Multiword expressions the Achilles' heel of natural language processing

Agata Savary

University of Tours, Blois, France

LCL Pavia, 10 May 2019

Multiword expressions (MWEs)

What is so special about the highlighted expressions?

*The **prime time** speech by **first lady** **Michelle Obama** **set** the house **on fire**. She made **crystal clear** which issues she **took to heart** but she was **preaching to the choir**.*

Multiword expressions (MWEs)

What is so special about the highlighted expressions?

*The **prime time** speech by **first lady** **Michelle Obama** **set** the house **on fire**. She made **crystal clear** which issues she **took to heart** but she was **preaching to the choir**.*

Definition [2]

Combination of at least **two words** which exhibits lexical, morphological, syntactic, semantic and /or statistical **idiosyncrasies**.

Sample idiosyncrasies in MWEs

- **Non-compositional semantics**: the meaning of a MWE is surprising, given the meanings of its component words
  - EN *to **pull** one's **leg*** 'to tease someone playfully'
  - IT ***lasciar perdere*** 'to let lose'⇒'to give up'
- Morphosyntactic **irregularity** (token[a]-specific):
  - FR ***grand-mères*** '$\text{grand}_{sing.masc}$-$\text{mothers}_{pl.fem}$' (defective agreement)
  - EN ***by and large*** 'mostly' (Prep Conj Adj is an irregular syntactic structure)
  - EN *to **go nuts*** 'to get crazy' (*go* alone is intransitive)
- Morphosyntactic **inflexibility** (type[b]-specific):
  - EN ***the die is cast*** 'a point of no-retreat has been passed' vs. *#someone cast the die*

---

[a] Token = individual occurrence

[b] Type = sets of surface realizations of the same expression

Defining idiosyncrasy

One usually tries to distinguish MWEs from "regular" or "free" constructions of the **same syntactic structure**.

Synt. structure	**Regular construction**	**MWE**
Adj N	*a hot soup*	*a **hot dog*** 'a hot sausage served in a long bread roll'
V Det N	*to pay a bill*, *to discuss a visit*	*to **pay** a **visit*** 'to visit'
V NP Prep Det N	*to throw fish to the dolphins*	*to **throw** Harry **to the lions*** 'to sacrifice or ruin Harry'
V Part NP	*to put up a flag*	*to **put up** a great performance* 'to show a great level of skill'
V Refl PP	*to wash oneself in the bath*	*to **find oneself** in times of trouble* 'to discover that one is in trouble'

Semantic non-compositionality

Semantic compositionality [6]

An expression E is semantically compositional if a **compositional semantic calculus** applies to it: given the meanings of E's components and E's **syntactic structure**, a grammar rule allows us to deduce the meaning of E.

Semantic non-compositionality – 3 cases

- A component has no individual meaning, it functions only within MWEs (*cranberry/fossil word*)
  - *to* ***go astray*** 'to become lost'
  - *to* ***let bygones be bygones*** 'to ignore a past offense'
- The syntactic structure is irregular
  - ***by and large*** 'mostly'
  - ***long live*** *the queen!* 'may she live for a long time'
  - *to* ***pretty-print*** 'use beautifying conventions for texts printing'
- The meaning is not deduced regularly
  - *a* ***hot dog*** 'a hot sausage served in a long bread roll' or 'a person showing off dangerous acts'
  - *to* ***pay*** *a* ***visit*** 'to visit'
  - *the* ***Black Sea*** 'a lake in Asia'

Inflexibility of MWEs

A MWE is (much) **less flexible** (variable) than a regular construction of the same syntactic structure.

Regular construction	**MWE**	**MWE property**
warm soup $\approx$[1] *hot soup* $\approx$ *warm stew*	***hot dog*** vs. *#warm dog* vs. *#hot terrier*	Lexical inflexibility
to throw meat to the lions $\approx$ *to throw meat to the lion*	*to* ***throw*** *someone* ***to the lions*** vs. *#to throw someone to the lion*	Morphological inflexibility
she held her elbow $\approx$ *she held his elbow*	*she* ***held*** *her* ***tongue*** 'she refrained from expressing her view' vs. *#she held his tongue*	Morpho-syntactic inflexibility

[1] '$\approx$' means that the meaning shift is predictable from the formal change

Inflexibility of MWEs

Regular construction	MWE	MWE property
to throw meat to the lions ≈ *to throw meat to the hungry lions*	*to **throw** someone **to the lions*** vs. *#to throw someone to the hungry lions*	Syntactic inflexibility
he made it for her ≈ *It was made for her by him*	*he **made it** to the station well in advance* 'he managed to get to the station . . . ' vs. *#it was made by him to the station . . .*	
the die is stolen ≈ *someone stole the die*	***the die is cast*** 'a point of no-retreat has been passed' vs. *#someone cast the die*	
a text in red and blue ≈ *a text in blue and red*	*a photo in **black and white*** 'a photo in shades of gray' vs. *#a photo in white and black*	

Partial (in)flexibility of MWEs

Property	MWE respecting the property	MWE violating the property
free subject	*John **held** his **tongue*** ≈ *Adam **held** his **tongue***	***fear lends wings*** 'fear gives you unusual capacities' vs. *#Panic lends wings*
free object	***a little bird told** Suzy* 'Suzy received the information from a secret source' ≈ ***a little bird told** Mary*	*Suzy **crossed** her **fingers** for Tim* 'Suzy wishes good luck to Tim' vs. *#Suzy crossed her thumbs*
verb inflection	*Suzy **crossed** her **fingers*** ≈ *Suzy will **cross** her **fingers***	***a little bird told** Suzy* ≈ *#a little bird will tell Suzy*
object inflection	*Luke **held** his **tongue*** ≈ *Luke and Sue **held** their **tongues***	*Suzy **crossed** her **fingers*** vs. *Suzy crossed her finger*
object modification	*John **broke** my **fall*** 'John made my fall less forceful' ≈ *John **broke** my sudden **fall***	*Suzy **crossed** her **fingers*** vs. *Suzy **crossed** her long **fingers***
free poss. det.	*John **broke** my **fall*** ≈ *John **broke** his/her/our **fall***	*Suzy **crossed** her **fingers*** vs. *#Suzy crossed our fingers*
passive	*John **broke** my **fall*** ≈ *My **fall** was **broken** by John*	***fear lends wings*** vs. *#wings are lent by fear*

(In)flexibility as a matter of scale

A MWE is **less flexible** than a regular construction of the same syntactic structure but it is often **not totally inflexible**.

Expression	Property						
	Free subject	Free object	Verb inflection	Object inflection	Object modif.	Free poss. det.	Passive
fear lends wings							
Suzy ***held*** *her* ***tongue***	✓		✓	✓			
Suzy ***crossed*** *her* ***fingers***	✓		✓				✓
a little bird told *Suzy*		✓		✓	✓	✓	
Suzy ***broke*** *my* ***fall***	✓		✓	✓	✓	✓	✓
Suzy lends her books	✓	✓	✓	✓	✓	✓	✓
Suzy held her book	✓	✓	✓	✓	✓	✓	✓
Suzy crossed the road	✓	✓	✓	✓	✓	✓	✓
a little girl told Suzy	✓	✓	✓	✓	✓	✓	✓
Suzy broke my car	✓	✓	✓	✓	✓	✓	✓

Lexicalization

MWE components

- **Lexicalized components** – mandatory components, always realized by the same lexemes; without them the MWE cannot occur. They are marked **in bold**.
- **Open slots** – mandatory components which can be realized (relatively) freely
- Example: *she set the house on fire* 'she made the people very excited'
  - *Michelle put the house on fire*, *His wife put the house on fire* → *she* is not lexicalized
  - *#she put the house on fire*[a], *#she set the house in fire*, *#she set the house in blaze* → *set*, *on* and *fire* are lexicalized
  - *she set the assembly/many lobbies on fire* → *the house* is not lexicalized
  - **she put on fire* → the direct object of *put* is an open slot
  - ⟹ *NP* ***set*** *NP* ***on fire***

[a] '#' and '*' signal the loss of idiomatic meaning and ungrammaticality, respectively.

Challenges for NLP

Pervasiveness

Up to 40% of words in a text belong to MWEs. [5, 10]

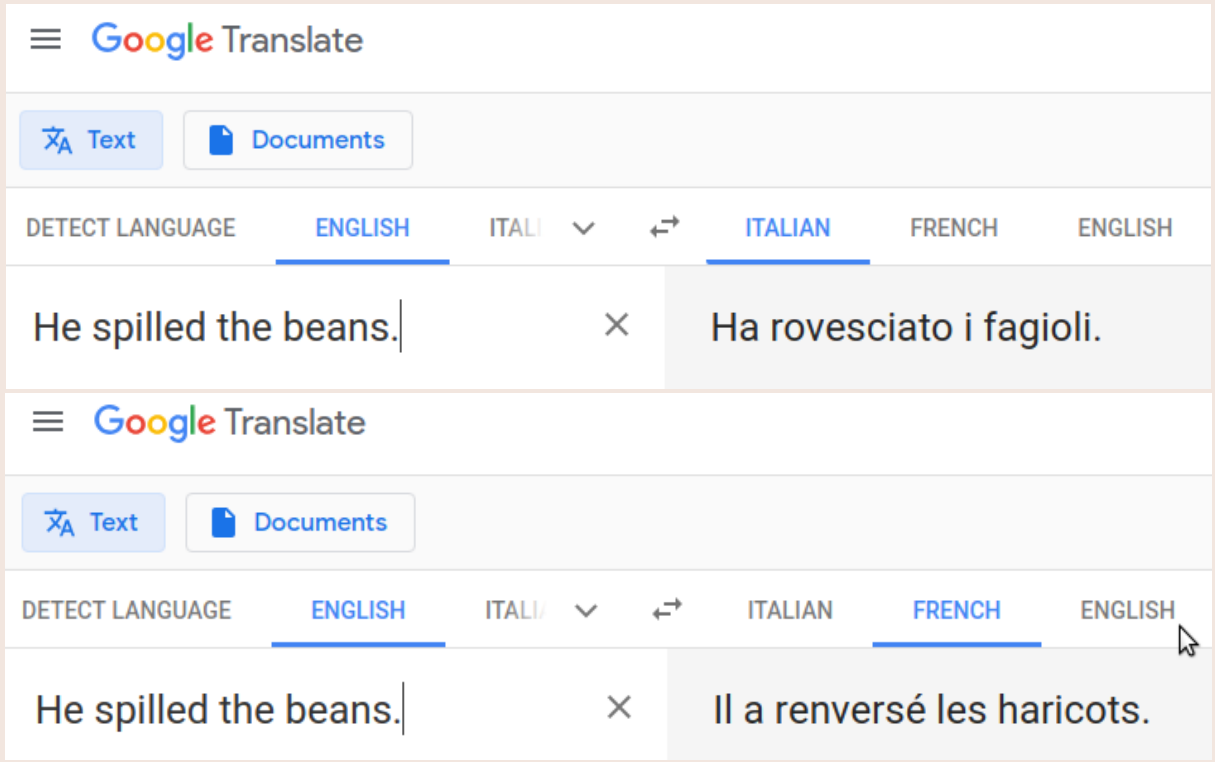
> *The **prime time** speech by **first lady** **Michelle Obama** **set** the house **on fire**. She made **crystal clear** which issues she **took to heart** but she was **preaching to the choir**.*

Here: 18 MWE components for 31 words of the text → 58%

Non-compositionality

Computational methods are mostly **compositional**. Complex phenomena are decomposed into simpler subproblems. Subproblems receive independent solutions, which are then composed to provide global solutions.
MWEs are **semantically non-compositional**. They are challenging for **semantically-oriented NLP applications**.

Machine translation



Word-to-word translations do not capture the idiomatic meaning.

Information retrieval

- The task: for a given query (one or more words), automatically find the relevant documents
- Bag-of-words approach:
  - Eliminate stop words, lemmatize the text, create an **index** (list of words contained in the text with their frequencies)
  - Example: *He **took the bull by the horns*** $\rightarrow$ {bull – 1, horn – 1, take – 1}
  - Each query word is looked up in the index. The documents containing the query words are weighted and returned.
- Challenges from MWEs:
  - A document contains *He **took the bull by the horns*** 'He dealt decisively with a difficult situation'
  - The query contains *horns of a bull*
  - The document is irrelevant but it will likely be returned

Opinion mining (= sentiment analysis)

- The task: automatically predict the valency (positive, neutral ou negative) of an opinion expressed by a text
- Examples:
  - *Huge respect to the French people for believing in better lives.*
  - *Nothing justifies violence or intimidation against an elected representative of the Republic.*
- Simple bag-of-word technique:
  - Single words are annotated with elementary valency: *respect* → 1, *violence* → -2, *justify* → 1, ...
  - Local rules modify elementary valency:
    - *huge*, *extreme* multiply the valency; *huge respect* → 2*1 = 2; *extreme violence* → 2*(-2) = -4
    - negation inverses valency: *nothing justifies* → -1*1=-1

Opinion mining – challenges from MWEs

Text	Comp. valency	True valency
kick$_0$ *the bucket*$_O$ 'die'	0	-2
go nuts$_O$ 'get crazy'		
make a mountain$_0$ *out of a molehill*$_0$ 'exaggerate'		
it's in the bag$_0$ 'success will obviously be achieved'		
kill$_{-2}$ *two birds*$_O$ *with one stone*$_O$ 'solve two problems with one single action'		
the sky's the limit$_{-1}$ 'there is no limit'		
beyond one's wildest$_{*(-1)}$ *dreams*$_1$ 'much better than expected'		
dark$_{-1}$ *horse* 'a person with a surprising ability'		

Opinion mining – challenges from MWEs

Text	Comp. valency	True valency
kick$_0$ *the bucket*$_O$ 'die'	0	-2
go nuts$_O$ 'get crazy'	0	-2
make a mountain$_0$ *out of a molehill*$_0$ 'exaggerate'	0	-1
it's in the bag$_0$ 'success will obviously be achieved'	0	2
kill$_{-2}$ *two birds*$_O$ *with one stone*$_O$ 'solve two problems with one single action'	-2	1
the sky's the limit$_{-1}$ 'there is no limit'	-1	2
beyond one's wildest$_{*(-1)}$ *dreams*$_1$ 'much better than expected'	-1	2
dark$_{-1}$ *horse* 'a person with a surprising ability'	-1	2

Solutions

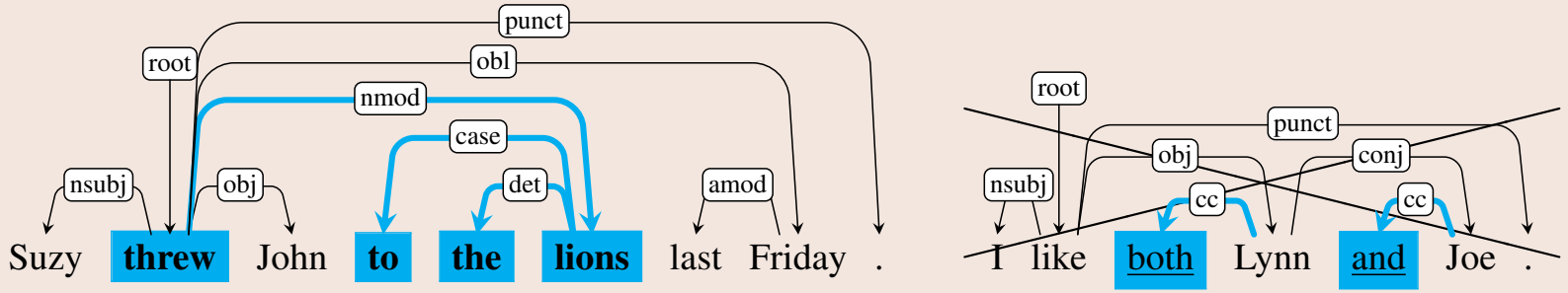
- Automatically identify the MWEs in the text, apply dedicated treatment
- Machine translation
  - rephrase the MWE prior to translation
  - *he **spilled** the **beans*** → *he revealed the secret* → *ha rivelato il segreto*
- Information retrieval
  - don't add the MWE components to the index
  - add the expression as a whole
  - *the re-election was **in the bag*** → {re-election – 1, in the bag – 1}
- Opinion mining
  - assing a valency to the whole expression
  - *[**kill two birds with one stone**]*$_2$

Focus on **verbal** MWEs

Verbal MWEs (VMWEs)

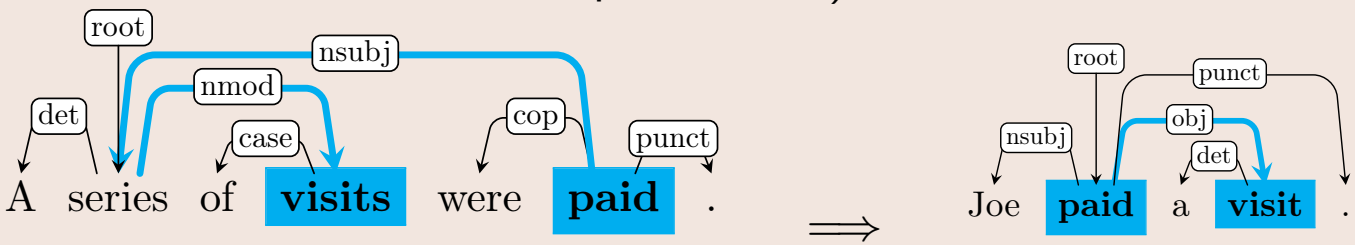
Verbal MWEs – MWEs whose **canonical form** is such that:

- its syntactic head is a verb V
- its other lexicalized components form phrases directly dependent on V, i.e. the **dependency subgraph** of the lexicalized components is weakly **connected**



Canonical form

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



Challenges from **verbal** MWEs

- Discontinuity:
  - EN *Trying hard to **bear** all these more or less important indications **in mind***
  - DE *Klaus Kinkel (FDP) **ging** in seiner Würdigung des Mauerfalls zumindest auf den 9. November 1938 **ein**.*
- Variability: morphological, syntactic, lexical
  - EN *he **broke** my **fall*** vs. *both of my **falls** were hard to **break***
- Ambiguity: idiomatic vs. literal readings
  - EN *she **takes the cake*** 'she is the most outstanding' vs. *she takes the cake*
- Overlaps:
  - EN ***take** a **walk** and then a long **shower*** (coordination)
  - EN ***take** the fact that I **gave up** **into account*** (interleaving)
  - EN ***let** **the cat** **out** **of the bag*** (nesting)
- Multiword tokens
  - ES ***abstener**|**se*** 'abstain oneself'⇒'abstain' vs. *me abstengo*
  - DE ***auf**|**machen*** 'out|make'⇒'open' vs. *macht auf*
- Different languages ⇒ different behavior, linguistic traditions...

VMWE: state of the art in NLP

VMWE modeling via corpus annotation

- PARSEME corpus of verbal MWEs [11]

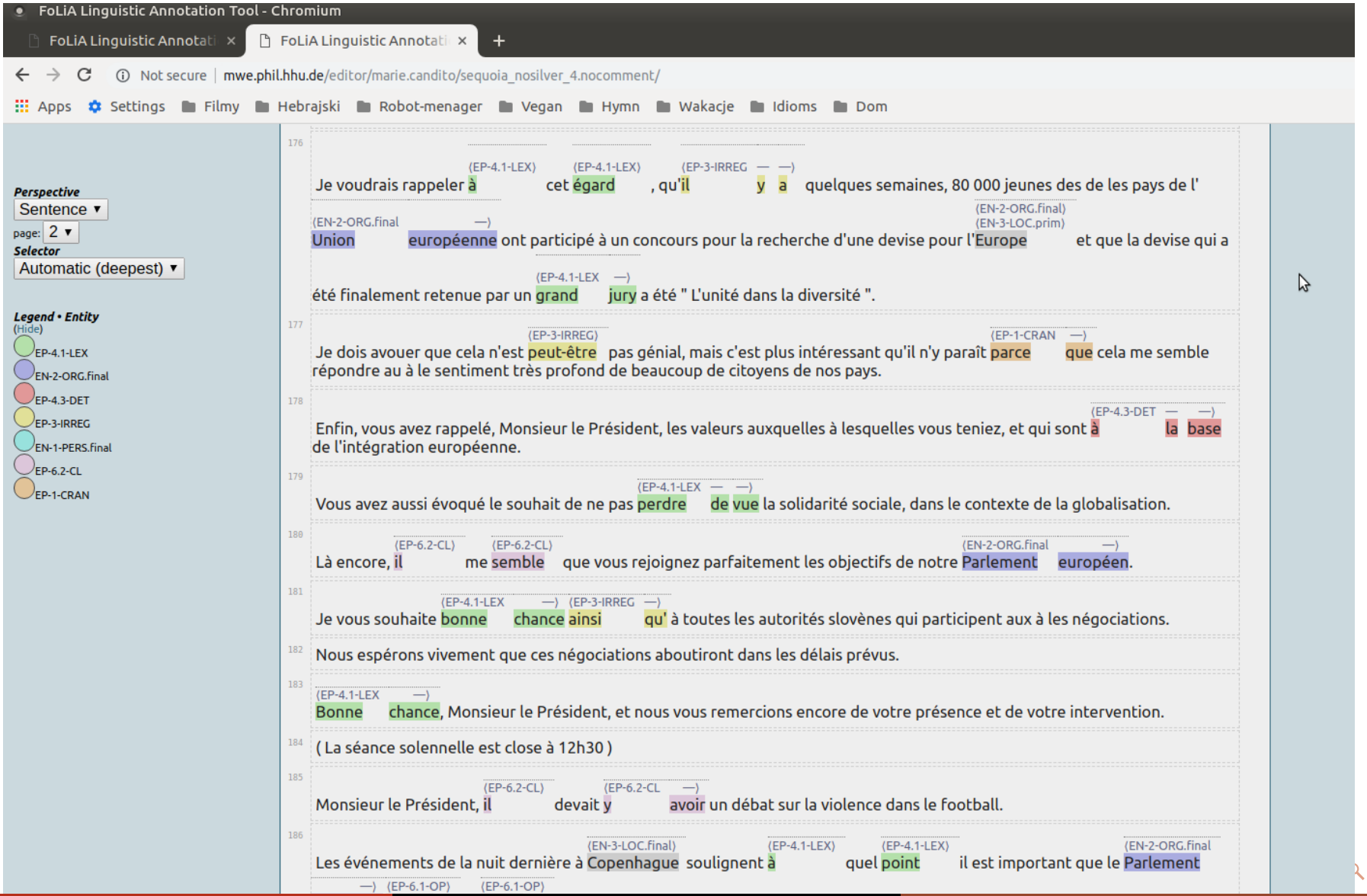
VMWE processing – identification in running text

- PARSEME shared task on automatic identification of verbal MWEs [8]

MWE game

Assign the 4 common senses to the literally translated multilingual MWEs.

Annotating MWEs in corpus



PARSEME multilingual corpus of verbal MWEs

International cooperation [11, 8]

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

Language groups

- **Balto-Slavic**: Bulgarian (BG), Croatian (HR), Lithuanian (LT), Polish (PL), Slovene (SL), Czech (CZ)
- **Germanic**: German (DE), English (EN), Swedish (SV)
- **Romance**: French (FR), Italian (IT), Romanian (RO), Spanish (ES), Brazilian Portuguese (PT)
- **Others**: Arabic (AR), Greek (EL), Basque (EU), Farsi (FA), Hebrew (HE), Hindi (HI), Hungarian (HU), Turkish (TR), Maltese (MT)

VMWE typology

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)'

VMWE typology

Language-specific categories

- inherently clitic verbs (**LS.ICV**) [7]

  IT ***prenderle*** 'to take it'⇒'to be beaten'

Unified multilingual annotation guidelines ▸ [link]

the fate of the republic rests on your shoulders

Annotation exercise

- Step 1: identify the candidate and its canonical form: *rests on your shoulders*
- Step 2: determine the lexicalized components
  - ***rests on*** *your/our* ***shoulders***, ***rests on*** *the* ***shoulders*** *of the deputies*, etc.
- Follow the ▸ decision tree
  - S.1 [1HEAD] (YES): *rests* is the only verbal head of the whole phrase
  - S.2 [1DEP] (YES): *on shoulders* is the only lexicalized dependent of *rests*
  - S.3 [LEX-SUBJ] (NO): *on* shoulders is not the subject of *rests*
  - S.4 [CATEG] (extended NP): *on shoulders* is a prepositional phrase
  - LVC.0 [N-ABS] (NO): *shoulders* is not abstract
  - VID.1 [CRAN] (NO): all components function also as stand-alone words
  - VID.2 [LEX] (YES): #*remains on your shoulders*, #*rests on your back/arms/head*
- Outcome: **VID**

Format and split

CUPT: extension of the CoNLL-U format

1	-	-	PUNCT	_	_	4	punct	_	_	*
2	si	si	SCONJ	_	_	4	mark	_	_	*
3	vous	il	PRON	_	_	4	nsubj	_	_	*
4	présentez	présenter	VERB	_	_	0	root	_	_	1:LVC.full
5	ou	ou	CCONJ	_	_	8	cc	_	_	*
6	avez	avoir	AUX	_	_	8	aux	_	_	*
7	récemment	récemment	ADV	_	_	8	advmod	_	_	*
8	présenté	présenter	VERB	_	_	4	conj	_	_	2:LVC.full
9	un	un	DET	_	_	10	det	_	_	*
10	saignement	saignement	NOUN	_	_	4	obj	_	_	1;2

Corpus	Sent.	Tokens	VMWE
train	208,420	**4,553,431**	**59,460**
dev	31,947	**672,102**	**9,250**
test	40,471	**846,798**	**10,616**
total	280,838	**6,072,331**	**79,326**

MWE identification (MWEI) [4]

- INPUT: text
- OUTPUT: text annotated with MWEs

PARSEME shared task on automatic identification of VMWEs [12, 8]

Goal

Automatically identify all VMWE occurrences in running text.

Two tracks

- **Closed**: only use the provided training/dev data
- **Open**: use the provided data + any external resource
  - corpora, lexicons, grammars, language models, word embeddings, ...

Evaluation dimensions

- Precision, recall and F1-measure
- **Per-language** scores vs. **cross-lingual** macro-averages
- **Precise-span** measure vs. **partial-match** measure
- **General** measure (all VMWEs) vs. **phenomenon-specific** measure (e.g. discontinuous VMWEs)

Systems and techniques

System	Lexicon matching	Phrase extraction & classification	Parsing	Sequential tagging	
				CRF	Neural networks
CRF-DepTree			✓	✓	
CRF-Seq			✓		
Deep-BGT			✓	✓	✓
GBD-NER					✓
Milos	✓			✓	
MWETreeC		✓		✓	
Polirem	✓				
Mumpitz				✓	✓
SHOMA			✓		✓
TRAPACC				✓	✓
TRAVERSAL			✓	✓	
varIDE		✓		✓	
Veyn					✓

MWE identification by sequential tagging

BIO tagging

The	prime	minister	paid	an	important	visit	to	the	president	.
O	B	I	B	O	O	I	O	O	O	O

BIO tagging with nesting

The	prime	minister	paid	a	few	important	visits	to	the	president	.
O	B	I	B	**b**	**i**	O	I	O	O	O	O

b and **i** stand for *begin* and *inside* of nested MWEs

Sequential tagging

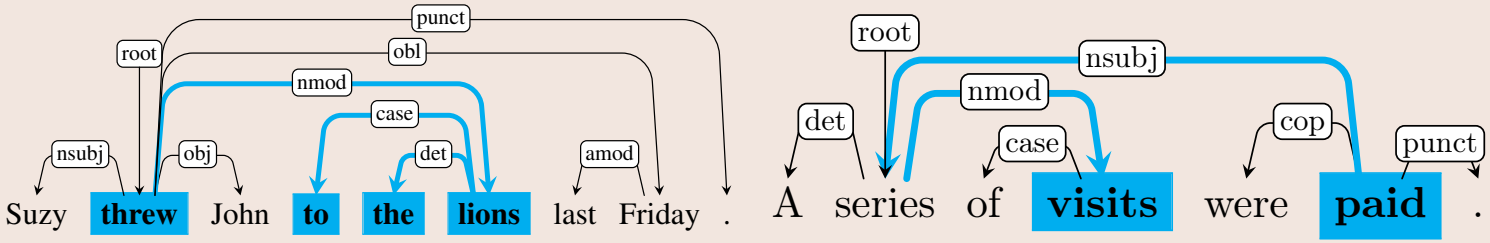
Decoding (finding the most probable sequence of tags) can be done by a sequential tagger (a model trained on annotated data), based on:

- Hidden Markov Model (+ Viterbi algorithm)
- Conditional Random Fields
- bi-directional Long Short Term Memory networks

Parsing-based MWE identification

Parsing data

They are available on input (columns 1-10 of the .cupt format, slide 27). Examples:



Technique 1: sequential tagging with parsing features

- The parsing data (e.g. the dependency labels) are used as **features** in CRF [CRF-Seq]
- The parsing data are **attached to word embeddings** on input of a neural net [Deep-BGT, Mumpitz]
- Discontinuities in VMWEs are handled by [9] self-attention[a] and a **graph convolutional network**[b] which takes on input all the words syntactically connected to the current word

---

[a] Self-attention = an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence

[b] A CNN = a NN in which neuron in one layer is connected to only a subset of neurons in the preceding layer

Parsing-based MWE identification

Technique 2: (parsing-based) candidate extraction + classification [MWETreeC]

- Parsing data are used to extract **VMWE candidates** as connected dependency subtrees
- The candidates are **classified** based on various morpho-syntactic features (including dependencies)

Technique 3: candidate extraction + parsing-based classification [varIDE]

- Words from seen VMWEs are used to identify **VMWE candidates** (disregarding syntax)
- The candidates are **classified** based on various morpho-syntactic features (including dependencies)

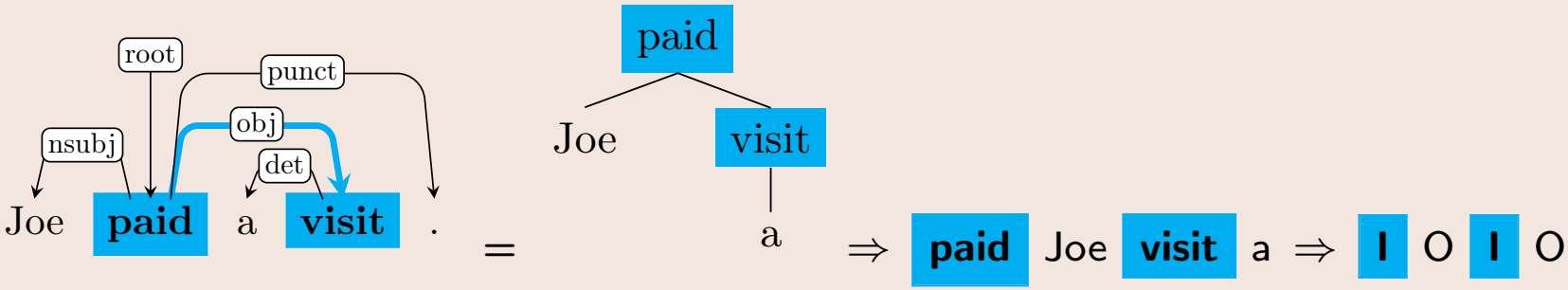
Technique 4: joint parsing and MWEI [TraPacc]

- An elaborate dependency parser is designed to perform **parsing and MWE identification in parallel**
- Parsing data from .cupt are used to train the parser

Parsing-based MWE identification

Technique 5: parsing tree traversal + node labeling [TRAVERSAL]

- The sentence is read not linearly but following the traversal of the syntactic tree
- CRF-based sequential IO (B is omitted) tagging is applied to this traversal



Some results (more ▸ online)

Cross-lingual macro-average scores

submission	track	P	R	F1
TRAVERSAL	closed	67.58	44.97	**54.00**
TRAPACC-S	closed	62.28	41.40	49.74
TRAPACC	closed	55.68	44.67	49.57
CRF-Seq	closed	56.13	39.12	46.11
varIDE	closed	61.49	36.71	45.97
SHOMA	open	66.08	51.82	**58.09**

Best phenomenon-specific F-scores

Seen-in-train	76.31
Unseen-in-train	28.46
Identical-to-train	87.63
Variants-of-train	65.02
Continuous	62.74
Discontinuous	44.36

Top scores for the languages with the biggest corpora and complete lemma annotation

	BG	PL	PT	RO
#VMWEs	6.7K	5.2K	5.5K	5.9K
unseen ratio	.33	.28	.28	.05
Best non-NN F1	.63	.67	.62	.83
Best NN F1	.66	.64	.68	.87

NER results for comparison (single words included but majority of MWEs)

Persons, locations, organisations, etc.

Language	NEs	CoNLL 2002–2003 [13, 14]			NN-NER survey 2018 [15]			
		Techniques	Gaz.	F1	NN	Model	Gaz.	F1
English	35,000	MaxEnt, HMM	✓	0.86	✓	word+char		0.92
German	20,000	MaxEnt, HMM	✓	0.71	✓	word+		0.79
Dutch	13,000	decision trees		0.77	✓	char.+		0.87
Spanish	18,000	decision trees	✓	0.81	✓	affix		0.87

		Reference corpus tool		
		Techniques	Gaz.	F1
Polish	87,000	CRF		0.77

Biomedical terms (English) [3]

Genes & proteins				Disorders				Chemicals			
NEs	Tech.	Gaz.	F1	NEs	Tech.	Gaz.	F1	NEs	Tech.	Gaz.	F1
24,000	CRF	✓	0.86	3,228	CRF	✓	0.81	20,000?	CRF		0.86

Unseen data: *Hard nuts to crack*

Seen data

A VMWE from the corpus is considered seen if a VMWE with the **same multi-set of lemmas** is annotated at least once in the training corpus.

Scores for languages with the biggest corpora and full lemma annotation

System	VMWEs	BG				PL				PT			
		IRV	LVC	VID	All	IRV	LVC	VID	All	IRV	LVC	VID	All
TRAVERSAL	seen	.89	.63	.55	.76	.92	.76	.57	.85	.89	.77	.69	.78
	unseen	.26	.06	.07	.13	.26	.20	.04	.17	.12	.25	.07	.20
SHOMA	seen	.92	.65	.58	.78	.90	.69	.58	.82	.86	.88	.84	.87
	unseen	.59	.21	.10	.31	.24	.19	.04	.18	.42	.35	.08	.31

NER results for comparison (English) [1]

On CoNLL-2003 unseen[a] data: from **0.81** to **0.94**

[a] There: unseen data = surface forms present only in the test

Why are MWEs so much harder to identify?

Idiosyncrasy

- **NEs**: idiosyncrasy of **tokens**
  - **trigger** words (*lake*, *association*, *Mr.*)
  - **graphical** features (uppercase, digits)
- **MWEs**: idiosyncrasy of **types**
  - **no/few trigger** words
  - **no graphical** features

Semantic similarity

- **NEs**: semantic similarity of component words
  - if we have seen *association* an NE component, *organisation*, *counsel*, etc. are simpler to predict as (unseen) MWE components
- **MWEs**: weak semantic similarity between component words of different MWEs

Other challenges

- **NEs**: continuity, slight variation
- **VMWEs**: discontinuity, strong variation

Future work

- Explicitly address the strong sensitivity of MWEI to **unseen data** (e.g. via eval. measures and shared subtasks)
- Couple MWE identification with MWE **discovery**, via **NLP-applicable syntactic lexicons** of MWEs
- Extend **discovery** methods so as to
  - not only lists of candidates but also some of their **syntactic structures**
  - cover **all syntactic types** of MWEs
  - adapt to **many languages**
  - **enrich** the existing **lexicons** rather than extract from scratch
- Such a lexicon should:
  - contain lemmas and POS of the lexicalized components, and **syntactic structures** of some morpho-syntactic variants
  - be distributed in **extensional** corpus-compatible formats
  - encode **rare MWEs** with high priority

Mid-term objective

Unified multilingual reference datasets with **MWE-annotated corpora** (extended to new, non-verbal MWE categories) and NLP-oriented **MWE lexicons**.

Keep an ear to the ground 'keep informed'

MWE community

- PARSEME - European network on parsing and MWEs
- MWE section of SIGLEX (special interest group at the ACL) - join both

Keep an ear to the ground 'keep informed'

MWE events

- Yearly MWE workshop co-located with major NLP conferences
  - Joint event with the Linguistic Annotation Workshop community (LAW-MWE-CxG at COLING 2018)
  - Joint event with the WordNet community (MWE-WN at ACL 2019)
- PARSEME shared task on automatic identification of MWE
  - Editions 1.0 and 1.1
  - New edition planned for 2020-21 (new languages and MWE categories)
- Yearly EUROPHRAS conferences
- MUMTTT workshops (on MWEs in MT)

Keep your nose to the wind 'keep informed'

Book series

Phraseology and Multiword Expressions, at Language Science Press, Berlin

- 2 volumes out, 3 others in the pipeline

MWE resources

- DIMSUM shared task dataset
- SIGLEX-MWE resource list
- PARSEME corpus of verbal MWEs edition 1.0 and 1.1 (18 & 19 languages) - open-ended project:
  - Italian team: Johanna Monti, Federico Sangati, Valeria Caruso, Manuela Cherchi, Anna De Santis, Maria Pia di Buono, Annalisa Raffone, ...
  - New languages and annotators are welcome
  - New MWE categories (adverbials, nominals, ...) will be addressed
- PARSEME annotation guidelines
- PARSEME surveys
  - On MWE annotation in treebanks
  - On lexical resources of MWEs
  - On multilingual MWE resources

*Why do we **eat, sleep and breathe** MWEs?*

'Why are we so enthusiastic and passionate about MWEs?'

- MWEs are fascinating!
  - They convey messages succinctly and efficiently
  - They hide traces of history, stereotypes, and surprising connotations
  - They can be very funny
- MWEs are challenging
  - They are hard to understand for non-native speakers
  - They are signs of a speaker's fluency
  - They behave differently than regular combinations of words
  - They are hard to tokenize, identify, parse, translate automatically
- They are prevalent

Bibliography I

Augenstein, I., Derczynski, L., and Bontcheva, K.
Generalisation in named entity recognition: A quantitative analysis.
Computer Speech & Language 44 (2017), 61 – 83.

Baldwin, T., and Kim, S. N.
Multiword expressions.
In *Handbook of Natural Language Processing*, N. Indurkhya and F. J. Damerau, Eds., 2 ed. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2010, pp. 267–292.

Campos, D., Matos, S., and Oliveira, J. L.
Biomedical named entity recognition: A survey of machine-learning tools.
In *Theory and Applications for Advanced Text Mining*, S. Sakurai, Ed. IntechOpen, Rijeka, 2012, ch. 8.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A.
Multiword Expression Processing: A Survey.
Computational Linguistics 43, 4 (2017), 837–892.

Gross, M., and Senellart, J.
Nouvelles bases statistiques pour les mots du français.
In *Proceedings of JADT'98, Nice 1998* (1998), pp. 335–349.

Kracht, M.
Compositionality: The very idea.
Research on Language and Computation 5, 3 (2007), 287–308.

Bibliography II

Monti, J., Cordeiro, S. R., Ramisch, C., Sangati, F., Savary, A., and Vincze, V.

Advances in Multiword Expression Identification for the Italian language: The PARSEME shared task edition 1.1.

In *Proceedings of Fifth Italian Conference on Computational Linguistics (CLiC-it)* (2018).

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.

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)* (2018), Association for Computational Linguistics, pp. 222–240.

Rohanian, O., Taslimipoor, S., Kouchaki, S., Ha, L. A., and Mitkov, R.

Bridging the gap: Attending to discontinuity in identification of multiword expressions.

CoRR abs/1902.10667 (2019).

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D.

Multiword Expresions: A Pain in the Neck for NLP.

In *Proceedings of CICLING'02* (2002), Springer.

Bibliography III

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

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., and Doucet, A.
The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions.
In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (Valencia, Spain, April 2017), Association for Computational Linguistics, pp. 31–47.

Tjong Kim Sang, E. F.
Introduction to the conll-2002 shared task: Language-independent named entity recognition.
In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20* (Stroudsburg, PA, USA, 2002), COLING-02, Association for Computational Linguistics, pp. 1–4.

Tjong Kim Sang, E. F., and De Meulder, F.
Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition.
In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* (2003), pp. 142–147.

Bibliography IV

Yadav, V., and Bethard, S.

A survey on recent advances in named entity recognition from deep learning models.
In *Proceedings of the 27th International Conference on Computational Linguistics* (Santa Fe, New Mexico, USA, Aug. 2018), Association for Computational Linguistics, pp. 2145–2158.