PARSEME-FR Annotation guide for MWEs

Marie Candito, Matthieu Constant, Yannick Parmentier, Carlos Ramisch, Agata Savary

October 3, 2016

PARSEME-FR Working Group 1

- Define a set of operational criteria to identify and categorize multiword expressions
- Build a gold standard with comprehensive and deep annotation of MWEs, for evaluation purpose only
- Build a lexicon of annotated MWEs with sophisticated information (syntactic structure, list of properties, ...)

$\rightarrow$ Common basis of the project, adaptable to everyone's needs

$\rightarrow$ possibly correlated with "filtering" work on existing MWE lexicons

$\rightarrow$ in parallel, shared task PARSEME

Corpus

- Sequoia Corpus (Candito and Seddah 2012)
- Already annotated in deep syntax
- Subparts: Europarl, Wikipedia, Est-Républicain newspaper, European Medicine Agency reports
- 67000 tokens, 3099 sentences
- Also planned to work on part of French UD treebank

$\rightarrow$ Free corpus

Methodology to write the guidelines

Participants: Agata, Carlos, Marie, Mathieu, Yannick

1. For each type of MWEs, a person in charge of reading bibliography (focus on linguistic properties) + summary
2. First draft of **sufficient criteria**
3. Pilot annotation phase + adjudication
4. Criteria update
5. If not satisfying, return to (3)

$\rightarrow$ Currently: 4th pilot annotation phase

Manual identification of MWEs

- **Important choice**: we consider a list of **sufficient criteria**

→ Solution to the well-known problem of deciding the MWE status on a binary basis

→ Ideally, once an MWE is identified thanks to at least one criteria, the properties associated with other criteria should be encoded

→ Filters will be applicable to the list of annotated MWEs, according to people's needs

- Multiword Named Entities are also considered and annotated, but treated with different criteria (organizations, locations, person names, products)

Manual identification of MWEs

1. By sequentially reading text, one intuitively identifies sequences in the text that could be an MWE
2. This candidate sequence is then categorized with a POS tag
3. One successively applies different linguistic criteria (some being specific to the POS tag)
4. If one criterion works, one marks the sequence as MWE (as well as the criterion used)

If one criterion should remain...

Semantic non-compositionality

One of the (non-empty) items of the expression does not have its usual meaning

Clear cases

- *kick the bucket*: corresponds to *die*
- *arme blanche* (lit. weapon white) = cutting weapon
- *dans la foulée* (lit. in the stride) = immediately afterwards

If one criterion should remain...

Semantic non-compositionality

One of the (non-empty) items of the expression does not have its usual meaning

Difficult cases: additional meaning

- *red wine*: is a wine, which has a red color, but has some additional features in relation with the way it is made
- *devoir conjugal* (conjugal duty): seems to be a duty linked to married couples (conjugal); but isn't there something more?

$\rightarrow$ Need for other criteria to capture this additional meaning!

Formal criteria: principles

Principle

- A MWE exhibit **irregular** morphological, syntactic and/or semantic properties
- Identifying MWEs = tracking irregularities on different linguistic levels
- Approach very close to the lexicon-grammar methodology (M. Gross 1983)

Example

- *casser les pieds* (lit. break the feet) = to annoy
- passive is impossible while possible with verb *casser* (break)
- *pieds* must be plural etc...

List of criteria

1. Cranburry word [CRAN]
2. Semantic
   - Identity [ID]
   - Predicativity [PRED]
3. Irregular morphosyntactic structure [IRREG]
4. Lexical fixity
   - lexical substitution [LEX]
   - determiner [DET]
   - zero determiner [ZERO]
   - preposition [PREP]
5. Morphosyntactic fixity
   - morphosyntactic features [MORPHO]
   - insertions [INSERT]
6. Specific criteria
   - Operator [OP] $\rightarrow$ to capture light verb constructions
   - Pronominal verbs [seV]
   - fixed clitics [CL]

"Cranberry" word

- A word in the expression cannot work as isolated word
- Examples:
  - verb: *prendre la poudre d'**escampette*** (to run away)
  - adverbial: *en **catimini*** (in the sly, in secret)
  - conjunction: ***tandis** que* (whereas)

Semantic tests

1. ID: X Y => Z? with Z being the syntactic head

   cordon bleu (lit. string blue) $\neq$ cordon (string)
   nager dans le bonheur (lit. swim in the happiness) = to be happy
   $\neq$ nager (swim)

2. PRED: one cannot find any predicative relation

   arme blanche (lit. weapon white) = cutting weapon
   **une arme qui est blanche* (a weapon that is white)

Irregular internal syntax

- The internal morphosyntatic structure of the candidate MWE is irregular with respect to its POS tag
- adverbial: *en outre* (Prep Prep)
- noun: *à-coup* (Prep Noun); in context, *un à-coup* (*Det Prep Noun)

Lexical fixity

1. LEX: no possible lexical substitution of a full word by a "neighbor"
   - *eau/#boisson de vie* (lit. water/drink of life) = brandy
   - *prendre/*saisir un virage* (lit. take/*grab a turn) = take a bend
2. DET: fixed determiner (zero determiner included)
   - *perdre la face* (lit. lose the face) = lose face
   - *garde du/#d'un corps* (lit. guard of the/a body) =bodyguard
3. ZERO (for verbal expressions): the object **can** have a zero determiner
   - *prêter attention/prêter une attention particulière*
     (lit. lend attention/lend a particular attention) = pay attention
4. PREP (for adverbials): the introducing preposition is fixed and the noun does not commute with a whole class of nouns
   - *dans la foulée/à la suite (de)* $\rightarrow$ MWE
   - vs. *en (avion+train+voiture)* $\rightarrow$ not an MWE

Morphosyntactic fixity

1. MORPHO: one cannot modify morphosyntactic features
   - *perdre *la/les pédale(*ɛ+s)* – lit. loose *the.SING/the.PL pedal = loose control
   - *dans la/*les foulée* – lit. in the.SING/*the.PL stride(*s) = immediately afterwards
2. INSERT: one cannot insert material that is a priori possible to insert (syntactically and semantically)
   - *(*très) bien que* – lit. (*very) well that – = although
   - *en cours (*normal)* – lit. in course *normal – ongoing

Specific criteria for verbal expressions

cf. shared task guidelines

1. **OP**: verb has a neutral meaning (does not bring any additional meaning) AND noun phrase reduction AND impossibility to have two "actants"
   - both narrower and broader class than light verb constructions
2. **seV**: the addition of reflexive clitics is mandatory (*se suicider* = commit suicide) or change the meaning (*s'agir* (=be about) $\neq$ agir (= act)) or the valency (*confesser X, se confesser de X* (=confess)) of the verb
3. **CL**: same as *seV* but with non-reflexive clitics: *l'emporter, il y avoir, s'en aller*

Annotation scheme

Luc	
fait	1+OP
un	
faux	1/2+PRED
pas	1/2

- Annotation of lexical items of MWEs (*faire faux pas*)
- deep annotation: MWE embeddings *(faire (faux pas))*
- Possibility to annotate complex overlapping (pay_1 $close_2$ $attention_{1/2}$)

Thank you for your attention!!
Questions/Comments?