Être ou ne pas être … une expression polylexicale verbale : Recherche de traits discriminants

Caroline Pasquer, Agata Savary, Jean-Yves Antoine, Carlos Ramisch, Nicolas Labroche et Arnaud Giacometti

Blois, 13 juin 2019

Plan

- Contexte et motivation
- Recherche de traits discriminants
- Résultats
- Conclusion & perspectives

Contexte et motivations

Expressions polylexicales (EP) : caractéristiques

- **≥ 2 unités graphiques :**
  - “*avoir*$_1$ *du*$_2$ *pain*$_3$ *sur*$_4$ *la*$_5$ *planche*$_6$”
  - “*court*$_1$ *- circuiter*$_2$”

- **Idiosyncrasie : ‘AB’ ≠ ‘A’ + ‘B’**
  - “*faire-valoir*”$_{\text{Nom}}$ ≠ *faire*$_{\text{Verbe}}$ + *valoir*$_{\text{Verbe}}$
  - “*avoir du pain sur la planche*” ≠  + 

EP et TAL

- **EP fréquentes** [JACKENDOFF]
  - ✓ Lexiques — ▪ Populaire. Manger les pissenlits par la racine, être mort et enterré.
  - ✗ Variabilité
  - ✗ Nouvelles EP

© https://www.charentelibre.fr

- **Impact négatif**
  - Traduction : *"avoir un chat dans la gorge"* vs. *"to have a frog in one's throat"*
  - Parsing : 8% d'erreurs [BALDWIN]

⇨ Identification préalable des EP

EP : variabilité

- **≥ années 1980 [GROSS] … ≥ 2002 [SAG *et al.*]**
- **Éléments *lexicalisés***

  "il mange / bouffe maintenant$_{ADV}$ les pissenlits par la racine"
  # dents-de-lion
  ↑ discontinuité

- PARSEME Shared Task : **catégories d'EP verbales (EPV)**
  - LVC = *Light Verb Constructions* — *"prendre décision"*
  - VID = *Verbal IDioms* — *"avoir du pain sur la planche"*
  - IRV = Verbes Intrinsèquement Réflexifs — *"se prélasser"* (**prélasser*) , *"se rendre"* (= 'aller' ≠ 'rendre')
  - MVC = *Multi-Verb Constructions* — *"laisser tomber"* (= 'abandonner')

- **Variabilité des EPV**
  - 👍 *"Je prends une décision"* = *" Les décisions sont prises"*
  - 👎 *"Il jette l'éponge et divorce"* ≠ *"Il jette les éponges"*
    "Il jette l'éponge sale$_{ADJ}$"

⇨ Identifier les EPV grâce à leur profil de variabilité

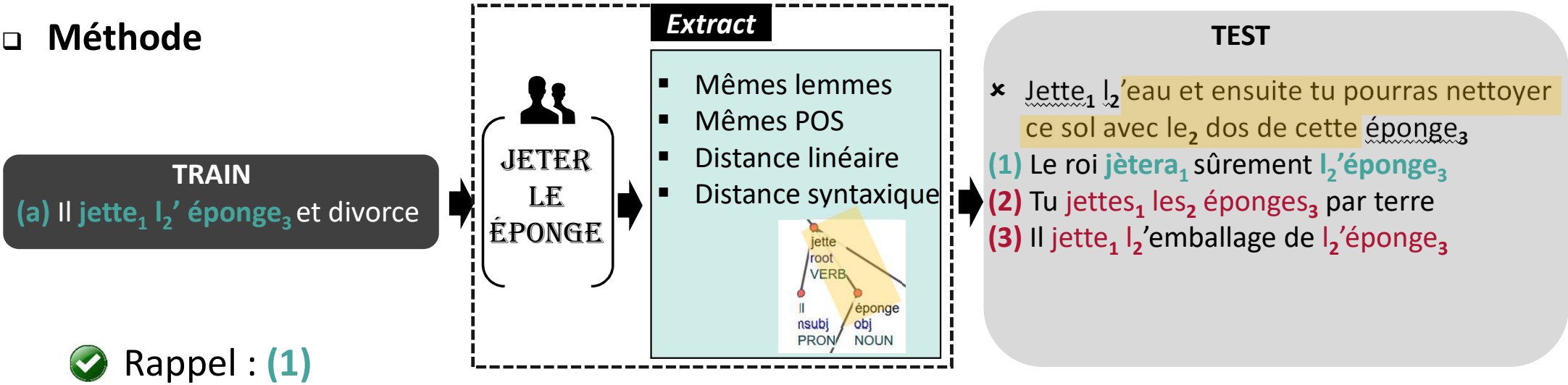
Tâche = Identification d'EPV déjà vues

- **Exemple**

  Corpus d'entraînement (TRAIN) annoté en EPV: "J'ai **pris** la **décision** de **faire appel** à cet artisan".

  Corpus de TEST : "Les **décisions** qu'ils **prennent** **jouent** un **rôle** pour l'avenir du pays."

- **Méthode**



- Rappel : (1)
- Précision : lectures littérales **(2)** + co-occurrences fortuites **(3)**

⇨ **Traits discriminants ?**

Recherche de traits discriminants

Traits discriminants +EPV vs. -EPV

- **Exemples +/- :**

- ***Extract* sur TEST :** **R = 1,00** **P = 0,57**

- **Traits :** lemmes, morphologie, insertions, dépendances syntaxiques (⇢ passives…)

TRAIN

(a) Il jette$_1$ l$_2$' éponge$_3$ et divorce
(b) Le roi jetera$_1$ sûrement l$_2$' éponge$_3$
(c) Jette$_1$ cette serpillère, mais garde les$_2$ éponges$_3$ propres$_{ADJ}$

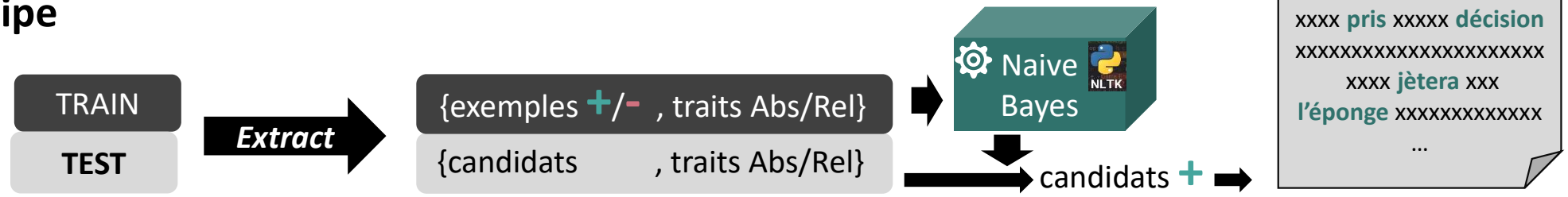
	Traits RELATIFS (même EPV)		
	Nombre NOM	Insert (POS)	Dep. NOM
(a) vs (b)	True	False	True
(b) vs (a)	True	False	True
(c) vs (a,b)	False	False	False

	Traits ABSOLUS		
	Nombre NOM	Insert (POS)	Dep. NOM
(a)	Sing	Ø	DET
(b)	Sing	ADV	DET
(c)	Plur	DET-NOUN PUNCT-CCONJ VERB	DET+ADJ

- EP ≥ 2 occ. ⇨ 78% TRAIN

Baseline = VarIDE [PASQUER et al.]

- PARSEME Shared Task **(2018)** : systèmes d'identification d'EPV en 19 langues [RAMISCH]
- **Principe**

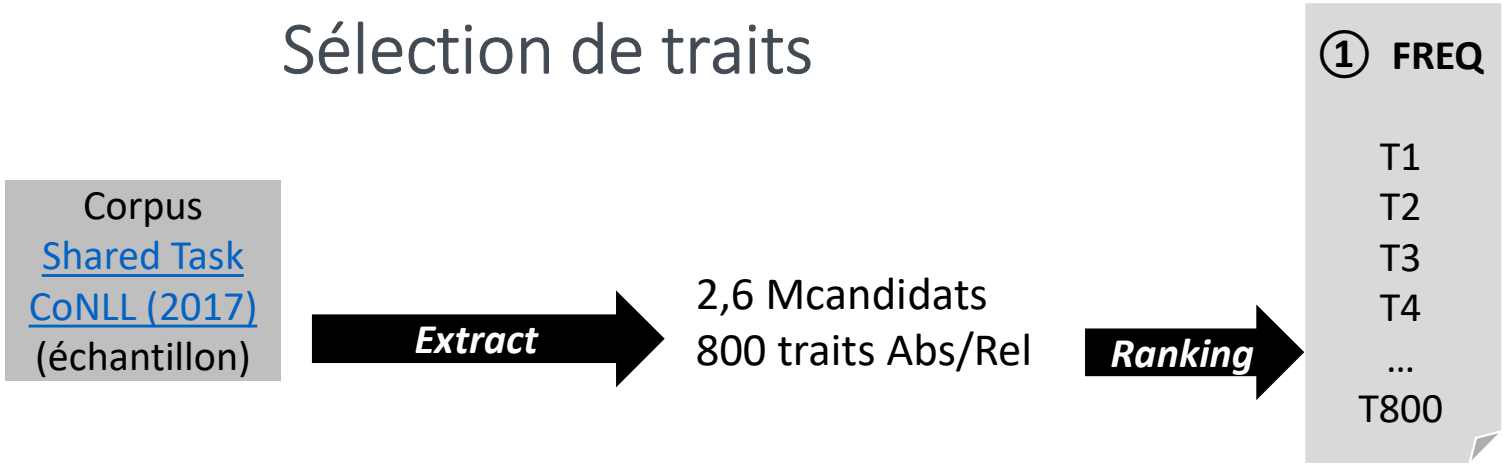


- Traits créés automatiquement parfois inappropriés
- Temps d'exécution

- **Si moins de traits :**
  - ⇨ classification performante & rapide ?
  - ⇨ mise en évidence de traits pertinents pour la Tâche ?

Sélection de traits

- **Traits utiles**



- **Traits classés**

① FREQ : fréquence ⬊
② GAIN : valeur de gain d'information ⬊
③ CHI² : valeur de Chi² ⬊
④ FOREST : pertinence ⬊ des traits

new **Sélection de traits & plusieurs classifieurs**

- Entraînement sur $TRAIN_{90\%}$
- Choix configuration d'après $TRAIN_{10\%}$

FREQ
CHI2
GAIN
FOREST

Top1
Top2
Top3…

Naïve Bayes
Arbre de décision
SVM linéaire

Evaluation

- **Corpus TEST** PARSEME Shared Task (**498** occ. **+ EPV**)

- **'WebCorpus' : extrait de Shared Task CoNLL (2017)** [ZEMAN *et al.*]
  - Sources = *Wikipedia* + webcrawling
  - Prétraitement automatique
    - Segmentation, tokenization,
    - Annotation morphosyntaxique

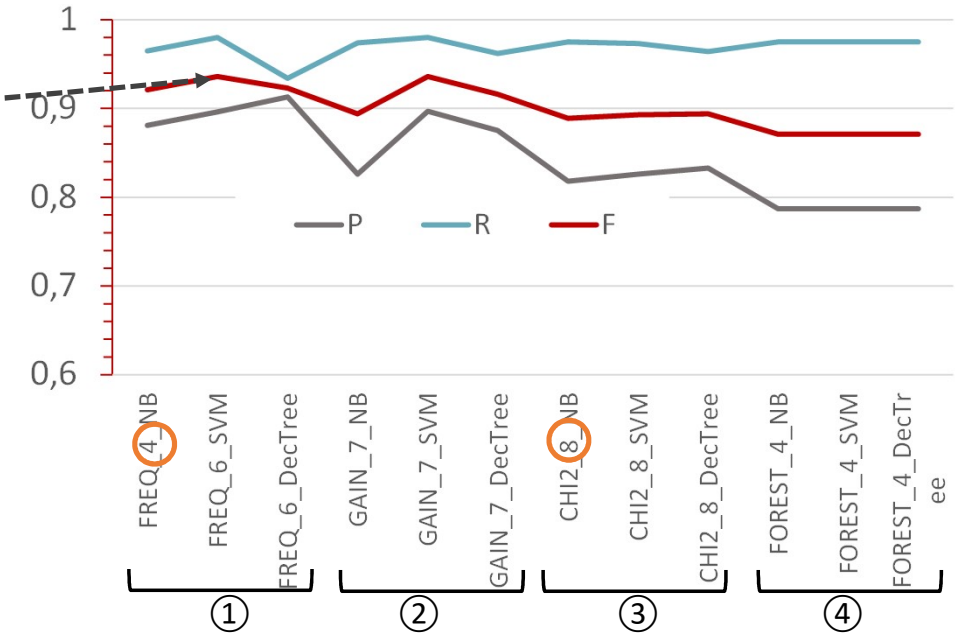
  - Sélection d'EPV représentatives :
    - Par catégorie
    - Par fréquence (haute / médiane / basse)

    ⇨ 90 EPV (4618 occ.)  69% **+EPV**

Résultats

Configuration optimale sur $TRAIN_{10\%}$ (EPV vues 2 fois)

- ① FREQ + 6 traits + SVM linéaire
  - ⇨ $\bar{F}_{10\%TRAIN} = 0{,}936 \quad \sigma = 0{,}015$
- Traits pertinents



F↘ 4pt

	ABS	REL (③ ④)
Lemmes (*'se' + 'emparer'*)	✓	
Lemme du verbe	✓	✓
Catégorie (LVC,…)	✓	
Séquences d'insertions (Ø, …)	✓	✓
Insertions spécifiques (VERB, PUNCT…)		✓
Nombre d'insertions		✓
Distance syntaxique		✓

Modifieur : Je **prends** une $grande_{ADJ}$ **décision**
Passive : Ma **décision** est_{AUX} **prise**
Relative : C'est la **décision** que_{PRON} je_{PRON} **prends**

Pas de traits morpho / dép. syntax
⇨ pas assez de VID ?

Performances

- **TEST** PARSEME Shared Task : VarIDE **vs. TRAVERSAL** [WASZCZUK]
  - F_{TEST} (vues + non vues) = 0,55 < F_{TEST} = 0,56
  - F_{TEST} (vues) = 0,8207 ≈ F_{TEST} = 0,8172 (> VarIDE (0,70), réseaux neurones)
  - F_{TEST} (vues : variantes) = 0,7317 > F_{TEST} = 0,7123
  - ⚠ EP vues ≥ 2 fois
- **WebCorpus :** F = 0,922
  - F_{LVC} (0,90) < F_{VID} (0,94)

Conclusion

Conclusion et perspectives

- **Contributions**
  - Performances ≈ meilleur système
  - Traits pertinents : traits relatifs, insertions
  - 6 traits ⇨ Interprétabilité & rapidité de classification
  - Généralisation sur corpus externe & large
- **Perspectives**
  - Analyse d'erreurs
  - Stratégie pour EPV vues 1 fois
  - Classifieur par catégorie d'EPV
  - Couplage identification d'EPV + découverte par *word-embeddings*
  - …

Références

- Baldwin,T., Bender, E.M., Flickinger,D., Kim, A. et Oepen, S. (2004). Road-testing the English Resource Grammar over the British National Corpus. In Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal
- Gross, G. (1988). Degré de figement des noms composés. Langages, 90:57–72.
- Gross, M. (1982). Une classification des phrases "figées" du français. Revue québécoise de linguistique, 11(2)
- Jackendoff, R. (1997). The Architecture of the Language Faculty. Numéro 28. Linguistic Inquiry Monographs.
- Pasquer, C. et al. (2018). Are Variants Really as Alikeas Two Peas in a Pod? In Proceedings of the Joint Workshop on Linguistic Annotation, Multi-word Expressions and Constructions (LAW-MWE-CxG-2018), pages 283–289. Association for Computational Linguistics.
- Ramisch, C. et al (2018). 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), pages 222–240
- Sag, I., Baldwin, T., Bond, F., Copestak, A. et Flickinger, D. (2002). Multiword expressions : A pain in the neck for NLP. In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), page 1–15, Mexico City, Mexico.
- Waszczuk, J. (2018). TRAVERSAL at PARSEME Shared Task 2018: Identification of Verbal Multiword Expressions Using a Discriminative Tree-Structured Model. In Proceedings of the Joint Work-shop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 275–282. Association for Computational Linguistics.
- Zeman D. et al. (2018) CoNLL 2018 Shared Task: Multilingual parsing from raw text to Universal Dependencies. Proceedings of the CoNLL 2018 SharedTask: Multilingual Parsing from Raw Text to Universal Dependencies, pages 1–21.

Merci de votre attention

Des questions ?

caroline.pasquer@univ-tours.fr