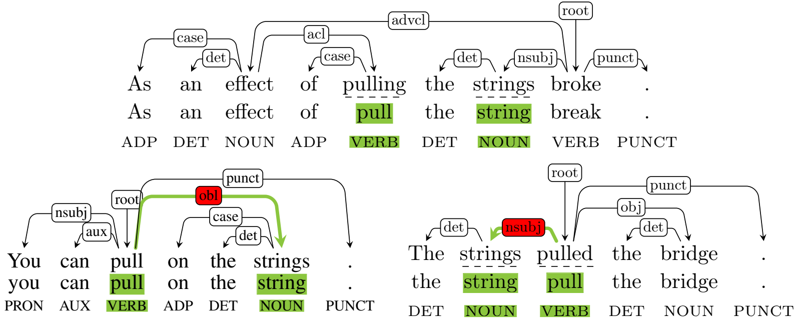
(c) e' is **not** an **idiomatic** occurrence of a MWE

Literal occurrence II



Coincidental occurrence

Condition (b) is not fullfilled.



Outline

1 Intro

2 SOA

3 Definitions

4 Data

5 Results

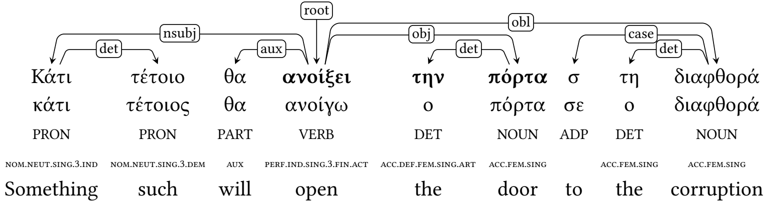
6 Interesting cases

7 Wrap-up

Corpus

- PARSEME shared task v1.1 corpora
- Manual annotation for VMWEs, no annotation of their literal occurrences
- Manual or automatic annotation of lemmas, POS tags, morphological features, dependency trees (following the UD standard)

Lang.	**Sent.**	**Tokens**	**VMWEs**	**Morphology**	**Syntax**
Basque	11,158	157,807	3,823	partly manual	partly manual
German	8,996	173,293	3,823	automatic	automatic
Greek	8,250	224,762	2,405	automatic	automatic
Polish	16,121	274,318	5,152	partly manual	partly manual
Portuguese	27,904	638,002	5,536	partly manual	partly manual



Identification of potential literal occurrences

Heuristics

For each annotated VMWE $e = e_1, \ldots, e_n$,

1. Extract all co-occurrences of n tokens with the **same lemma or surface forms** which are **not annotated** as VMWEs
2. Apply 4 filters in parallel (high recall, reasonable precision):
   - all tokens fit into a window with **2 gaps**

     nsubj
     No **strings** were **pulled** for him
     no string is pull for he
     $\Longrightarrow$
     obj conj cc
     pulling wires and **strings**
     pull wire and string

   - all tokens form a **connected graph** as e but dependency **labels and directions** are **neglected**

     nsubj
     No **strings** were **pulled** for him
     no string is pull for he
     $\Longrightarrow$
     acl
     strings which you have been **pulling**
     string which you have be pull

   - as above but **dependency directions** are preserved

     nsubj
     No **strings** were **pulled** for him
     no string is pull for he
     $\Longrightarrow$
     obl
     Pull as strongly as you can on the **string**
     pull as strongly as you can on the string

   - as above but **dependency labels** are preserved

     nsubj
     No **strings** were **pulled** for him
     no string is pull for he
     $\Longrightarrow$
     nsubj
     The **strings** **pull** the bridge
     the string pull the bridge

3. Merge the outcome of the 4 filters
4. NB: POS, dependencies and even VMWEs may be **noisy**

Manual annotation

- err-false-idiomatic:
  - *She [. . .] brought back a branch of dill.*
- err-skipped-idiomatic:
  - *Any insult [. . .]* ***brings*** *us all* ***down***
- nonverbal-idiomatic:
  - *After the major* ***kill-offs****, wolves [. . .]*
- wrong-lexemes:
  - *Then take your finger and place it under their belly*
- missing-context:
  - *Enron is blowing up.*
- coincidental: (***do the job***)
  - *[. . .] why you like the job and do a little bit of [. . .]*
- literal: (***get going***)
  - *At least you get to go to Florida [. . .]*

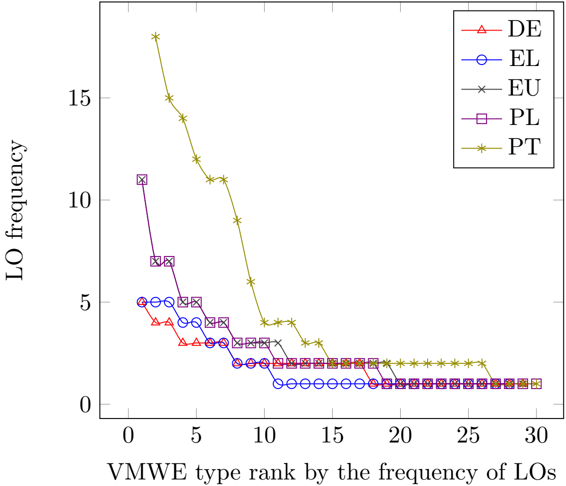
Outline

1. Intro
2. SOA
3. Definitions
4. Data
5. Results
6. Interesting cases
7. Wrap-up

Overall results

	DE		EL		EU		PL		PT	
Annotated VMWEs	3,823		2,405		3,823		4,843		5,536	
LO candidates	926		451		2,618		332		1,997	
err-false-id.	21.5%	(199)	12.0%	(54)	9.4%	(246)	0.0%	(0)	3.8%	(76)
err-skipped-id.	27.0%	(250)	47.5%	(214)	17.3%	(453)	5.4%	(18)	10.7%	(213)
nonverbal-id.	0.0%	(0)	0.0%	(0)	0.2%	(6)	0.0%	(0)	0.5%	(9)
wrong-lexemes	40.1%	(371)	0.9%	(4)	26.7%	(700)	1.8%	(6)	38.1%	(760)
missing-context	0.3%	(3)	0.2%	(1)	0.5%	(12)	2.1%	(7)	0.7%	(13)
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)
Idiomaticity rate	**98%**		**98%**		**98%**		**98%**		**96%**	

Distribution of LOs



Outline

1 Intro

2 SOA

3 Definitions

4 Data

5 Results

6 Interesting cases

7 Wrap-up

Literal occurrences of IRVs

(1) 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.'

(2) 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.'

Finding: Some IRVs are ambiguous with true **reflexive**, **reciprocal**, **impersonal** and **middle-passive** uses of the reflexive clitic.

Literal occurrences of LVCs

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

(4) Kierowcy mieli sfałszowane zezwolenia. (PL)
Drivers had falsified permissions.
'The drivers had false driving licenses.'

Finding: LOs of LVCs occur when the **predicative noun** has a **non-predicative couterpart**

Polish-specific literal occurrences of LVCs

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

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

Finding: An negated LVC is homonimic with a **negated existential construction** (*są powody* (lit. 'are reasons.NOM') 'there are reasons' ⇒ *nie ma powodów* (lit. 'not has reasons.ACC') 'there are no reasons')

Portuguese-specific literal occurrences of LVCs

(7) a **força** que o Alvinegro **tem** (PT)
the strength that the Alvinegro has
'the strength of Alvinegro'

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

Finding: **Resultative constructions** are homonimic (and hard to distinguish from) LVCs.

Literal occurrences of VIDs I

(9) 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.'

(10) 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.'

Finding: Figurative idioms have a strong **potential** of literal occurrences but are **rare**.

Literal occurrences of VIDs II

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

(12) 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.'

Finding: Domain-specific uses can be literal occurrences of general-purpose idiomatic occurrences.

Basque-specific literal occurrences of VIDs

(13) **Kontu-a-n** **hartu** du lagun-a-ren iritzi-a. (EU)

Kontu-a-n	**hartu**	du	lagun-a-ren	iritzi-a.
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.'

(14) Diru-a hartu du kontu-tik. (EU)

Diru-a	hartu	du	kontu-tik.
money-ART.ABS	take	AUX	account-ABL

Took money from the account.
'He/She withdraw money from the account.'

Finding: In agglutinative languages post-positions vanish in the process of lemmatization.
lemma-based definition of a literal occurrence ⇒ high number of literal occurrences (with different post-positions)

Greek-specific literal occurrences of VIDs

(15) 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 angry at the kids.'

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

Finding: The way of lemmatizing pronouns (e.g. *τα* 'them' ⇒ *εγώ* 'I', *την* 'her' ⇒ *εγώ* 'I') influences the proliferation of literal occurrences for VIDs with expletives.

Properties of coincidental occurrences and frequent errors

- They influence **automatic processing** of VMWEs (especially **identification**)
- See the **paper** ...

Outline

1. Intro
2. SOA
3. Definitions
4. Data
5. Results
6. Interesting cases
7. Wrap-up

Take-home message

- We provide a **formal definition** of a literal occurrence (missing so far)
- Literal occurrences should be understood in semantic and **syntactic** terms (treebanks needed)
- Literal occurrences are **rare** in practice (2-4%)
- **Coincidental** occurrences are more frequent but can be eliminated with **good parsers**
- **Simple heuristics** perform well on multilingual identification of seen VMWEs (Pasquer et al., 2020).
- Language-specific special treatment of a **few frequent cases** can boost identification performances

Take-home message

- We provide a **formal definition** of a literal occurrence (missing so far)
- Literal occurrences should be understood in semantic and **syntactic** terms (treebanks needed)
- Literal occurrences are **rare** in practice (2-4%)
- **Coincidental** occurrences are more frequent but can be eliminated with **good parsers**
- **Simple heuristics** perform well on multilingual identification of seen VMWEs (Pasquer et al., 2020).
- Language-specific special treatment of a **few frequent cases** can boost identification performances

Literal occurrences of VMWEs are **rare birds** that **cause a stir**

Research question

Why are literal occurrences of VMWEs **rare**, although **ambiguity** is a **norm** in natural language?

Outline

8 Bibliography

9 Heuristics

10 Guidelines

Bibliography I

Abeillé, A. and Schabes, Y. (1989). Parsing idioms in lexicalized tags. In H. L. Somers and M. M. Wood, eds., Proceedings of the 4th Conference of the European Chapter of the ACL, EACL'89, Manchester, pp. 1–9. The Association for Computer Linguistics.

Bott, S., Khvtisavrishvili, N., Kisselew, M., and Schulte im Walde, S. (2016). $G_h$ost-PV: A Representative Gold Standard of German Particle Verbs. In Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon, Osaka, Japan.

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**.

Cook, P., Fazly, A., and Stevenson, S. (2008). The vnc-tokens dataset. In Proceedings of the Workshop on Multiword Expressions.

Cordeiro, S., Villavicencio, A., Idiart, M., and Ramisch, C. (2019). Unsupervised compositionality prediction of nominal compounds. Computational Linguistics. (to appear).

Ehren, R., Lichte, T., Kallmeyer, L., and Waszczuk, J. (2020). Supervised disambiguation of German verbal idioms with a BiLSTM architecture. In Proceedings of the Second Workshop on Figurative Language Processing, pp. 211–220, Online. Association for Computational Linguistics.

Fazly, A., Cook, P., and Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. Computational Linguistics, **35**(1), 61–103.

Geeraert, K., Baayen, R. H., and Newman, J. (2018). "Spilling the bag" on idiomatic variation. 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. 1–33. Language Science Press., Berlin.

Bibliography II

Grice, H. P. (1989). Studies in the Way of Words. Harvard University Press, Cambridge, Mass.

Hashimoto, C. and Kawahara, D. (2008). 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, pp. 992–1001. Association for Computational Linguistics.

Köper, M. and Schulte im Walde, S. (2016). 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, pp. 353–362, San Diego, California.

Lichte, T. and Kallmeyer, L. (2016). Same syntax, different semantics: A compositional approach to idiomaticity in multi-word expressions. In C. Piñón, ed., Empirical Issues in Syntax and Semantics 11, pp. 111–140.

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In Proceedings of the 28th International Conference on Computational Linguistics, pp. 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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 III

Popiel, S. J. and McRae, K. (1988). The figurative and literal senses of idioms, or all idioms are not used equally. Journal of Psycholinguistic Research, **17**(6), 475–487.

Ramisch, C., Cordeiro, S., Zilio, L., Idiart, M., Villavicencio, A., and Wilkens, R. (2016). 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), pp. 156–161, Berlin, Germany. ACL. CORE2018 rank: A*. https://aclweb.org/anthology/P16-2026.

Recanati, F. (1995). The alleged priority of literal interpretation. Cognitive Science, **19**, 207–232.

Savary, A. and Cordeiro, S. (2018). 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, Pr pp. 64 – 72, Prague, Czech Republic.

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. (2019). Verbal MWEs: Idiomaticity and flexibility. In Y. Parmentier and J. Waszczuk, eds., Representation and Parsing of Multiword Expressions, pp. 5–38. Language Science Press, Berlin.

Tu, Y. and Roth, D. (2011). Learning English light verb constructions: Contextual or statistical. In Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World, pp. 31–39. Association for Computational Linguistics.

Bibliography IV

Tu, Y. and Roth, D. (2012). Sorting out the most confusing English phrasal verbs. In Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings o pp. 65–69. Association for Computational Linguistics.

Waszczuk, J., Savary, A., and Parmentier, Y. (2016). Promoting multiword expressions in A* TAG parsing. In COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, pp. 429–439.

Outline

8 Bibliography

9 Heuristics

10 Guidelines

Relaxed non idiomatic occurrences (RNOs)

Goal: extract potential LOs from the corpus for annotation

Procedure

- extract each VMWE token $e = \{e_1, \ldots, e_{|e|}\}$ in each sentence s
- for each extracted e, for each sentence $s' = \{s'_1, s'_2, \ldots, s'_{|s'|}\}$:
- r is a *relaxed non-idiomatic occurrence* (RNO) of e in s', if:
  - r is a subsequence of s'
  - $|r| = |e|$
  - there is a bijection $\text{rno}_e^r : \{1, \ldots, |e|\} \to \{1, \ldots, |e|\}$ such that:
    - for $i \in \{1, 2, \ldots, |e|\}$ and $j = \text{rno}_e^r(i)$, $cf(\text{lemmasurface}(e_i)) \in \{cf(\text{lemma}(r_j)), cf(\text{surface}(r_j))\}$
    - r is not a VMWE token

LO candidates

- **WindowGap**: all matched tokens of the RNO must fit into a sliding window with no more than g external elements (gaps). We use $g = 2$.
- **BagOfDeps**: the RNO must corresponding to a weakly connected unlabeled subgraph with no dummy nodes
- **Unlabeled**: the RNO must correspond to a connected unlabeled graph with no dummy nodes, that is, the dependency labels are ignored but the parent relations are preserved.
- **Labeled**: the RNO must be a connected labeled graph with no dummy nodes, in which both the parent relations and the dependency labels are preserved.

The resulting set of LO candidates is the **union of the 4 heuristics** output

Outline

8 Bibliography

9 Heuristics

10 Guidelines

First phase: initial checks

- $e = \{e_1, e_2, \ldots, e_{|e|}\}$ is a VMWE token annotated in a sentence s
- cs is the canonical structure of e's type
- $c = \{c_1, c_2, \ldots, c_{|c|}\}$ is an LO candidate extracted by the heuristics
- **[FALSE]** Should e have been annotated as an IO of an MWE at all?
  - NO $\rightarrow$ annotate c as err-false-idiomatic
  - YES $\rightarrow$ go to the next test
- **[SKIP]** Is c an IO of an MWE that annotators forgot/ignored?
  - YES, it is a verbal MWE $\rightarrow$ annotate c as err-skipped-idiomatic
  - YES, but a non-verbal MWE $\rightarrow$ annotate c as nonverbal-idiomatic
  - UNSURE, not enough context $\rightarrow$ annotate c as missing-context
  - NO $\rightarrow$ go to the next test
- **[LEX]** Do c's components have the same lemma and POS as cs's?
  - NO $\rightarrow$ annotate c as wrong-lexemes
  - YES $\rightarrow$ go to the next test

Performance of the heuristics

Language	WindowGap			BagOfDeps			Unlabeled			Labeled			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