# Introduction to annotating verbal multiword expressions in the PARSEME framework


Carlos Ramisch, Agata Savary

Aix-Marseille Université, Université Paris-Saclay, France

UniDive webinar, 19 June 2023

# PARSEME

## Network

- COST Action on **Parsing and Multiword Expressions** (MWEs) funded by European Commission in **2013-2017**, still **active**
- 31 countries, 30 languages and 6 dialects from 10 language genera
- Outcomes: publications, resources, tutorials, methodologies, PMWE book series

## MWE corpora (https://gitlab.com/parseme/corpora/-/wikis/)

- **Collaborative** effort: 26 language teams, 35 language leaders, 200 annotators
- Annotation **guidelines** for **verbal** MWEs **unified** across 26 languages
- Corpora **manually** annotated for MWEs: **26 languages**, open licenses
- **Continuous enhancements** of the guidelines and corpora

# Multiword 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

*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**.*

## A definition

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

## Idiosyncrasy

A mode of behaviour or a property which is **particular** to an (few) individual(s). An **unusual** feature.

# Major idiosyncrasy 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'

# Major idiosyncrasy 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'

## Challenge

Semantic non-compositionality is **hard to test directly**.

# Inflexibility: a proxy for semantic non-compositionality

**Hypothesis**

A MWE is **less flexible** 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 |
| *the die is stolen* $\approx$ *someone stole the die* | ***the die is cast*** vs. *#someone cast the die* | Syntactic inflexibility |

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

# Focus on **verbal** MWEs – some challenges

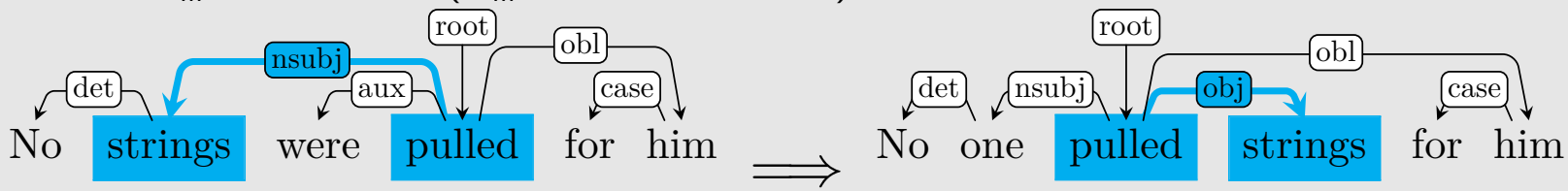
- Discontinuity:
  - EN *Trying hard to **bear** all these more or less important indications **in mind***
- Interleaving:
  - EN ***take** the fact that I **gave up** **into account***
- Multiword tokens
  - DE ***auf**/**machen*** (lit. '*out*/*make*') 'open' vs. *macht auf*
- Flexibility: morphological, syntactic, lexical
  - EN *he **broke** my **fall*** vs. *both of my **falls** were hard to **break***

# Neutralizing flexibility

## Canonical form

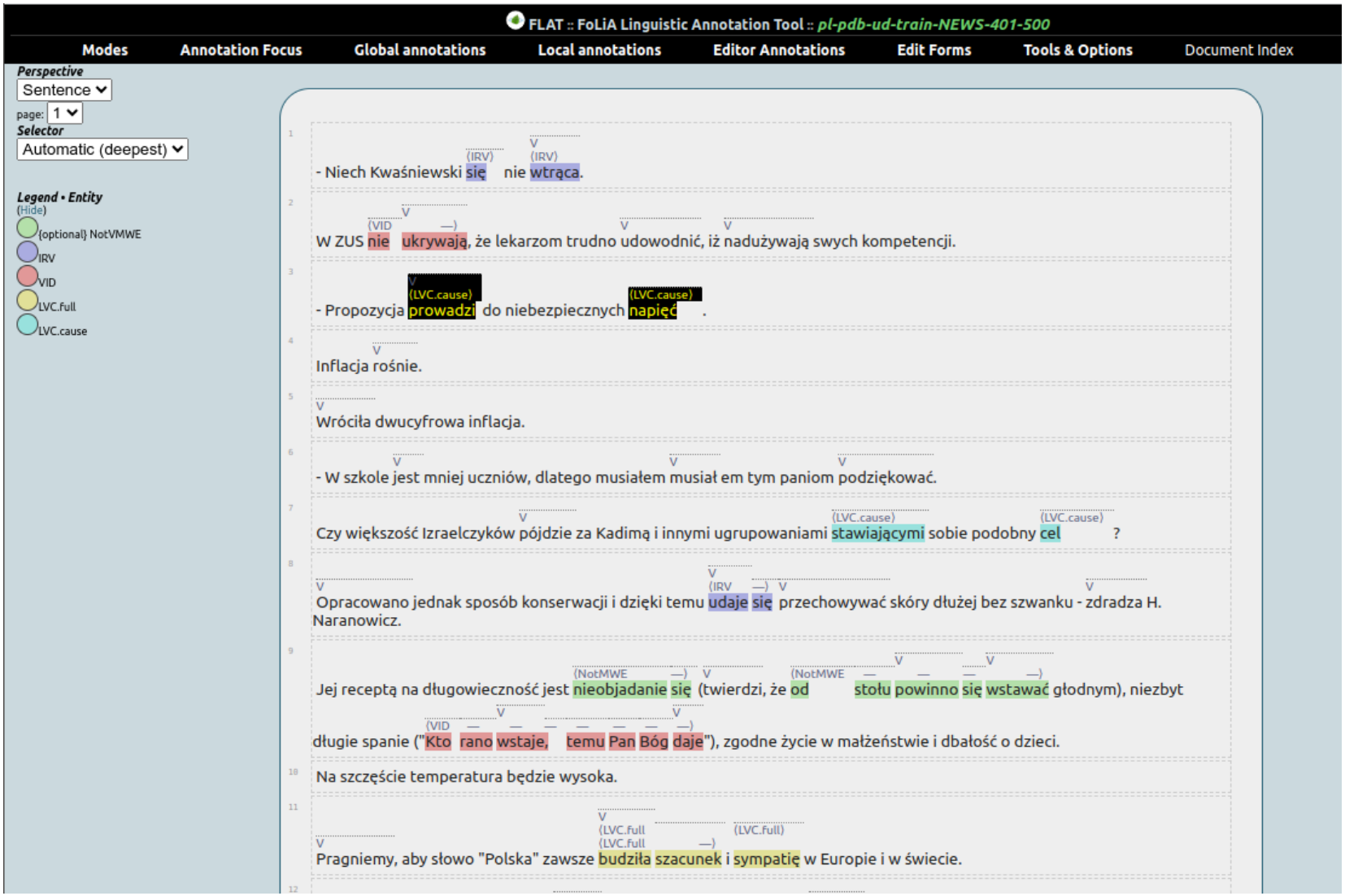
Least syntactically marked syntactic variant which preserves the idiomatic reading.

finite verb $<_m$ infinitive/participle; active voice $<_m$ passive v.; non-negated form $<_m$ negated f.; no extraction $<_m$ extraction, ... ($<_m$ = less marked than)




Canonical forms are useful for **formalizing** the morpho-syntactic properties of MWEs. This is useful e.g. for **annotation guidelines**.

# Annotating MWEs in a corpus



# PARSEME annotation guidelines

(https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3)

## Objectives

- Formalise idiomaticity in a **cross-linguistically unified** and **computationally tractable** way
- Unify what is truly **similar**, emphasise what is **language-specific**
- Make the annotation **reproducible**

# VMWE typology (v. 1.3)

- **Universal** categories (valid for all languages):
  - light verb constructions (**LVCs**)
    - **LVC.full**: EN *to **give** a **lecture***
    - **LVC.cause**: EN *to **grant rights***
  - verbal idioms (**VIDs**)

    EN *to **call it a day***
- **Quasi-universal** categories (valid for many languages):
  - inherently reflexive verbs (**IRVs**)

    FR ***s'évanouir*** 'to faint'
  - 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**)

    HI ***kar le-na*** (lit. *'do take.INF'*) 'to do something (for one's own benefit)'
- **Experimental** (optional) category
  - inherently adpositional verbs (**IAVs**)

    EN *to **come across** sth/sb*, *to **rely on** sth/sb*

# Towards reproducibility – guidelines as decision diagrams

If you are annotating **Italian** or **Hindi**, go to the Italian-specific decision tree or Hindi-specific decision tree. I

- ↳ Apply test S.1 - [**1HEAD**: Unique verb as functional syntactic head of the whole?]
  - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
    - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
    - ↳ **NO** ⇒ It is not a VMWE, **exit**
  - ↳ **YES** ⇒ Apply test S.2 - [**1DEP**: *Verb v has exactly one lexicalized dependent d?*]
    - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
      - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
      - ↳ **NO** ⇒ It is not a VMWE, **exit**
    - ↳ **YES** ⇒ Apply test S.3 - [**LEX-SUBJ**: *Lexicalized subject?*]
      - ↳ **YES** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
        - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
        - ↳ **NO** ⇒ It is not a VMWE, **exit**
      - ↳ **NO** ⇒ Apply test S.4 - [**CATEG**: *What is the morphosyntactic category of d?*]
        - ↳ **Reflexive clitic** ⇒ Apply IRV-specific tests ⇒ *IRV tests positive?*
          - ↳ **YES** ⇒ Annotate as a VMWE of category **IRV**
          - ↳ **NO** ⇒ It is not a VMWE, **exit**
        - ↳ **Particle** ⇒ Apply VPC-specific tests ⇒ *VPC tests positive?*
          - ↳ **YES** ⇒ Annotate as a VMWE of category **VPC.full** or **VPC.semi**
          - ↳ **NO** ⇒ It is not a VMWE, **exit**
        - ↳ **Verb with no lexicalized dependent** ⇒ Apply MVC-specific tests ⇒ *MVC tests positive?*
          - ↳ **YES** ⇒ Annotate as a VMWE of category **MVC**
          - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
            - ↳ **YES** ⇒ Annotate as a VMWE of category **ID**
            - ↳ **NO** ⇒ It is not a VMWE, **exit**
        - ↳ **Extended NP** ⇒ Apply LVC-specific decision tree ⇒ *LVC tests positive?*
          - ↳ **YES** ⇒ Annotate as a VMWE of category **LVC**
          - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
            - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**

# QA

Any questions?

# Annotation exercise 1 ▸ [FLAT] ▸ [guidelines]

*the fate of the republic rests on your shoulders* (sentence 4)

# Annotation exercise 1 ▸ [FLAT] ▸ [guidelines]

*the fate of the republic rests on your shoulders* (sentence 4)

- 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**

# Annotation exercise 2

*I hate to put a little pressure on you* (sentence 4)

# Annotation exercise 2

*I hate to put a little pressure on you* (sentence 4)

- Step 1: identify the candidate and its canonical form: *put a little pressure on you*
- Step 2: determine the lexicalized components
  - ***put*** *a little* ***pressure*** *on you*, *put more/no/a lot of pressure*, etc.
- Follow the ▸ decision tree
  - S.1 (YES) → S.2 (YES) → S.3 (NO) → S.4 (extended NP) →
  - LVC.0 [N-ABS] (YES): *pressure* is abstract
  - LVC.1 [N-PRED] (YES): 2 semantic arguments: (i) the person putting pressure, (ii) the person subject to the pressure
  - LVC.2 [V-SUBJ-N-ARG] (YES): I is the subject of *put* and the agent of *pressure*
  - LVC.3 [V-LIGHT] (YES): ***put pressure*** ≈ force
  - LVC.4 [V-REDUC] (YES): *my pressure on you*
- Outcome: **LVC.full**

# Annotation exercise 3

*Mr Osborne signed up with a US speakers agency* (sentence 26)

# Annotation exercise 3

*Mr Osborne signed up with a US speakers agency* (sentence 26)

- Steps 1-2: identify the candidate and its canonical form: *sign up* or *sign up with*
  - A VMWE in its proptotypical form is a verbal phrase in active voice whose head verb is in a finite for and whose other lexicalized components depend either on the verb or on another lexicalized component
  - In the lexical (as opposed to functional) approach to dependency grammar prepositions depend on the nouns they introduce ⇒ *sign up*
- Follow the ▸ decision tree
  - S.1 (YES) → S.2 (YES) → S.3 (NO) → PREP.EN.1 (YES) → S.4 (particle) →
  - VPC.1 [PART-REDUC] (YES): *sign up* 'to sign one's name (as to a contract) in order to join something' and *sign* refer to the same event
  - VPC.2 [PART-SPATIAL] (NO): *up* is not spatial in the context of *sign*
  - VPC.3 [PART-SPATIAL-LIT] (NO): there is no literal reading of *sign up* with *up* being spatial
- Outcome: **VPC.semi**

# Annotation exercise 4

*Opportunity for Beijing to demonstrate its ambitions* (sentence 44)

# Annotation exercise 4

*Opportunity for Beijing to demonstrate its ambitions* (sentence 44)

- Steps 1-2: identify the candidate and its canonical form: *demonstrate its ambitions*
- Follow the ▸ decision tree
  - LVC.0 [N-ABS] (YES): *ambition* is abstract
  - LVC.1 [N-PRED] (YES): 2 semantic arguments: (i) the person/group having the ambition, (ii) object of the ambition
  - LVC.2 [V-SUBJ-N-ARG] (YES): *Bejing* is the subject of *demonstrate* and the agent of *ambitions*
  - LVC.3 [V-LIGHT] (NO): *ambitions* can exist without being demonstrated
  - VID.1, VID. 2, VID.3, VID.4, VID5 (NO): *demonstrate*/*show*/*display its ambitions*/*aspirations*/*desires*
- Outcome: **no VMWE**

# Homework

Take the text in English here and try to annotate the following sentences (FLAT is not required):

- *We face a lot of competition* (sentence 14)
- *There are parallels to draw* (sentence 17)
- *Vote was cast* (sentence 18)
- *The date for cutting the first steel* (sentence 27)
- *To charge passengers an access fee* (sentence 36)
- *put new limits* (sentence 54)
- *The few ruin it for the many* (sentence 45)
- *Took down popular websites* (sentence 46)
- *We are moving in the right direction* (sentence 50)

# PARSEME infrastructure

- PARSEME ▸ [wiki] - extensive documentation of corpora and tools
- Language leaders guide
- User guides
- Gitlab repositories for all languages
- Corpus validators, converters, filters, release automation ...
- Data quality tools
  - Consistency checks
- Corpus browser: Grew-match $\longrightarrow$ see **next tutorial**

# PARSEME annotation framework – conclusions

## Principles and constraints

- The annotations guidelines are **unified** accross 26 languages with relatively few **language-specific** sections
- Annotation follows a **decision diagram** (unique starting point), for the sake of **reproducibility**
- Non-compositionality is a matter of **scale** but decisions must be **binary**
- **Semantic non-compositionality** is the major property to capture but is **hard to test directly**
- Lexical and morpho-syntactic **inflexibility** is considered a **proxy** for semantic non-compositionality
- Inflexibility tests are driven by the **syntactic structure**
- Strong dependence on the underlying **syntactic theory**
- PARSEME annotation largely **relies on the UD** annotation of morpho-syntax

# QA

Any questions?