

Identifying variants of multiword expressions

Caroline Pasquer, Agata Savary, Jean-Yves Antoine
Carlos Ramisch, University of Aix-Marseille, France

✉ first.last@univ-tours.fr
first.last@lif.univ-mrs.fr

Verbal Multiword Expressions (VMWEs)

- MWEs = word combinations with unpredictable behavior
e.g. semantics: *to have egg on one's face* = 'to seem stupid' ≠
- VMWE Challenges
 - Scale-wise variability (≠ regular phrases): (1)(2) vs. (3) ! auto. identification
 - Ambiguity: (1) ≠ (4)
 - Discontinuity: (1) ≠ (2) ! sequence labeling

✓(1) *I broke_{pret} her heart_{sing} when I left*

✓(2) *Lives lost and hearts_{plur} broken_{past}*

✗(3) **I break my heart*

✗(4) *I broke her heart in chocolate when I stepped on it*

Each VMWE has a
Variability profile
(literal reading)

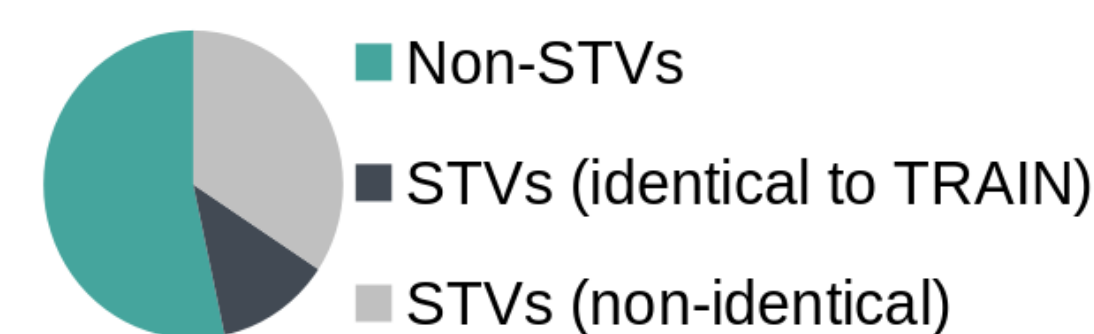
⇒ VMWE identification required by NLP applications

PARSEME corpus

- Rich verbal inflection in French
- manually annotated VMWEs
- syntactic dependencies
- lemmas, POS, morph. features
- UD v2 tagsets
- VERB-(DET)-NOUN VMWEs
 - Light Verb Constructions
jouer rôle 'play role'
 - Idioms
fermer la porte 'close the door'
- TRAIN: 854 VMWEs (2098 occ.)
- TEST: 177 VMWEs (283 occ.)

Focus on variant identification in corpus

- Goal = identification of known VMWEs whatever their form
Why? Variants are pervasive e.g. *Seen-in-train variants* (STVs)
- STV definition = same meaning but possibly different surface form e.g. *break heart* in (1)(2)
- Hypothesis = any VMWE_x shares similarities with VMWE_y already seen in corpus
 - VMWE_x = VMWE_y ⇒ VMWE-specific properties
 - VMWE_x ≠ VMWE_y ⇒ VMWE-generic properties, e.g. same restricted inflection in *I dance on my/your grave* as in (3)



Baseline ⇒ same lemmas + syntactic connection

- Strong baseline: F1 = 0.88

(a) *Il joue son propre rôle*
He plays his own role

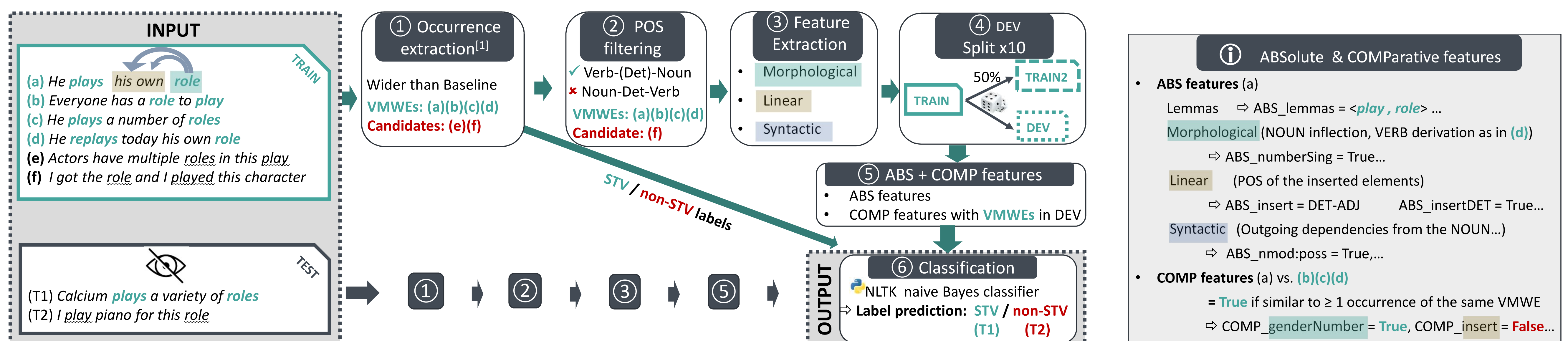
(b) *Chacun a son rôle à jouer*
Everyone has a role to play

(c) *Il joue un tas de rôles*
He plays a pile of roles

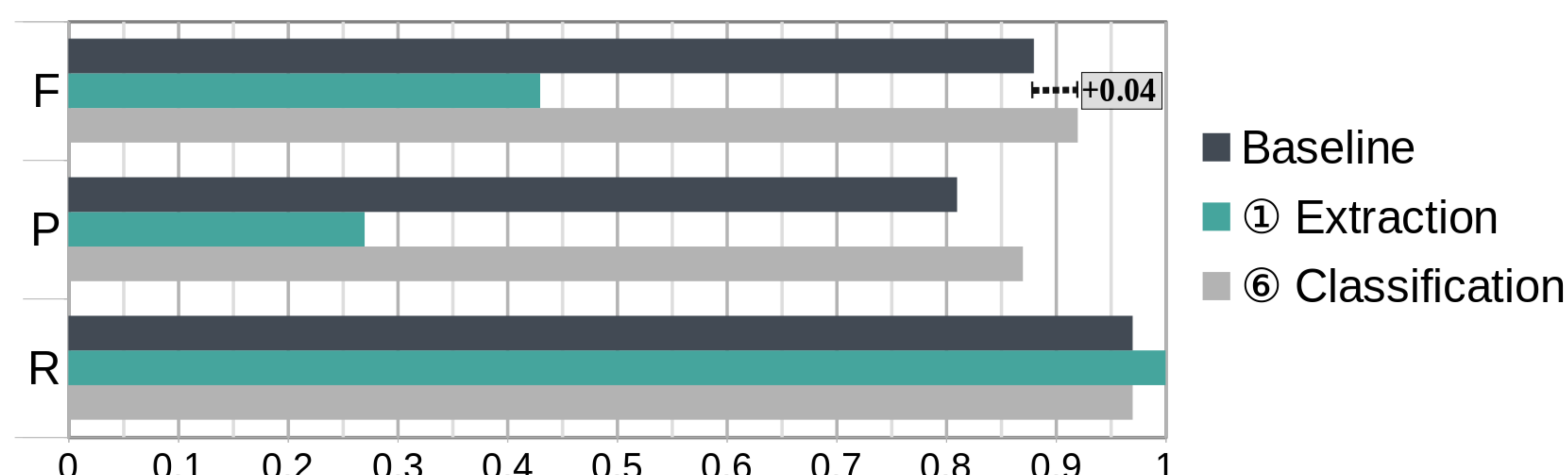
(c) Indirect syntactic connection

Our method ⇒ classification based on the Morpho-syntactic variability profiles

- Same lemmas (↑R but P↓) + ABSolute/COMParative VMWE features ⇒ (1)(2) gathered + (3)(4) discarded



Extraction & Classification results vs. Baseline



Discriminative features

- Linear features more discriminative than morphological features
- COMP features for **STV** prediction
e.g. same insertions
- ABS feat. for **non-STV** prediction
e.g. insertion of VERB, PUNCT...

Conclusions & Perspectives

- F1 = 0.92 > Baseline
- Good reproducibility $\sigma_{F1} = 0.01$ (10 samples)
- Larger corpora & not only Verb-(Det)-Noun ⇒ Workshop Poster, Aug 25^[2]
- Other features: orthographic similarity, lexical similarity

References, acknowledgements

- [1] SAVARY and CORDEIRO, TLT 2018
- [2] PASQUER *et al.*, COLING Workshop 2018
- French PARSEME-FR grant
- Thanks to S.Cordeiro for the extraction script