UNIVERSITÉ DE TOURS

École Doctorale MIPTIS

Équipe BdTln – LIFAT

THÈSE présentée par :

Caroline PASQUER

soutenue le : 13 novembre 2019

pour obtenir le grade de : Docteur de l'Université de Tours

Discipline/ Spécialité : Informatique

Garder la trace, mettre de l'ordre et relier les points : modéliser la variation et l'ambiguïté des expressions polylexicales

THÈSE DIRIGÉE PAR :

SAVARY Agata	Maître de Conférences, HDR, Université de Tours
ANTOINE Jean-Yves	Professeur, Université de Tours

RAPPORTEURS :

DAILLE Béatrice	Professeure, Université de Nantes
KALLMEYER Laura	Professeure, Université Heinrich Heine de Düsseldorf, Allemagne

JURY :

ANTOINE Jean-Yves	Professeur, Université de Tours
DAILLE Béatrice	Professeure, Université de Nantes
DIAS Gaël	Professeur, Université de Caen
KALLMEYER Laura	Professeure, Université Heinrich Heine de Düsseldorf, Allemagne
RAMISCH Carlos	Maître de Conférences, Université Aix-Marseille (co-directeur)
SAVARY Agata	Maître de Conférences, HDR, Université de Tours

Remerciements

Ce travail atteint son accomplissement grâce à plusieurs collaborations et soutiens, notamment celui de l'Agence Nationale de la Recherche dans le cadre du projet PARSEME-FR (projet ANR-14-CERA-0001), et je tiens à exprimer mes remerciements à plusieurs personnes.

J'exprime tout d'abord une reconnaissance toute particulière à ma directrice de thèse, Agata Savary, et à mes co-directeurs, Jean-Yves Antoine et Carlos Ramisch, qui ont dirigé cette recherche avec autant de rigueur que de gentillesse. Ils ont toujours su se rendre disponibles, malgré l'éloignement géographique. Leurs conseils avisés, leurs relectures minutieuses et leurs remarques méthodologiques m'ont permis d'avancer et de surmonter les difficultés qui pouvaient se présenter. Ils ont su me donner les clés pour apprendre à faire face à de nouveaux défis.

Je remercie Béatrice Daille, Laura Kallmeyer et Gaël Dias qui m'ont fait l'honneur d'accepter de faire partie du jury de thèse.

Je remercie Nicolas Labroche et Arnaud Giacometti pour leurs précieux conseils dans le choix et l'utilisation d'algorithmes de classification et de sélection de traits.

Je remercie Silvio Cordeiro pour le partage de codes et pour les échanges intéressants auxquels cela a donné lieu.

J'exprime ma gratitude à François Laurand pour sa disponibilité et son aide précieuse dans l'utilisation des ressources informatiques du Laboratoire et la gestion de données volumineuses.

Un grand merci va ensuite à mes collègues de l'équipe BdTln et du Laboratoire d'Informatique du site de Blois, à Lucie Kuhnel pour avoir rendu agréables les trajets Vendôme-Blois, aux doctorants et stagiaires passés et présents pour le climat chaleureux qu'ils ont instauré : Jakub Waszczuk, Adnan El Moussawi (pas seulement pour les viennoiseries), Mahfoud Djedaini, El Arby Sidi Aly, Rayene Bessrour, Lamine Diop, Clément Moreau, Frederick Bisone, Adam Lion-Bouton, Najet Hadj Mohamed...

Je rends ici hommage à mon grand-père Maurice pour avoir, durant mon enfance, alimenté mes centres d'intérêt *a priori* divergents mais aujourd'hui réunis dans le Traitement Automatique du Langage.

Merci à mes filles qui m'ont soutenue à leur façon dans ce travail et pour qui les expressions polylexicales n'ont (presque) plus de secrets. Merci Emmeline pour tes listes d'expressions polylexicales parfois improbables telles que ***croûte que croûte*** (Chaulet, 1975). Merci Héloïse pour ton interprétation très personnelle des expressions polylexicales.

Grâce à toi, j'ai pris la mesure de l'opacité de ces expressions : un jour, tu apprendras qu'une ***peur bleue*** ne se ressent pas obligatoirement devant un fantôme de cette couleur.

Last but not least, merci à mon conjoint Éric pour m'avoir supportée (au sens propre et figuré) durant toutes ces années où j'ai cherché ma voie. Qu'il soit rassuré, je l'ai trouvée !

Résumé

L'identification automatique d'expressions polylexicales (EP) est un pré-requis pour de nombreuses applications de traitement automatique des langues. Cette tâche représente un défi car les EP, et en particulier les verbales (EPV) telles que ***casser sa pipe*** (signifiant *mourir*), ont des formes de surface très variables (***cassera-t-il*** *un jour* ***sa pipe*** ?). Cependant, comparée à des constructions libres, cette variabilité est généralement plus restreinte (p. ex. certains noms non modifiables par un adjectif), d'où des profils de variabilité distincts. On se penche ici sur un sous-problème de l'identification d'EPV, à savoir l'identification d'occurrences d'EPV vues dans d'autres contextes, quelque soit leur forme de surface, ce qui nécessite de prendre en compte l'ambiguïté pour éviter des lectures littérales (*casser sa vieille pipe*) ou des co-occurrences fortuites (*casser le tuyau de sa pipe*).

On considère pour cela deux approches : la première se fonde sur une mesure de la variabilité des EPV indépendante de la langue. La seconde consiste à modéliser le problème comme une tâche de classification d'après des traits pertinents pour la variabilité morpho-syntaxique des EPV, ce qui nous a conduit à développer un système (VarIDE), qui a participé à la compétition PARSEME d'identification automatique d'EPV en 2018.

Mots clés : Expression polylexicale, variabilité, ambiguïté, traitement automatique des langues

Abstract

Automatic identification of multiword expressions (MWEs) is a pre-requisite for many natural language processing applications. This task is challenging because MWEs, especially verbal ones (VMWEs) like *to* ***kick the bucket*** (which means *to die*), exhibit surface variability (*no* ***buckets*** *were* ***kicked***). However, compared with regular constructions, this variability is usually more restricted (e.g. some nouns cannot be modified by an adjective), hence various variability profiles. We address here a subproblem of VMWE identification, namely the identification of occurrences of VMWEs previously seen in corpora, whatever their surface form, which requires to take ambiguity into account to avoid literal (*he kicked the old bucket*) or coincidental occurrences (*he kicked the ball and the bucket fell down*).

To this end, we considered two main approaches : The first one is based on a language-independent measure of VMWE variability. The second one consists in modeling the problem as a classification task on the basis of features relevant to the VMWE morpho-syntactic variability, which led to a system (VarIDE) that participated in the PARSEME shared task on automatic identification of VMWEs in 2018.

Keywords : Multiword expression, MWE, variability, ambiguity, natural language processing

Table des matières

Introduction **25**

I *Appelons un chat un chat* : les expressions polylexicales du point de vue de la linguistique et de leurs enjeux **29**

1 Définition **33**
1.1 Lexicalisation des composants 33
1.2 Comportement idiosyncratique des EP 34

2 Hétérogénéité des EP **39**
2.1 Consensus et divergences 39
2.1.1 Mots composés 39
2.1.2 Termes polylexicaux 40
2.1.3 Expressions idiomatiques ou idiomes 41
2.1.4 Collocations 41
2.1.5 Verbes à particules ou *verb-particle constructions* (VPC) 42
2.1.6 Entités nommées polylexicales (ENP) 43
2.1.7 Constructions à verbe support (CVS) 44
2.2 Fréquence des EP 44
2.3 Classer les EP : ***la quadrature du cercle*** ? 45

3 Variabilité des EP **49**
3.1 Type vs. token 49
3.2 Variabilité imprévisible des EP 50
3.2.1 Variabilité et maintien de lecture idiomatique 51
3.2.2 Variabilité diachronique et diatopique 51
3.3 Continuum de variabilité des EP 54
3.4 Niveaux de variabilité des EP 55

4 Focalisation sur l'ambiguïté : EP vs. non-EP **57**
4.1 Différents types d'ambiguïté 57
4.2 Désambiguïsation 58

II *Du pain sur la planche* en matière de traitement automatique des EP **61**

5 Modélisation et traitement automatique de la variabilité des EP **65**
5.1 Lexique 65
5.1.1 Dictionnaires 65
5.1.2 Tables du Lexique-Grammaire 66
5.2 Découverte 67
5.2.1 Définition 67
5.2.2 Méthodes générales de découverte 68
5.2.3 La variabilité au service de la découverte 71
5.2.4 Évaluation 73
5.3 Méthodes transférables à l'identification d'EP 74

6 Identification d'EP **75**
6.1 État de l'art sur la tâche d'identification 75
6.1.1 Définition 75
6.1.2 Méthodes d'identification 76
6.1.3 Évaluation 89
6.2 Lien entre l'identification et des tâches annexes 90
6.2.1 Parsing 91
6.2.2 Traduction 95

7 Identification de variantes d'EP **101**
7.1 Motivation 101
7.2 EP de référence vs. variantes : ***l'œuf ou la poule ?*** 103
7.2.1 Variantes d'entités nommées (EN) 103
7.2.2 Variantes de termes 104
7.3 Variantes d'EP verbales : notre définition 106
7.4 Esquisse de méthode : la variabilité au service de l'identification de variantes 109

8 Corpus utilisés **111**
8.1 Corpus PARSEME 111
8.1.1 Processus d'annotation 111

8.1.2 Catégories d'EP utilisées . . . 113
8.1.3 Corpus PARSEME du français . . . 114
8.2 Corpus `CoNLL17` et `WebSample` . . . 116

III Où la variabilité des EP *a voix au chapitre* 119

9 Variabilité observée en corpus 123
9.1 Variabilité morphologique des composants d'EP de patron VERB-(DET)-NOUN 123
9.1.1 Variation verbale . . . 123
9.1.2 Variation nominale . . . 124
9.2 Discontinuités . . . 125
9.2.1 Discontinuités par catégories d'EP . . . 126
9.2.2 Variation syntaxique *via* les discontinuités . . . 127
9.3 Distance syntaxique . . . 130

10 Modélisation de la variabilité 133
10.1 Profil de variabilité multidimensionnel . . . 133
10.2 Limites de la définition de profil . . . 135
10.2.1 Qualité du corpus source . . . 135
10.2.2 Qualité d'annotation d'EP . . . 135
10.2.3 Validité de la *LemmNorm* pour la fusion de types d'EP . . . 135
10.2.4 Représentativité limitée du profil de variabilité . . . 136
10.3 Représentation graphique du profil de variabilité . . . 136
10.3.1 Choix des traits . . . 136
10.3.2 Représentation graphique . . . 137
10.4 Correspondance transformation-trait(s) . . . 138

11 Quantification de la variabilité des EP 143
11.1 Niveaux de variabilité d'EP VERB-(DET)-NOUN . . . 143
11.1.1 Variabilité des EP d'après leurs dépendances . . . 143
11.1.2 Comparaison avec Tutin (2016) . . . 144
11.2 Score de variabilité d'EP VERB-(DET)-NOUN . . . 145
11.2.1 Mesure de similarité et de variabilité . . . 146
11.2.2 Validation de la méthode . . . 148
11.2.3 Utilisation de la méthode : discrimination $LVC_{1.0}$ vs. $ID_{1.0}$ . . . 149
11.2.4 Limites et biais . . . 150
11.3 Bilan . . . 151
11.3.1 Retour sur les hypothèses . . . 151

11.3.2 Variabilité des EP et performance des systèmes 151

IV Vers une identification multilingue de variantes d'EP *pour faire avancer le schmilblick* 155

12 VarIDE : *un ballon d'essai* 159

12.1 Hypothèses 159
12.2 Extraction de candidats (`ExtractCands`) 160
12.2.1 Normalisation et génération de patrons 160
12.2.2 Extraction de candidats positifs ('EP') et négatifs ('non-EP') 162
12.3 Choix des traits pour la classification 163
12.3.1 Traits adaptables par langue 163
12.3.2 Traits absolus (ABS) vs. relatifs (REL) 164
12.3.3 Classification d'EP et attribution de catégories 168
12.4 Évaluation de VarIDE 168
12.5 Bilan 172
12.5.1 Nature des données 172
12.5.2 Remarques sur la phase de `Classif` 173
12.5.3 Pertinence des traits 174
12.5.4 Disponibilité de VarIDE 175

13 VarIDE+ avec sélection automatique de traits : *comment séparer le bon grain de l'ivraie ?* 177

13.1 Extraction de candidats (`ExtractCands2`) 178
13.2 Méthodologie de sélection de traits 180
13.2.1 Motivation 180
13.2.2 Suppression de traits invalides : `ActiveFeat` 182
13.2.3 Classement des traits : `RankFeat` 183
13.2.4 Optimisation du choix de traits : `SelectFeat` + `Classif` 184
13.3 Résultats 184
13.3.1 Ensembles optimum de traits 184
13.3.2 Évaluation comparative 186
13.3.3 Évaluation manuelle sur un corpus externe 187
13.3.4 *Bootstrap* 189
13.4 Bilan 189

14 Perspectives 193

Conclusion **195**

Annexe **201**

A Liste d'EP de `WebSample` **201**

Index **217**

Liste des tableaux

5.1 Extrait du Lexique-grammaire pour les EP ***faire équipe*** et ***faire joujou***. . 67

5.2 Comparaison entre les EP verbales candidates découvertes automatiquement et un lexique de référence. Le candidat *mettre en service* n'est pas, de notre point de vue, une EP selon l'argumentation développée dans la section 5.1.1. 74

6.1 Comparaison de l'identification automatique d'EP dans la phrase 6.10 par trois systèmes par rapport à une annotation de référence. Les indices numériques et alphabétiques des colonnes d'annotation signalent les composants d'une même EP. L'annotation du troisième système est illustrée par l'exemple 6.10. 90

7.1 Différences entre les conceptions d'une *variante* dans l'état de l'art telle que définie par (a) (Jacquemin, 2001), (b) (Daille, 2017), (c) pour les *variant-of train* durant la compétition PARSEME 2018 et celles que nous avons utilisées lors de l'identification de variantes d'EP verbales de patron VERB-(DET)-NOUN dans (d) (Pasquer *et al.*, 2018c) et (e) pour tout patron par notre système VarIDE (Pasquer *et al.*, 2018a). 107

8.1 Correspondances approximatives entre les catégories des éditions 1.0 et 1.1 . 115

8.2 Statistiques du corpus français de l'édition PARSEME 1.0 : nombre de phrases, de tokens et distribution des EP par catégorie. 116

8.3 Statistiques du corpus français de l'édition PARSEME 1.1 : nombre de phrases, de tokens et distribution des EP par catégorie. 116

8.4 `WebSample` par rapport à `FR-train1.1` (EP vues au moins deux fois) 117

9.1 EP de patron VERB – (DET) – NOUN les plus fréquentes dans `FR-train1.0` 125

9.2 Nature des discontinuités de longueur 1. Les pourcentages les plus significatifs sont mis en gras. 128

9.3 Quelques aspects de la variabilité et de figement des $LVC_{1.0}$ vs. $ID_{1.0}$ dans `FR-train1.0` 129

10.1 Profil (sommaire) de variabilité de l'EP ***jeter l'éponge*** vu comme une combinaison de traits. 134

11.1 Comparaison des degrés de variabilité observés en corpus avec ceux établis par Tutin (2016). Les cellules grisées indiquent des invariabilités. Le niveau 0 correspond à la variabilité minimale et le niveau 4 à la variabilité maximale. 145

11.2 Distribution des EP extraites du corpus `FR-train1.0` parmi les classes de Tutin (2016) 149

12.1 Exemples d'EP, de leur *LemmNorm*, des séquences de POS observées et de leur *POSnorm*. Les patrons d'extraction autorisés (d'après ces exemples) sont : ⟨NOUN ;VERB⟩⇒ {NOUN − VERB, VERB − NOUN}, ⟨PRON ;VERB⟩⇒ {PRON − VERB} 163

12.2 Aperçu de traits ABSolus et RELatifs pour les exemples 12.2 (RELatif à 12.1), 12.3 et 12.4. Les traits relatifs portant sur la *LemmNorm* et la catégorie sont indiqués comme n'étant pas applicables car, par définition, ils seraient systématiquement *vrai*. 167

12.3 Pour chaque famille de langues (romanes, germaniques, slaves et autres) : performances d'identification d'EP de VarIDE sur l'ensemble des EP (vues + non-vues), sur les *variant-of-train*, proportion de variantes dans le `test`, résultat d'extraction de candidats (quantité par étiquette 'EP' vs. 'non-EP', proportion d"EP'), rappel avant et après classification pour les *variant-of-train*. Les trois dernières colonnes fournissent des détails sur les spécificités des corpus (présence d'étiquettes au format UD, relations de dépendances syntaxiques disponibles) ainsi que le recours (ou non) à *Filtre20* pour limiter la longueur des discontinuités, ce choix étant effectué d'après les performances observées sur le `dev`. 169

12.4 Évaluation de l'impact de l'ajout d'une phase de classification `Classif` après la phase d'extraction de candidats `ExtractCands` pour chaque langue : on précise la F-mesure – sur les *seen-in-train* et les *variant-of-train* – obtenue après chacune de ces étapes et les deux dernières colonnes mettent en évidence la différence $F_{\texttt{Classif}}$ - $F_{\texttt{ExtractCands}}$. 174

13.1 Performance de `ExtractCands2`, en se fondant sur les EP attestées vues au moins deux fois dans `FR-train1.1`. La F-mesure sur le corpus `FR-test1.1` serait alors de 73. 180

13.2 Traits pertinents pour l'identification d'EP et performances pour les EP préalablement vues pour notre systèmes vs. les systèmes de la compétition PARSEME. (Ramisch *et al.*, 2018) : (a) (Waszczuk, 2018), (b) (Taslimipoor et Rohanian, 2018), (c) (Boros et Burtica, 2018), (d) (Stodden *et al.*, 2018), (e) (Moreau *et al.*, 2018) (CRF-DepTree-categs system), (f) (Moreau *et al.*, 2018) (CRF-Seq-nocategs system), (g) (Berk *et al.*, 2018), (h) (Ehren *et al.*, 2018), (i) (Pasquer *et al.*, 2018a), (j) (Zampieri *et al.*, 2018) 181

13.3 Meilleure performance moyenne sur 10% de `FR-train1.1` (validation croisée) : P, R et F sont les scores moyens obtenus sur les 10 partitions, et σ la variance correspondante. Les ensembles de traits pris en compte – dont le nombre figure dans la colonne `RankFeat` – sont représentés sous la forme de chiffres entourés d'un cercle dans la colonne `SelectFeat`. Ils sont détaillés dans la Table 13.4. 185

13.4 Ensembles de traits optimum mentionnés dans la Table 13.3 186

13.5 Résultats d'évaluation dans la configuration d'évaluation de la compétition PARSEME, et avec un corpus externe manuellement annoté 188

13.6 Résultats de VarIDE+ par catégorie sur `WebSample` (EP vues au moins deux fois) 188

A.1 Types d'EP des catégories $\text{IRV}_{1.1}$, $\text{LVC}_{1.1}$ et $\text{VID}_{1.1}$ utilisés pour constituer le corpus `WebSample`. 201

A.2 Types d'EP de la catégorie $\text{MVC}_{1.1}$ utilisés pour constituer le corpus `WebSample`. 202

Table des figures

2.1 Aperçu de la fréquence et de l'hétérogénéité des EP dans un extrait de roman (Leblanc, 1907). Les indices numériques signalent les composants d'une même EP. 45

2.2 Illustration de la loi de Zipf : la majorité (resp. la minorité) des types d'EP sont rares (resp. fréquents) dans le corpus `FR-train1.1`. 46

6.1 Exemples de données linéairement séparables (en haut) et non-linéairement séparables (en bas). 83

6.2 Exemple de transformation de données non linéairement séparables (en haut) en données linéairement séparables (en bas) : la dimension de l'ordonnée ne permettant pas de différencier les deux classes est redéfinie en tant que fonction quadratique. 84

6.3 Exemple d'arbre de décision : deux tests permettent l'attribution d'une étiquette de classe. Comme chaque arbre se subdivise en deux sous-arbres, il s'agit d'un arbre binaire. 85

6.4 Exemple de phrase parsée via UDPipe (Straka et Straková, 2017). 92

6.5 Représentation sous la forme d'arbre syntaxique de la phrase ***Il pleut (des cordes)*** *de Paris à Marseille*, obtenue à partir de UDPipe (Straka et Straková, 2017) 93

6.6 Représentation sous la forme d'arbre syntaxique de dépendances de la phrase *Il* ***prend*** *également, et cette fois-ci seul et sans vote de l'assemblée, des* ***décisions*** *réglementaires.* Représentation obtenue à partir de UDPipe (Straka et Straková, 2017) 94

6.7 Connexion syntaxique entre le verbe *jouer* et le nom *rôle* dans (a) mais pas dans (b) 94

6.8 Alignement entre la version originale française d'un extrait de roman (à gauche) et sa traduction anglaise (à droite). 98

7.1 Nombre de variantes observées dans le corpus `FR-train1.1` en fonction de la fréquence des types d'EP. 102

8.1 Exemple d'annotation sur la plate-forme FLAT : l'annotateur humain sélectionne les composants de chaque EP (ici en surbrillance) et leur attribue une étiquette en fonction de leur catégorie (LVC.full, ID, etc.). 113

9.1 Fréquence de la longueur des discontinuités par catégorie d'EP dans `FR-train1.0` 126
9.2 Discontinuités (sous forme de séquences de POS) les plus fréquentes dans `FR-train1.0` 127
9.3 Distance syntaxique d'EP avec une valeur de 1 entre éléments surlignés : en série (à gauche) ou en parallèle (à droite). 131
9.4 Distance syntaxique avec une valeur de 2 entre éléments surlignés : en série (à gauche) pour une EP ou en parallèle (à droite) pour une co-occurrence fortuite. 131

10.1 Exemple de dépendance entrante (*jouer*) et sortante (*rôles*) à partir du nom *multitude*. 133
10.2 Profil de variabilité de l'EP ***prendre l'habit*** (en vert) par rapport aux lectures littérales (en bleu) et co-occurrences fortuites (en noir) dissociées ou considérées conjointement comme des non-EP (en rouge). Le profil de couleur verte est d'autant plus réduit que l'EP est figée. 139
10.3 Profil de variabilité de l'EP ***prendre décision*** (en vert) par rapport aux co-occurrences fortuites (en rouge). 140

11.1 Deux exemples de l'EP ***prendre*** *une* ***décision*** étiquetés en POS et avec parsing en dépendances. 146
11.2 Boxplots de Tukey de V^L (a) et V^S (b) (en ordonnée) en fonction des niveaux de Tutin (en abscisse). 150
11.3 Boxplots de Tukey de V^L (en ordonnée) en fonction des catégories d'EP (en abscisse) : $\text{ID}_{1.0}$ et $\text{LVC}_{1.0}$ 150
11.4 F-mesure des 15 systèmes de la compétition PARSEME 1.1 pour l'identification d'EP déjà vues selon que l'on considère des variantes identiques à celles vues (*identical-to-train*) ou différentes (*variant-of-train*). 152

12.1 Performances des 15 systèmes ayant participé à la ST pour le français avec une focalisation sur les données vues (*seen-in-train*) selon qu'il s'agit de *identical-to-train* ou de *variant-of-train*. 172
12.2 Démonstrateur mis au point par le laboratoire ATILF : exemple d'identification par notre système VarIDE. Les éléments lexicalisés des EP ***prendre décision*** et ***il convenir*** sont mis en évidence. 176

13.1 Candidats d'EP ayant une chaîne de dépendance discontinue : (a) extrait, (b) non-extrait. 179

Liste des conventions et abréviations

Conventions typographiques :

- ***escampette*** : élément lexicalisé (dit *composant*) d'une expression polylexicale (EP)
- Forme canonique d'une EP : `manger les pissenlits par la racine`
- Forme lemmatisée normalisée d'une EP : ⟨le ;le ;manger ;pissenlit ;racine⟩
- Lecture idiomatique : *Il* ***jette l'éponge*** *et divorce.* ⇒ 'Il abandonne et divorce'.
- Lecture idiomatique avec traduction littérale : *He* ***kicked the bucket*** *'Il a frappé le seau'*⇒'Il est mort'.
- Lecture littérale : *Il jette l'éponge dans l'évier.*
- Co-occurrence fortuite : *Il jette les gants usagés et nettoie l'éponge.*
- Parties de discours et relations de dépendances entrantes au format *Universal Dependencies* : dans l'exemple $Il_{\text{PRON.NSUBJ}}$ $\boldsymbol{jette}_{\text{VERB.ROOT}}$ $\boldsymbol{l}_{\text{DET.DET}}$' $\boldsymbol{éponge}_{\text{NOUN.NOBJ}}$, *éponge* est un nom (NOUN) et le complément d'objet (NOBJ) du verbe.

Jugements de grammaticalité :

- Exemple agrammatical : **Il faut que je viens*
- Exemple grammaticalement correct mais généralement non acceptable : ? ?*Il pleut des chiens et des chats*
- Phrase dont le glissement de sens va au-delà de ce qui est attendu de la modification formelle : # *Il jette une éponge* vs. *Il* ***jette l'éponge***

Langues utilisées dans les exemples :

- `BE` français de Belgique
- `CA` français du Canada
- `CN` chinois
- `DE` allemand
- `EN` anglais
- `ES` espagnol
- `FR` français de France
- `IT` italien
- `JP` japonais
- `LA` latin
- `PL` polonais
- `TR` turc

Autres abréviations :
— ABS : trait absolu
— CVS : construction à verbe support
— EN(P) : entité nommée (polylexicale)
— EP : expression polylexicale
— $ID_{1.0}$: idiome verbal
— $IReflV_{1.0}$, $IRV_{1.1}$: *inherently reflexive verb* (verbe intrinsèquement réflexif)
— *LemmNorm* : forme lemmatisée normalisée
— $LVC_{1.0/1.1}$: *light verb construction* (construction à verbe support)
— $MVC_{1.1}$: *multi-verb construction* (construction multi-verbes)
— NB : Naïve Bayes
— POS : *part of speech* (partie de discours)
— *POSnorm* : séquence de POS normalisée
— *POSseq* : séquence de POS observée
— REL : trait relatif
— ST : compétition PARSEME 1.1 sur l'identification d'expressions polylexicales verbales
— SVM : *support vector machine* (séparateur à vaste marge)
— TAL : traitement automatique du langage
— $VID_{1.1}$: idiome verbal
— $VPC_{1.1}$: *verb-particle construction* (verbe à particule)

Liste des expressions polylexicales

Aller dans le bon sens ⇒ 'Évoluer positivement'

Appeler un chat un chat ⇒ 'Ne pas avoir peur d'appeler les choses par leur nom'

Apporter des oranges ⇒ 'Rendre visite à quelqu'un en prison ou à l'hôpital'

Arrête ton char ⇒ 'Arrête de dire n'importe quoi'

Autant/au temps pour moi ⇒ 'Admettre avoir commis une erreur'

Avoir du pain sur la planche ⇒ 'Avoir beaucoup de travail'

Avoir un chat dans la gorge ⇒ 'Être enroué'

Avoir un petit pois à la place du cerveau ⇒ 'Être peu intelligent'

Avoir voix au chapitre ⇒ 'Avoir le droit de s'exprimer / d'influencer'

Ballon d'essai ⇒ 'Premier jet d'une chose rendue publique'

Boire le bouillon de onze heures ⇒ 'Mourir'

Casser sa pipe ⇒ 'Mourir'

C'est de l'hébreu/du chinois ⇒ 'C'est complètement incompréhensible'

Casser du sucre sur le dos ⇒ 'Dire du mal de quelqu'un en son absence'

Cerise sur le gâteau ⇒ 'L'avantage supplémentaire'

Coiffer Sainte Catherine ⇒ 'Le 25 novembre, les femmes portent un chapeau jaune et vert si à 25 ans elles sont encore célibataires'

Courir la prétentaine ⇒ 'Vagabonder'

Crier haro sur le baudet ⇒ 'Exprimer sa révolte envers un individu ou quelque chose'

En avoir le cœur net. ⇒ 'Savoir à quoi s'en tenir'

En rester comme deux ronds de flan ⇒ 'Être stupéfait'

Enfoncer une porte ouverte ⇒ 'Énoncer une banalité'

Être (heureux) comme un poisson dans l'eau ⇒ 'Être heureux'

Être comme le poisson hors de l'eau ⇒ 'Être malheureux'

Faire avancer le schmilblick ⇒ 'Faire avancer un sujet'

Faire face ⇒ 'Affronter'

Filer le *parfait* ***amour*** ⇒ 'Vivre un amour sans histoire'

Filer un *mauvais* ***coton*** ⇒ 'Faire de mauvaises affaires'

Honni soit qui mal y pense ⇒ 'Honte à celui qui y voit du mal'

Il pleut des cordes ⇒ 'Il pleut très fort'

Jeter l'éponge ⇒ 'Abandonner'

L'œuf ou la poule ? ⇒ 'Lequel des deux est la cause de l'autre ?'

La quadrature du cercle ⇒ 'Un projet irréalisable'

Manger les pissenlits par la racine ⇒ 'Être mort et enterré'

Ne pas en croire ses oreilles ⇒ 'Être surpris par des propos'

On aura tout vu ⇒ 'Avoir vu quelque chose d'original'

Passez-moi la rhubarbe, je vous passerai le séné ⇒ 'Se faire des concessions mutuelles'

Perdre la raison/la tête/la boule/le nord/les pédales etc. ⇒ 'Être désorienté'

Poser un lapin ⇒ 'Ne pas venir au rendez-vous'

Prendre acte ⇒ 'Retenir formellement une information'

Prendre de court ⇒ 'Prendre par surprise'

Prendre la poudre d'escampette ⇒ 'S'enfuir'

Prendre la tête ⇒ 'Se retrouver en première position d'une compétition non encore terminée / Faire perdre patience'

Prendre le taureau par les cornes ⇒ 'S'attaquer à une difficulté avec détermination'

Prendre ses jambes à son cou ⇒ 'Se sauver'

Quand on parle du loup, (on en voit la queue) ⇒ 'Une personne arrive au moment où on en parle'

Retourner au charbon ⇒ 'Reprendre une activité désagréable'

Revenons à nos moutons ⇒ 'Revenons à notre sujet'

Se faire du mouron ⇒ 'S'inquiéter'

Séparer le (bon) grain de l'ivraie ⇒ 'Séparer le mal et le bien, les gentils et les méchants'

Se taper des barres ⇒ 'Rire'

Sucrer les fraises ⇒ 'Être gâteux'

Tirer sa révérence ⇒ 'S'en aller'

Tirer son épingle du jeu ⇒ 'Se dégager adroitement d'une situation délicate'

Tourner la page ⇒ 'Commencer une nouvelle étape d'une vie en oubliant le passé'

Vider son sac ⇒ 'Avouer'

Introduction

Des modules de prédiction de mots sur les téléphones portables aux enceintes connectées, l'informatique et le langage se rejoignent dans de nombreux domaines de la vie quotidienne. Le Traitement Automatique du Langage (TAL) se situe à la croisée de ces domaines et cette discipline doit répondre à de nombreux défis, parmi lesquels figure notamment la prise en charge des expressions polylexicales (*multiword expressions* en anglais) désormais abrégées sous l'acronyme EP. Ces EP sont au cœur du réseau PARSEME[1] (*PARSing and Multi-word Expressions*) qui réunit des experts interdisciplinaires (linguistes, psycholinguistes, informaticiens, linguistes informaticiens, industriels) de 31 pays dans le but d'accroître la qualité et d'améliorer la prise en compte des EP dans les systèmes de TAL. Cette thèse est d'ailleurs financée par l'Agence Nationale pour la Recherche dans le cadre de la spin-off française (PARSEME-FR) du réseau PARSEME.

Les EP se distinguent par un ensemble de propriétés sur lesquelles nous reviendrons en détail dans la partie I mais dont nous pouvons d'ores et déjà esquisser les contours. Considérons la phrase *Alea jacta est, le manuscrit est fin prêt en vue de la soutenance de thèse*, elle contient à elle-seule quatre EP[2]. La difficulté d'accès au sens (opacité sémantique) s'observe fréquemment dans les EP et cette propriété est particulièrement flagrante dans le cas de la locution latine ***alea jacta est***. Néanmoins, même lorsque le sens de chaque mot d'une EP telle que ***poser un lapin*** est connu, il arrive que le sens global de cette EP n'entretienne aucune relation évidente avec le sens isolé de chacun des mots la composant : dans cette expression il n'est nullement question de *lapin*. De même, alors qu'en français, on ***soutient*** *une* ***thèse*** (de doctorat), en anglais on la *défend* (`(EN)` ***thesis defense***).

Outre cette dimension arbitraire, les EP sont dotées d'un ensemble de spécificités qui posent des problèmes importants dans le cadre d'applications de TAL, telles que la traduction automatique, car une traduction mot-à-mot est rarement pertinente (*Il a* ***posé un lapin*** *à son amie* $\neq$ `(EN)` *He put a rabbit to his friend*). Pour ne rien arranger, ces EP sont extrêmement fréquentes (Jackendoff, 1997), ce qui explique les initiatives (dont PARSEME) visant à améliorer la connaissance sur ces expressions. Sag *et al.* (2002) qualifiaient les EP de ***pain in the neck*** '*douleur dans la nuque*'⇒'**casse-pieds**', et c'est toujours le cas en 2019. Depuis quinze ans en effet, un atelier MWE (*MultiWord Expression Workshop*) leur est intégralement dédié sous le parrainage de SIGLEX (*Special Interest Group on the LEXicon*) de l'*Association for Computational Linguistics* (ACL). Cela témoigne à la fois de l'importance des EP et des défis complexes qui leur sont associés.

1. `https://typo.uni-konstanz.de/parseme/`
2. ***Alea jacta est, fin prêt, en vue de, soutenance de thèse.***

Bien que les EP puissent être perçues comme sources de problèmes pour les applications de TAL, nous proposons au contraire de tirer bénéfice de leurs spécificités pour mieux les identifier dans des corpus textuels. L'hypothèse sous-jacente est qu'au-delà de leur diversité apparente, les EP partagent des caractéristiques et comportements utiles pour leur identification. Savoir qu'une séquence donnée est une EP permet ensuite de lui appliquer le traitement approprié, comme la recherche d'équivalents dans la langue cible au lieu d'une traduction littérale (*Il a* ***posé un lapin*** *à son amie* = `(EN)` *He* ***stood*** *his friend* ***up***). Une autre dimension à prendre en compte pour garantir une identification correcte des EP est la gestion de leur ambiguïté. Comme l'illustre l'exemple 1, sa résolution permettrait là encore d'améliorer les performances de traduction automatique, car la seule séquence *il y a* bénéficie de trois traductions différentes. Un repérage préalable des EP faciliterait par conséquent le choix de la traduction la plus adéquate.

(1) `(FR)` ***Il y a***$_{EP\_verbale}$ *un festival en juin, il y a*$_{non\ EP}$ *déjà participé* ***il y a***$_{EP\_adverbiale}$ *cinq ans* $\Longrightarrow$ `(EN)` There is a festival on June, he already participated five years ago.

Dans cette thèse, nous nous concentrons exclusivement sur l'identification d'EP connues (préalablement *vues* dans d'autres contextes) et écartons de ce fait les EP inconnues (dites *non vues*). Ce choix est motivé par le constat qu'il est généralement difficile, voire impossible, de tirer profit des caractéristiques d'EP vues pour acquérir des connaissances sur des EP non vues. Cette difficulté est d'ailleurs attestée par les faibles performances obtenues sur ces EP non vues lors de la compétition PARSEME de 2018 [3] pour l'identification d'EP verbales (Ramisch *et al.*, 2018). Notre approche consiste donc à optimiser, dans un premier temps, l'identification des EP vues. Quoique cette tâche semble *a priori* simple, il n'en est rien car les EP verbales apparaissent rarement sous une forme de surface unique (p. ex. flexion temporelle, passivation, etc.). Considérons l'EP ***casser sa pipe*** (dont la lecture idiomatique signifie *mourir*), le verbe peut être fléchi à différents temps mais cette EP présente des contraintes : l'association d'une forme au passé et de la première personne du singulier oriente vers une lecture littérale [4], comme dans l'extrait de roman (Ex. 2) rappelant l'origine de cette expression.

(2) *Quand on met la pipe entre les dents de l'opéré, il serre tant qu'il peut. S'il meurt sa pipe tombe et se casse. Moi, j'ai cassé ma pipe et je suis toujours vivant.* (Schlogel, 1992)

Cet exemple témoigne également du fait que la seule présence des éléments d'une EP dans une même phrase peut être purement fortuite [5] (*sa pipe tombe et se casse*). La gestion des différentes ambiguïtés représente donc un premier défi à relever pour favoriser une identification fiable d'EP verbales connues. A cela s'ajoute le fait que les systèmes d'identification automatique rencontrent davantage de difficultés pour les EP apparaissant sous une forme inédite, autrement dit jamais vue au préalable, ce qui occasionne des baisses de performance allant de 15 à 20 points de F-mesure [6] selon les systèmes soumis à la compétition PARSEME de 2018. Puisque la variabilité morpho-syntaxique des EP verbales est au

3. F-mesure en français de 18,87% pour le meilleur système.
4. Matérialisée par un soulignement sous forme de vagues.
5. Matérialisée par un soulignement sous forme de tirets.
6. Cette métrique est explicitée à la page 73.

cœur de ce problème, c'est à ce phénomène spécifique que nous nous intéresserons, notre objectif étant d'améliorer l'identification globale des EP en maximisant les performances sur les EP vues.

Pour cela, notre thèse s'appuie sur les hypothèses principales suivantes :

H1 Chaque EP a un profil de variabilité (c.-à-d. un ensemble de transformations autorisées) qui lui est propre et qui est attesté par un éventail plus ou moins étendu de réalisations (*tokens*). A l'inverse, des séquences ressemblantes mais ne relevant pas du statut d'EP ne respectent pas ce profil. Par exemple, l'EP ***jeter l'éponge*** ⇒ 'abandonner' ne conserve pas son statut d'EP en cas de modification du nom par un adjectif (*jeter l'éponge verte*), alors que cette transformation n'a aucune incidence pour la séquence *jeter la serpillière (verte).*

H2 On peut modéliser ce profil de variabilité comme un phénomène multidimensionnel d'après les informations morphosyntaxiques fournies par des corpus annotés, par exemple le type de discontinuités, l'existence de modifieurs ou bien encore la variation interne des composants.

H3 Connaître le profil de variabilité de chaque EP peut aider à identifier les variantes d'une EP apparaissant dans un nouveau texte. Par exemple, nous considérerons la séquence ***jeter l'éponge*** comme étant une EP mais pas *une éponge jetée, jeter des éponges, etc.* Il serait donc possible d'utiliser cette modélisation pour développer un système d'identification d'EP centré sur les variantes.

Nous nous focalisons sur l'identification de variantes d'EP connues, d'après un processus se déroulant en deux phases. Dans un premier temps, nous recherchons des EP potentielles dans un corpus. Un filtrage est ensuite effectué sur la base du profil de variabilité afin que seules les candidates compatibles avec ce profil soient étiquetées 'EP'. Les contributions escomptées se situent sur deux plans interdépendants : d'une part, l'analyse du phénomène de variabilité d'EP verbales devrait permettre de proposer un système d'identification de variantes avec une extension multilingue. D'autre part, les atouts et défauts d'un tel système sont également intéressants pour affiner notre modélisation de la variabilité.

Ce mémoire s'organise selon 14 chapitres répartis sur 4 parties, chacune introduisant sous la forme d'un chapeau les chapitres qui lui sont relatifs. Dans la partie I, nous établissons un compte-rendu de l'état de l'art sur les propriétés linguistiques des EP, la terminologie associée et les enjeux majeurs. Dans la partie II, nous nous intéressons aux méthodes mises en œuvre pour la gestion des EP par les applications de TAL. La partie III porte sur la modélisation de la variabilité. La partie IV est dédiée aux systèmes d'identification de variantes développés durant cette thèse.

Première partie

Appelons un chat un chat : les expressions polylexicales du point de vue de la linguistique et de leurs enjeux

La communication humaine vise à transmettre de l'information entre un émetteur (qui encode le message) et un récepteur (qui le décode). Ce message peut être représenté sous la forme d'un agencement spécifique de mots d'après les règles morphosyntaxiques propres à chaque langue. La teneur du message, autrement dit son *sens* découle, d'après Dubois *et al.* (1997), à la fois du *contexte verbal*, fourni par les éléments situés dans son entourage, et du *contexte situationnel*, lié aux informations partagées entre l'émetteur et le récepteur, et qui peuvent influer sur la construction du sens. La phrase (DE) ***Ich bin ein Berliner*** ⇒ 'je suis un Berlinois' acquiert ainsi un sens inédit lorsque ce n'est pas un Allemand, mais l'Américain John Kennedy qui la prononce comme marque de soutien aux Ouest-Allemands.

Dans certains cas, pour accéder au sens, il ne faut pas considérer chaque mot individuellement mais prendre en compte des groupes de mots. On parle alors d'expressions polylexicales (EP). Dans cette partie, nous présentons leurs caractéristiques (chapitre 1), différentes approches de classification (chapitre 2), notamment certaines fondées sur leur variabilité (chapitre 3), une facette des EP qui constitue le fil conducteur de cette thèse puisque nous proposons d'en tirer profit pour distinguer les EP de séquences ressemblantes mais qui n'en sont pas (chapitre 4).

Chapitre 1

Définition

La définition des EP adoptée dans ce travail repose sur des caractéristiques à la fois formelles, en lien avec la lexicalisation de leurs composants (section 1.1), et comportementales liées à des idiosyncrasies par rapport à des combinaisons ordinaires de lexèmes (section 1.2).

1.1 Lexicalisation des composants

Les EP (parfois nommées *expressions multi-mots* dans la littérature) sont des unités de sens généralement constituées d'au moins deux unités graphiques à la fois requises et non substituables. Compte-tenu des divergences terminologiques sur la notion de *mot*, nous qualifions de *tokens* les unités graphiques, c.-à-d. toute séquence de caractères que l'on peut distinguer d'une autre séquence (on compte par exemple trois tokens dans la phrase *voici trois tokens*). De ce fait, on restreint l'usage de *mot* aux unités de sens (ou *lexèmes*). Mots et tokens coïncident parfois (p. ex. *trois*), mais un token peut aussi être constitué de plusieurs mots : *du* = *de* + *le*.

 [1] Les unités d'une EP sont qualifiées d'*éléments lexicalisés* (ou *composants*) et elles apparaissent en caractères gras dans les exemples donnés dans ce mémoire. Cette notion de lexicalisation des composants s'entend, à l'instar de Savary *et al.* (2018), comme suit :

> Note that lexicalisation is traditionally defined as a diachronic process by which a word or a phrase acquires the status of an autonomous lexical unit, that is, "a form which it could not have if it had arisen by the application of productive rules" (Bauer et Laurie, 1983; Lipka *et al.*, 2004). In this sense all expressions considered VMWEs [verbal multiword expressions] in this work are lexicalized. Our notion of lexicalisation extends this standard terminology, as it applies not only to VMWEs but to their components as well. [2]

1. Ce pictogramme signale les définitions utilisées dans cette thèse.

2. Il faut noter que la lexicalisation est traditionnellement définie comme un processus diachronique dans lequel un mot ou une phrase acquiert le statut d'unité lexicale autonome, c'est-à-dire "une forme qu'il n'aurait pas dû présenter s'il avait émergé de l'application de règles de productivité" (Bauer et Laurie, 1983; Lipka *et al.*, 2004). En ce sens, toutes les expressions considérées comme EP verbales dans ce travail

En français, les composants d'une EP peuvent être séparés par une espace, un tiret ou une apostrophe comme dans ***aller et venir***, ***contre-braquer*** ou ***s'entr'aimer***. Ils peuvent également être soudés (***contrebraquer*** et ***s'entraimer***, après la réforme orthographique de 1990). La soudure demeure marginale en français par comparaison avec des langues agglutinantes (japonais, basque, etc.) ou des langues dont le système d'écriture n'utilise pas d'espaces pour séparer les mots (p. ex. le chinois). La soudure estompe les limites entre mots mais un locuteur natif peut les rétablir sans difficulté (***tirebouchon*** → ***tire-bouchon***) sauf lorsqu'il ne connaît pas les composants (***boutentrain*** → ***boute-entrain***) ou lorsqu'il ignore l'existence d'une soudure (***betterave*** → ***bette-rave***). Certaines séquences de caractères décomposables en lexèmes autonomes sont qualifiées d'*EP monotokens* ou, en anglais, *single-token multi-word expressions* (MWEs) (Baldwin et Kim, 2010). Ce principe d'autonomie exclut les compositions par préfixation ou suffixation (*parapluie* → **para* + *pluie*, *acidophile* → *acid(e/o)* + **phile*). ***Pont-levis*** est pourtant considéré comme une EP bien que *levis* n'aie pas d'existence autonome. Loin de remettre en question la définition des EP, cette particularité constitue au contraire l'une de leurs spécificités comportementales.

1.2 Comportement idiosyncratique des EP

La caractéristique majeure des EP est de présenter des comportements inhabituels que l'on qualifie d'*idiosyncratiques*, tels que l'incapacité de prédiction des propriétés de l'expression à partir des celles de ses composants. C'est ce que l'on nomme le principe de *non-compositionnalité* puisque dans ce cas, si l'on note A et B les propriétés respectives des composants, et $A + B$ leur association, alors $A + B \neq AB$, où AB représenterait les propriétés de l'ensemble de l'EP. Cette idiosyncrasie des EP s'observe à différents niveaux (Baldwin et Kim, 2010) :

— *Idiosyncrasie lexicale* : une EP peut comporter un ou plusieurs composants qui ne figurent pas dans le lexique de la langue de communication comme dans le cas de la locution latine `(LA)` ***ad hoc***. De même, certaines EP comportent des composants dépourvus d'existence autonome, comme *levis* dans ***pont-levis***, *mouron* ou *prétentaine* dans ***se faire du mouron*** ⇒ 's'inquiéter' ou ***courir la prétentaine*** ⇒ 'vagabonder'. De tels lexèmes sont qualifiés de *cranberry word* selon la terminologie d'Aronoff (1976) car `(EN)` *cran* n'existe pas en tant que lexème autonome contrairement à *blue* ou *black* (p. ex. `(EN)` ***blueberry***, ***blackberry***).

— *Idiosyncrasie syntaxique* : une EP peut juxtaposer des composants *a priori* syntaxiquement incompatibles, comme dans la locution adverbiale `(EN)` ***by and large*** qui résulte de la coordination d'une préposition et d'un adjectif. Parfois, la partie de discours[3] d'une EP n'est pas déductible à partir de celles de ses composants :

sont lexicalisées. Notre notion de la lexicalisation élargit cette terminologie standard, puisqu'elle s'applique non seulement aux EP mais également à leurs composants [Notre traduction].

3. La partie du discours (en anglais *part of speech* ou POS) d'un mot correspond à sa nature (ou classe grammaticale) : nom (NOUN au format Universal Dependencies `https://universaldependencies.org`), verbe (VERB), déterminant (DET), numéral (NUM), adjectif (ADJ), adverbe (ADV), pronom (PRON), préposition (ADP ou adposition), ponctuation (PUNCT), auxiliaire (AUX), etc. Ce format sera exclusivement utilisé dans les parties III et IV

laissez-passer est un substantif (et est donc employé comme tel) alors qu'il se compose de deux verbes. Précisons à ce sujet qu'il ne suffit pas qu'une EP contienne un verbe pour la qualifier d'*EP verbale* : il faut également qu'elle fonctionne comme tel.

— *Idiosyncrasie morphologique* : l'EP ***grand-mère*** est un substantif de genre féminin constitué à partir d'un nom féminin et d'un adjectif masculin, ce qui fait figure d'anomalie. De même, l'unique substantif dans ***porte-monnaie*** est féminin mais la résultante est un substantif masculin. D'autres EP nécessitent une flexion particulière comme ***lance-flammes*** ou ***sucrer les fraises*** ⇒ 'être gâteux' dont le nom doit être au pluriel, ce qui traduit un *figement* morphologique.

— *Idiosyncrasie sémantique* : dans l'expression ***casser du sucre sur le dos*** ⇒ 'dire du mal de quelqu'un en son absence', il n'est en réalité pas plus question de *sucre* que de *dos*. Une telle expression présente une non-compositionnalité sémantique qui oblige les apprenants (enfants ou non-francophones) à la mémoriser, faute d'être capables de la produire spontanément. Les EP non compréhensibles de prime abord sont dites *opaques* (par opposition à *transparentes*). Une EP telle que ***en rester comme deux ronds de flan*** ⇒ 'être stupéfait' requiert effectivement une connaissance de l'expression lors de l'encodage et du décodage tandis que seul l'encodage est difficile pour les EP transparentes (***enfoncer une porte ouverte*** ⇒ 'énoncer une banalité') (Fillmore *et al.*, 1988). A mi-chemin entre les EP sémantiquement non-compositionnelles (***en rester comme deux ronds de flan***) et compositionnelles (`(EN)` ***fish and chips***), se situent les EP semi-compositionnelles pour lesquelles seule une partie des composants éclaire le sens de l'EP, par exemple du ***vin gris*** qui est bien un vin mais dont la robe réelle est saumonée.

— *Idiosyncrasie pragmatique* : une combinaison de tokens peut prendre un sens particulier dans un contexte d'énonciation spécifique comme le ***je vous en prie*** donné en réponse à un remerciement.

— *Idiosyncrasie statistique* : on observe que certaines combinaisons de tokens sont plus fréquentes que d'autres ou tendent à apparaître dans un ordre préférentiel comme ***noir et blanc*** ou `(EN)` ***black and white*** '*noir et blanc*'. La dimension arbitraire de cet ordre est soulignée par l'inversion des couleurs en japonais et en espagnol : `(JP)` ***shirokuro*** '*blanc noir*'⇒'blanc et noir' et `(ES)` ***blanco y negro*** '*blanc et noir*'⇒'blanc et noir' (Baldwin et Kim, 2010). Cet ordre préférentiel de mots est souvent motivé par des raisons sémantiques ou phonologiques (Abraham, 1950). Dans certains cas, toute modification de l'ordre des composants (souvent reliés par une coordination) est perçue comme non naturelle ou incorrecte. Si deux composants sont mis en jeu, on parle alors de *binômes irréversibles* (en anglais *irreversible binomials* selon la terminologie de Malkiel (1959)) : c'est ce que l'on observe pour la locution adverbiale ***bel et bien*** ⇒ 'réellement' vs. **bien et bel*. Cette irréversibilité n'est pas nécessairement permanente : certains binômes deviennent irréversibles, comme ***bel et bien*** dont les composants apparaissaient dans l'ordre inverse au XIII^{e} siècle (Larrivée et Moline, 2012), et d'autres réversibles, ce qui explique que l'on puisse leur attribuer un degré d'irréversibilité d'après les différentes combinaisons observées en corpus (Mollin, 2013). De notre point de vue, seuls les binômes irréversibles relèvent du cadre des expressions polylexicales.

L'idiosyncrasie se fonde sur la perception d'un fonctionnement irrégulier, anormal, d'une expression donnée. Or, cette perception s'appuie sur les connaissances du locuteur, variables selon son origine sociale, géographique ou l'époque à laquelle il ou elle vit. La connaissance du latin diminue ainsi le niveau d'idiosyncrasie perçu face à l'emploi de locutions latines (Baldwin et Kim, 2010). De même, l'exemple de ***grand-mère/grand-tante*** est aujourd'hui ressenti comme idiosyncratique car *grand* est un adjectif masculin. Un locuteur antérieur au XVI^{e} siècle n'y verrait cependant aucune idiosyncrasie puisque *grand* était initialement épicène (comme en témoignent ***grand-route, grand-messe***, etc.) à l'instar de *fort* dans ***avoir fort envie*** (Blanco et Bogacki, 2014). Somme toute relative, cette idiosyncrasie des EP est parfois soulignée par des critères formels tels que l'absence de déterminants là où des expressions libres en nécessiteraient : ***avoir ∅ besoin*** mais pas **avoir ∅ nécessité*.

L'idiosyncrasie est une propriété définitoire des EP, mais on peut en citer d'autres, comme le fait d'être substituables par un seul lexème (Baldwin et Kim, 2010). Par exemple, ***apporter*** *une* ***solution*** est paraphrasable par le néologisme *solutionner* et traduisible en anglais par *to solve*. Néanmoins, toutes les EP ne sont pas substituables par un monolexème – et la possibilité d'une telle substitution n'implique pas non plus qu'il s'agit nécessairement d'une EP – comme l'illustre l'EP ***coiffer Sainte Catherine*** dans l'exemple 1.1 (Haßler et Hümmer, 2005).

(1.1) *Ah ! vous pourrez bien vous mordre les doigts, si vos filles* ***coiffent Sainte-Catherine*** ⇒ 'portent un chapeau jaune et vert le 25 novembre pour signaler qu'à 25 ans elles sont encore célibataires' (Emile Zola, *Pot-Bouille*)

C'est pourquoi ce critère est considéré comme une propriété additionnelle, non suffisant à lui seul pour savoir si l'on a affaire à une EP. Notons par ailleurs qu'une EP donnée peut présenter une ou plusieurs idiosyncrasies, par exemple morphologique + sémantique pour la plante dénommée ***miroir des elfes*** (# *miroir de l'elfe*). Autrement dit, une EP est plus ou moins irrégulière par comparaison avec une expression libre qui, elle, ne présente aucune idiosyncrasie. Cette absence d'idiosyncrasie n'est toutefois pas simple à définir. Pour Lichte *et al.* (2017), derrière la plupart des idiosyncrasies se dissimulent en réalité des formes inédites de régularités. En effet, un comportement est jugé idiosyncratique par rapport aux comportements observés pour un ensemble d'expressions. Reste alors à définir si ce référentiel est exclusivement composé d'expressions libres ou non. L'expression ***vider son sac*** ⇒ 'avouer' requiert par exemple un accord entre le sujet du verbe et le possessif, ce qui est perçu comme un comportement idiosyncratique par rapport à une expression libre (p. ex. *il/je mange sa pomme*), bien que de nombreuses EP partagent cette contrainte (p. ex. ***tirer sa révérence*** ⇒ 's'en aller', ***casser sa pipe*** ⇒ 'mourir').

Nous retiendrons, pour définir la notion d'EP, les critères distinctifs les plus répandus (Savary, 2008) :

- une EP contient au moins deux unités lexicales ("mots") clairement identifiables (p. ex. ***prendre décision, court-circuiter***),
- une EP possède une dénotation unique [4] et constante,
- une EP présente une idiosyncrasie lexicale, morphologique, syntaxique, statistique [5]

4. Un nombre *a priori* réduit de contre-exemples dispose de dénotations multiples comme ***prendre la tête*** ⇒ 'se retrouver en première position d'une compétition non encore terminée / faire perdre patience'.

5. Notons cependant que, en accord avec les définitions établies par le réseau PARSEME, ce seul critère

ou sémantique.

L'ampleur des idiosyncrasies, dépendant de chaque EP, sous-entend un dégradé en matière de figement (et, inversement, de variabilité), ce qui est également un trait distinctif des EP puisque les EP complètement figées sont rares (p. ex. (LA) ***ad hoc***). Par ailleurs, en plus d'être hétérogènes du point de vue de leurs idiosyncrasies respectives, leurs caractéristiques donnent lieu à de nombreuses définitions d'EP dont nous dressons un aperçu dans le prochain chapitre.

ne serait pas jugé suffisant pour considérer qu'il s'agit bien d'une EP.

Chapitre 2

Hétérogénéité des EP

Ce chapitre passe en revue différentes classifications relatives aux EP (section 2.1) et met en évidence la diversité et l'importance numérique des EP (section 2.2). Une discussion sur cet état de l'art est proposée en section 2.3.

2.1 Consensus et divergences

Les EP apparaissent comme des expressions de nature et de figement très hétérogènes que la littérature associe à plusieurs dénominations (non-exclusives et parfois orthogonales les unes des autres) et dont nous présentons une synthèse en signalant les incohérences (contre-exemples ou divergences) grâce au symbole ⚠.

2.1.1 Mots composés

Les mots composés sont des juxtapositions de lexèmes libres permettant de former un nouveau lexème (Constant *et al.*, 2017). Autrement dit, ils se comportent comme des mots simples et n'autorisent pas d'insertion comme le montrent les exemples suivants : ***pomme de terre*** / ***pomme de terre*** *pourrie* / **pomme pourrie de terre*.

Par composition, la séquence finale peut avoir une partie de discours différente de ses éléments, comme dans ***laissez-passer*** (Verbe + Verbe → Nom). On retrouve des mots composés de structures diverses par exemple Adj-Nom (***grand-mère***), Nom de Nom (***pomme de terre***), Nom à Nom (***verre à pied***) voire des compositions plus originales comme *le* ***qu'en dira-t-on***. Pour Gross (1988a), les mots composés ne se restreignent pas aux *locutions*[1] nominales (ou *noms composés*) comme en attestent les locutions adjectivales (*un garçon* ***comme il faut***), adverbiales (***tout à fait***), verbales (***mettre fin***), etc. De nombreux mots composés (notamment les locutions nominales ou adverbiales) sont constitués d'unités contiguës, dans un ordre constant[2] et, en français écrit, séparées graphiquement (*word-with-spaces*), si bien qu'on pourrait les intégrer au lexique en leur substituant une

1. Pour Spence (1969), "les catégories du *composé* et de la *locution* se confondent en français à tel point que la distinction perd pour ainsi dire toute valeur scientifique".

2. ***Mère-grand*** est par conséquent un mot-composé différent de ***grand-mère***, d'autant plus que **tante-grand* n'existe pas.

forme concaténée (p. ex. ***tout_à_fait***). Cette méthode ne serait cependant guère satisfaisante pour les mots-composés non invariables, notamment en raison de leur variabilité morphologique (p. ex. ***pomme(s) de terre***, ***mettr(e/ai/as/...) fin***).

⚠ Des divergences existent sur la prise en compte des soudures : Mathieu-Colas (1995) considère les mots soudés comme étant des mots composés, contrairement à Savary (2008). De plus, certains mots composés tolèrent parfois que leurs composants ne soient pas systématiquement juxtaposés. 👆 Nous qualifions de *discontinuité* toute suite d'éléments non lexicalisés situés entre le premier et le dernier composant d'une EP, le nombre d'éléments non lexicalisés étant considéré comme la *longueur* de la discontinuité. Des cas d'insertions adverbiales attestent de discontinuités dans *un* ***livre*** *soi-disant* ***rare*** (Gross, 1988a) ou ***à*** *très* ***court terme*** (exemple cité par Green *et al.* (2013)). Cela explique le constat de Gross (1988a) selon lequel "il se peut que la notion de *mot composé* vole en éclats".

2.1.2 Termes polylexicaux

A la différence des mots de la langue générale, les *termes* font référence à des concepts liés à un domaine de spécialité (Constant *et al.*, 2017). On distingue les termes simples (mono-tokens) comme le substantif *photon* et les termes dits complexes car composés de plusieurs lexèmes, comme l'expression ***laser à solide***. Ces deux exemples nominaux sont liés au secteur de la physique mais seuls les termes complexes relèvent des EP. Les termes polylexicaux ne se restreignent pas aux EP nominales (p. ex. l'EP verbale ***rejeter de souche*** ⇒ 'émettre de nouvelles tiges à proximité de la souche d'un arbre' en botanique). La proportion de termes complexes dans un lexique spécialisé est bien plus élevée que celle de termes simples : en japonais, ces termes complexes représentent environ 80% d'un lexique spécialisé (Nakagawa et Mori, 2003) (cités par Morin et Daille (2012)), d'où l'intérêt de les prendre en considération, notamment pour la traduction technique. Or, comme le soulignent Morin et Daille (2012) :

> Les domaines de spécialités sont caractérisés par des ressources textuelles réduites en comparaison à la langue générale et par une grande proportion de vocabulaire spécifique qui n'est pas présent dans les dictionnaires monolingues ou bilingues de langue générale.

Ces termes peuvent être discontinus en raison d'une modification adjectivale par exemple : ***liaisons*** *multiples* ***par satellite*** (Savary et Jacquemin, 2003). Les *termes* sont habituellement monosémiques au sein de leur domaine à la différence des *mots* du langage général.

⚠ En linguistique, le terme *mot composé* n'est pas monosémique comme en attestent les divergences vis-à-vis des mots soudés précédemment mentionnées. Nous rejoignons donc le constat plus nuancé de Savary et Jacquemin (2003) selon lequel "[m]ulti-word terms are far less polysemous than single-word terms" [3].

3. Les termes complexes sont bien moins polysémiques que les termes simples [Notre traduction].

2.1.3 Expressions idiomatiques ou idiomes

Le sens global des expressions idiomatiques ou idiomes n'est pas déductible du sens individuel de chacun de leurs composants du point de vue de la représentation mentale (p. ex. ***poser un lapin*** ⇒ 'ne pas venir au rendez-vous' ≠ *poser + 1 +* 🐇). Les expressions idiomatiques comprennent des emplois métaphoriques (p. ex. ***tourner la page*** ⇒ 'commencer une nouvelle étape d'une vie en oubliant le passé') ou des comparaisons (*similes*) telles que ***saoul comme un Polonais***, ainsi que – dans la terminologie PARSEME décrite en section 8.1.3.2 – des constructions grammaticales telles que l'impersonnel ***il y a***. Ces expressions idiomatiques sont souvent spécifiques à une langue donnée, notamment lorsqu'elle mettent en jeu des préjugés sur d'autres cultures (Ex. 2.1-2.4).

(2.1) `(FR)` ***C'est de l'hébreu/du chinois*** ⇒ 'C'est incompréhensible'

(2.2) `(EN)` ***It's all Greek to me*** '*C'est du grec pour moi*'⇒'C'est incompréhensible'

(2.3) `(DE)` ***Das kommt mir Spanisch vor*** '*Ça me paraît être de l'espagnol*'⇒'C'est incompréhensible'

(2.4) `(PL)` ***Siedzieć jak na tureckim kazaniu.*** '*Être assis comme lors d'un sermon en turc*'⇒'C'est incompréhensible'

Sheinfux *et al.* (2018) qualifient de *figuratifs* les idiomes auxquels le locuteur associe une représentation imagée. Plus précisément, ils distinguent les *idiomes figuratifs transparents* de ceux qui sont *opaques*. Lorsqu'ils sont transparents, l'image mentale entretient une relation étroite avec le sens de l'expression (***avoir un petit pois à la place du cerveau*** ⇒ 'être peu intelligent', ***tirer sa révérence*** ⇒ 's'en aller'). En revanche, lorsqu'ils sont opaques, aucun lien évident ne s'établit entre l'image mentale et le sens de l'EP comme dans ***apporter des oranges*** ⇒ 'rendre visite à quelqu'un en prison ou à l'hôpital' ou ***poser un lapin***. A la différence de Nunberg *et al.* (1994) qui établissent une corrélation entre la décomposabilité sémantique des EP (voir Section 3.4) et leur flexibilité, Sheinfux *et al.* (2018) suggèrent que plus un idiome est figuratif et transparent, plus il est susceptible d'être flexible.

⚠ Le choix de l'adjectif *idiomatique* (c.-à-d. propre à une langue donnée) n'est pas toujours justifié car, même si cela est plutôt rare, certaines EP ont des correspondances exactes entre différentes langues comme ***prendre le taureau par les cornes*** ⇒ 's'attaquer à une difficulté avec détermination' ⇔ `(EN)` ***take the bull by the horns***.

2.1.4 Collocations

Le terme *collocation* est particulièrement ambigu car plusieurs définitions coexistent (Savary *et al.*, 2018; Seretan, 2011). On distingue cependant deux courants majoritaires dans la tradition européenne :

> [I]n the British contextualist framework (Firth, 1961; Halliday et Hasan, 1976; Williams, 2003), collocations can be broadly defined as recurrent lexical elements which contribute to the text cohesion. In the "continental" tradition (Williams, 2003), collocations are also called "restricted lexical collocations" and considered as lexicalised phrases where two recurrent lexical elements have

a syntactic relationship (Tutin, 2008). [4]

Dans la vision dite *continentale*, les collocations décrites sous forme binaire base-collocatif ont pour particularité que le *collocatif* n'est pas libre mais imposé ou restreint par la *base* : pour exprimer l'intensité on utilise des adjectifs distincts dans ***peur bleue***, ***colère noire***, ***gros fumeur***, etc. Ces adjectifs (collocatifs) sont sélectionnés en fonction du nom (base) qui est le seul à conserver son sens habituel (Mel'čuk, 1998; Tutin et Grossmann, 2002; Polguère, 2003; Pausé, 2017), les collocations sont donc semi-compositionnelles. En effet, une ***peur bleue*** est bien une peur mais qui n'a rien à voir avec une quelconque couleur, ce qui souligne le caractère arbitraire du collocatif. Les collocations ne sont pas limitées à deux composants, celles impliquant plus de deux éléments pouvant être vues comme des collocations distinctes ayant fusionné (p. ex. ***essuyer*** *un* ***échec*** + ***échec cuisant*** → ***essuyer*** *un* ***échec cuisant***) (Tutin et Grossmann, 2002).

La vision anglo-saxonne rejoint la définition des EP institutionnalisées (ou *collocations*) que Sag *et al.* (2002) distinguent des EP lexicalisées (de figement variable) décrites en Section 3.4. Ces collocations sont compositionnelles à la fois du point de vue sémantique et syntaxique mais présentent une idiosyncrasie statistique. Elles se caractérisent en effet par la tendance des locuteurs à préférer l'association de certains tokens (par exemple : ***aimer à la folie***) voire leur ordre comme `(EN)` ***fish and chips*** qui est plus courant que *chips and fish* en Grande-Bretagne (Villavicencio *et al.*, 2005). Nerima *et al.* (2016) proposent la version Web d'un outil multilingue d'extraction de collocations à partir de corpus (Seretan, 2011) : parmi les séquences découvertes, on trouve aussi bien des EP (***prendre décision***, ***lancer appel***), que des non-EP (*débrancher respirateur, disparaître dispensaire*).

⚠ Le fait d'inclure des co-occurences (c.-à-d. la présence simultanée de plusieurs tokens) statistiquement significatives peut avoir pour conséquence d'inclure des co-occurrences qui ne sont pas des EP comme *lire ci-dessous* ou `(EN)` *doctor–sick* (Church et Hanks, 1990). La question du nombre de tokens pris en compte pour constituer la fenêtre de recherche de co-occurrences se pose également. Notons enfin que le statut *collocation* vs. *non-collocation* constitue une typologie orthogonale aux autres classifications d'EP de cette section.

2.1.5 Verbes à particules ou *verb-particle constructions* (VPC)

Les VPC consistent en un verbe et une particule (préposition ou adverbe) qui en modifie le sens initial. Les VPC sont très fréquentes dans les langues germaniques comme `(EN)` ***to give up*** '*donner en haut*'⇒'abandonner', qui ne peut être compris à partir du sens individuel de *give* (*donner*) et *up* (préposition indiquant entre autres un mouvement ascendant). Les VPC françaises sont rares en français contemporain (p. ex. ***faire avec*** ⇒ 'accepter'). Elles sont majoritairement attestées en diachronie (XIII^{e} siècle : ***tourner arrière*** ⇒ 'retourner' (Le Marchant et Kunstmann, 1973)) ou dans des variantes de français hors métropole, comme en Belgique – sous l'influence du flamand – ou au Québec – sous l'influence de l'anglais : `(BE)` ***regarder après*** ⇒ 'chercher' ou `(CA)` ***back frapper*** ⇒

4. Dans le cadre contextualiste anglo-saxon [...], les collocations sont généralement définies comme des éléments lexicaux récurrents qui contribuent à la cohésion du texte. Dans la tradition "continentale" [...], les collocations sont également appelées "collocations lexicales restreintes" et considérées comme des expressions lexicalisées dans lesquelles deux éléments lexicaux récurrents entretiennent une relation syntaxique. [Notre traduction]

'renvoyer un coup' (Treffers-Daller, 2012).

2.1.6 Entités nommées polylexicales (ENP)

La polylexicalité concerne également les entités nommées polylexicales telles que ***Valéry Giscard d'Estaing***. L'intérêt pour les ENP, antérieur à celui porté aux EP, était motivé par des besoins en extraction d'information. La particularité des ENP réside dans leur richesse sémantique et pragmatique, représentable par des liens vers des référents d'entités et de concepts du monde réel ou du monde du discours. D'après Sekine *et al.* (2002), les entités nommées incluent :

- des noms de personnes : ***Valéry Giscard d'Estaing***
- des noms d'organisations : ***Laboratoire d'Informatique Fondamentale et Appliquée de Tours***
- des noms de lieux : ***Centre-Val de Loire***
- des noms d'objets : ***New Beetle***
- des noms d'événements : ***Fête des Lumières***
- des dates : ***30-11-2008, 21 janvier 2012***
- des pourcentages : ***20,5%***
- des données monétaires : ***100 millions d'euros***

La variabilité des ENP s'observe dans l'emploi d'ellipses dont la nature n'est pas toujours prévisible (***Giscard d'Estaing*** → ***Giscard*** mais pas *#d'Estaing*). Elles peuvent prendre la forme d'acronymes *(**VGE**)*[5] et tolérer des discontinuités (***Amérique du Nord** et **du Sud***) ou des enchâssements tels que le [***musée de*** [***Madame Tussauds***]$_{\text{personne}}$]$_{\text{lieu}}$, ce phénomène étant fréquent dans le domaine biomédical ([***EBV-transformed*** [***human B cell line***]$_{\text{CELL\_LINE}}$]$_{\text{CELL\_LINE}}$) (Katiyar et Cardie, 2018). Les ENP sont aussi soumises à des contraintes spécifiques en matière de traduction, certaines étant traduites (***Christophe Colomb*** = `(EN)` ***Christopher Colombus*** = `(ES)` ***Cristóbal Colón***), d'autres non (***George Bush*** ≠ *Georges Buisson*).

Les entités nommées ont fait l'objet de nombreux travaux visant à les identifier dans des corpus textuels, dans le cadre d'une tâche dite *Named-Entity Recognition* (ou NER). En effet, leur identification s'est avérée cruciale pour des applications telles que l'extraction d'informations (qui ? où ? combien ? etc.), comme l'illustre la phrase : ***Notre-Dame***$_{lieu}$ *: déjà plus de **880 millions d'euros***$_{monnaie}$ *de dons selon **Stéphane Bern***$_{personne}$.

⚠ On peut s'interroger sur le fait de considérer les expressions numériques (dates, monnaie, pourcentages) comme des ENP. Quoique soumises à des contraintes formelles, comme l'expression de la date dont l'ordre jour-mois-année diffère d'une langue à l'autre, ou la convention d'écriture de nombres qui varie selon le contexte (p. ex. ***300 millions d'euros*** vs. 3.10^8 en écriture mathématique) ou la préférence pour une écriture entière (p. ex. ***300 millions d'euros*** plutôt que 0,3 milliard), nous les considérons en définitive comme des cas tangents d'ENP.

5. De notre point de vue, seuls les sigles dont la forme étendue est une ENP relèvent à leur tour des ENP. C'est ainsi le cas pour ***VGE*** (= ***Valéry Giscard D'Estaing***), mais pas pour le monogramme de Charlemagne (CRLS = *CaRoLuS*).

2.1.7 Constructions à verbe support (CVS)

Dans les CVS, ce n'est pas le verbe qui remplit la fonction de prédicat de la phrase, mais un nom prédicatif. Nombre de CVS peuvent être substituées par le verbe dérivé du nom qui les composent. Par exemple : ***prendre décision*** (= *décider*), ***prendre la fuite*** (= *fuir*), ***venir en aide*** (= *aider*), etc. Le verbe peut alors s'effacer (*la **décision** que Jean **prend** est importante* ↔ *la décision de Jean est importante*). De fait, le verbe des CVS ne sert, en français, qu'à porter l'information de flexion personnelle et temporelle, d'où sa qualification de *verbe support* ou *verbe léger* en anglais (*light verb*).

⚠ A l'instar de Smadja (1993); Mel'čuk (1998), Tutin et Grossmann (2002) considèrent que les CVS telles que ***prendre décision*** sont des collocations dans lesquelles le nom prédicatif serait la base tandis que le verbe support en serait le collocatif. Cela souligne à quel point la typologie des EP ne fait pas consensus, notamment dès lors que l'on emploie le terme de *collocation*.

L'hétérogénéité des EP relevées dans cette section permet de prendre la mesure de l'importance qualitative du phénomène, mais son importance quantitative ne doit pas non plus être sous-estimée.

2.2 Fréquence des EP

D'après Jackendoff (1997), le lexique de chaque langue contient autant d'EP que de tokens isolés. Un extrait du roman *Arsène Lupin Gentleman Cambrioleur* (Leblanc, 1907) (Fig. 2.1) suffit à prendre la mesure de la quantité et de la diversité d'EP susceptibles d'apparaître dans un texte, surtout si l'on adopte une définition large des EP incluant les soudures (***monsieur***), les entités nommées polylexicales en relation avec des personnes, des lieux, des dates ou des informations monétaires (***Arsène Lupin, gare des Batignolles, jeudi 28 septembre, trente mille francs***), des collocations (***port payé***), des routines langagières telles que les formules de politesse (***veuillez accepter l'expression...***), des mots composés (***sus-indiqués***), des constructions à verbe support (***procéder à déménagement***) ainsi que deux autres catégories d'EP verbales décrites plus en détails dans la section 8.1.2 : idiomes verbaux (***il y a***) et verbes intrinsèquement réflexifs (***se contenter***). Les éléments lexicalisés, en caractères gras, constituent 42% des tokens du texte, ce qui est proche de la valeur obtenue par Gross et Senellart (1998) dans le journal *Le Monde* (40% des tokens). Toutefois, ce pourcentage est probablement sous-évalué si l'on prend en considération des domaines ayant une terminologie spécifique (scientifique, médicale, juridique, etc.) (Sag *et al.*, 2002). De plus, il est impossible d'établir la liste exhaustive des EP existantes car de nouvelles expressions apparaissent sans cesse, qu'il s'agisse d'entités nommées polylexicales (noms de personnes, d'organisations, etc.) ou de néologismes obtenus en réutilisant le lexique disponible en fonction de l'actualité sociale (***gilet jaune***) ou de nouvelles technologies (***objet connecté***) par exemple.

Cette thèse ne vise pas l'identification exhaustive de tous ces types d'EP, mais exclusivement celle d'EP verbales, leur profil de variabilité étant susceptible d'être particulièrement riche. De ce fait, nous excluons les mots composés ne relevant pas de locutions verbales ainsi que les ENP, sauf lorsqu'elles s'intègrent dans des EP verbales (p. ex. ***coiffer Sainte***

« ***Mon***$_1$ ***sieur***$_{1\text{-}2}$ ***le***$_2$ ***baron***$_2$,

« ***Il***$_3$ ***y***$_3$ ***a***$_3$, dans la galerie qui réunit vos deux salons, un tableau de ***Philippe***$_4$ ***de***$_4$ ***Champaigne***$_4$ *d'excellente facture et qui me plaît infiniment.* [...]

« *Pour cette fois, je* ***me***$_5$ ***contenterai***$_5$ *de ces objets qui seront, je crois, d'un écoulement facile.* ***Je***$_6$ ***vous***$_6$ ***prie***$_6$ *donc de les faire emballer convenablement et de les expédier à mon nom (*___port___$_7$ ___payé___$_7$*), en* ***gare***$_8$ ***des***$_8$ ***Batignolles***$_8$*, avant huit jours... faute*$_9$ ***de***$_9$ *quoi, je ferai* ***procéder***$_{10}$ ***moi***$_{11}$***-même***$_{11}$ ***à***$_{10}$ *leur* ***déménagement***$_{10}$ *dans la nuit du* ***mercredi***$_{12}$ ***27***$_{12}$ *au* ***jeudi***$_{13}$ ***28***$_{13}$ ***septembre***$_{12\text{-}13}$*. Et,* ***comme***$_{14}$ ***de***$_{14}$ ***juste***$_{14}$*, je* ***ne***$_{15}$ ***me***$_{16}$ ***contenterai***$_{16}$ ***pas***$_{15}$ *des objets* ***sus***$_{17}$***-indiqués***$_{17}$.

« ***Veuillez***$_{18}$ *excuser le petit* ***dérangement***$_{19}$ *que je vous* ***cause***$_{19}$, et ***accepter***$_{18}$ ***l***$_{18}$*'****expression***$_{18}$ ***de***$_{18}$ ***mes***$_{18}$ ***sentiments***$_{18}$ ***de***$_{18}$ ***respectueuse***$_{18}$ ***considération***$_{18}$.

« ***ARSÈNE***$_{20}$ ***LUPIN***$_{20}$. »

« ***P.***$_{21}$***-S.***$_{21}$ – ***Sur***$_{22}$***tout***$_{22}$ ***ne***$_{23}$ ***pas***$_{23}$ *m'envoyer le plus grand des Watteau.* ***Quoi***$_{24}$***que***$_{24}$ *vous l'ayez payé* ***trente***$_{25}$ ***mille***$_{25}$ ***francs***$_{25}$ *à l'****Hôtel***$_{26}$ ***des***$_{26}$ ***Ventes***$_{26}$*, ce* ***n***$_{27}$*'est* ***qu***'$_{27}$*une copie.*

FIGURE 2.1 – Aperçu de la fréquence et de l'hétérogénéité des EP dans un extrait de roman (Leblanc, 1907). Les indices numériques signalent les composants d'une même EP.

Catherine). Les formules épistolaires de politesse sont également écartées de l'étude.

La fréquence des EP dans le corpus `FR-train1.1` qui sera le support de ce travail (voir section 8.1.3) respecte la loi de Zipf : un nombre réduit d'EP apparaît très fréquemment tandis que la majorité apparaît rarement. Comme le rappelle Wyllys (1981), cette loi de Zipf se fonde sur l'observation que la fréquence d'utilisation d'un mot dans un texte volumineux est inversement proportionnelle à son rang r : $f(r) = \frac{C}{r}$ où C est une constante, ce qui se traduit par une droite en échelle logarithmique (Fig. 2.2). Autrement dit, d'après cette loi, dans un corpus[6] comportant 1508 types d'EP verbales instanciés par 4550 tokens, l'EP la plus fréquente (***il y avoir***) est deux[7] fois plus utilisée que la deuxième (***il falloir***), trois fois plus que la troisième (***il s'agir***), etc. Cette distribution zipfienne souligne la difficulté d'un recensement exhaustif des EP.

2.3 Classer les EP : *la quadrature du cercle* ?

La revue de l'état de l'art montre combien il est difficile de classer les EP. Une même EP peut d'ailleurs s'inscrire dans différentes catégories : la plante nommée ***lin de Nouvelle-Zélande*** remplit les conditions pour être classée comme un mot composé (*#lin bleu de Nouvelle-Zélande*), comme une expression idiomatique (aucune parenté avec le lin), comme un terme botanique et, de surcroît, elle contient une ENP (***Nouvelle-Zélande***).

De plus, il n'existe pas de dichotomie stricte entre expressions complètement figées (qui seraient des EP) et expressions libres (qui seraient des non-EP), comme le souligne la

6. `FR-train1.1`.

7. Ce facteur dépend en réalité de la constante C, elle-même fonction de la taille du corpus et de la nature du phénomène étudié.

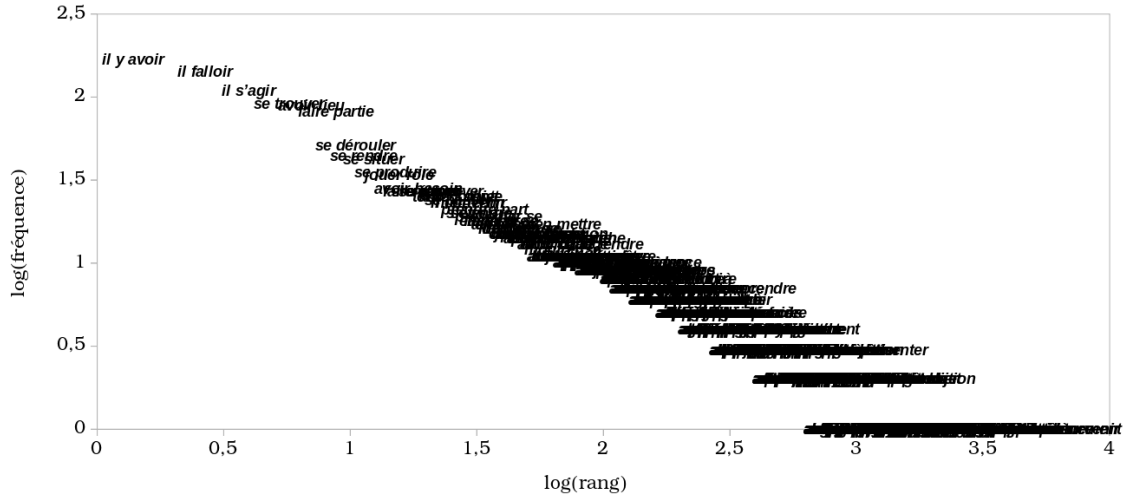


FIGURE 2.2 – Illustration de la loi de Zipf : la majorité (resp. la minorité) des types d'EP sont rares (resp. fréquents) dans le corpus `FR-train1.1`.

classification tripartite de Sag *et al.* (2002) mettant en évidence des modalités de figement partiel. A la différence des expressions dites libres tolérant toutes les transformations possibles (substitution d'un composant par un synonyme, flexion nominale ou verbale, etc.), une expression figée est, par principe, dénuée de toute forme de variabilité y compris morphologique. Rares en sont les exemples : on citera les locutions étrangères (`(LA)` ***ad hoc***, `(IT)` ***al dente***) ou les proverbes (***chose promise, chose due***). Toutefois, l'ellipse est fréquente dans les proverbes (***quand on parle du loup...*** ⇒ 'une personne arrive au moment où on en parle'), sans qu'il soit possible de prédire de façon certaine la création de nouvelles EP incomplètes à partir d'idiomes, de proverbes ou de citations (Riehemann, 2001). Par ailleurs, certaines EP présentées comme figées tolèrent en réalité des arguments au sein d'une classe d'équivalences relativement étendue, comme illustré dans (Ex. 2.5) (Gross, 1988b). Ici, le mot *raison* dans ***perdre la raison*** ⇒ 'être désorienté' alterne avec *tête* ou *boule* qui entretiennent une relation métaphorique (*tête* ↔ *boule*) ou métonymique (*tête* ↔ *raison*). En revanche, le lien est moins évident avec *nord* (analogie avec le fonctionnement d'une boussole) ou *pédales* (analogie avec le déséquilibre sur un vélo).

(2.5) ***perdre la raison/la tête/la boule/le nord/les pédales etc.*** ⇒ 'être désorienté'

De même, ***donner lieu*** qui fait partie des EP les moins variables d'après Tutin (2016), tolère la flexion temporelle et personnelle (***donnait/ donneront lieu***).

Contrairement à la plupart des mots composés, les expressions polylexicales verbales ne peuvent pas être modélisées sous la forme de *word-with-spaces* car elles autorisent fréquemment l'insertion de modifieurs adverbiaux (Ex. 2.6) ou d'arguments [8] (Ex. 2.7).

8. On distingue les *arguments* qui sont obligatoires des *modifieurs* qui sont optionnels (voir la présentation d'Agnieszka Patejuk intitulée *Verbal phraseology... To a parser* lors de la formation dispensée durant la réunion PARSEME du 27 juin au 1er juillet 2016 à La Rochelle, consultable à l'adresse `https://typo.uni-konstanz.de/parseme`). Dans la phrase *Jean rencontre souvent Marie*, *Marie* est un argument (**Jean rencontre souvent*) tandis que *souvent* est un modifieur (*Jean rencontre Marie*).

(2.6) *Il ne* ***prendra*** *jamais* ***le taureau par les cornes***

(2.7) *Il a* ***pris*** *tout le monde* ***de court*** ⇒ 'prendre par surprise'

Les EP verbales ne sont d'ailleurs pas les seules à autoriser des insertions, les prépositions par exemple y sont également sujettes (***sans*** *...* ***pour autant***).

En conclusion, l'absence de consensus sur les EP est révélée par des classifications qui se heurtent à des contre-exemples (mots composés tolérant des insertions, expressions figées qui ne le sont pas) et par des définitions recouvrant des approches différentes (collocations). Un compromis entre les différentes terminologies s'avère donc nécessaire et sera présenté en section 8.1.2. De façon orthogonale, d'autres classifications employées dans la littérature reposent davantage sur la variabilité des EP que sur leur constitution.

Chapitre 3

Variabilité des EP

Ce chapitre commence par rappeler la distinction entre *type* d'EP et *token* d'EP (section 3.1), cette dernière notion étant particulièrement utile pour représenter la diversité de variantes observables. La dimension imprévisible de la variabilité des EP est ensuite évoquée (section 3.2), puis la façon dont cette variabilité est perçue : soit de façon quantitative (section 3.3), soit qualitative (section 3.4).

3.1 Type vs. token

Hormis les EP entièrement figées, la plupart des EP sont susceptibles d'apparaître sous d'autres formes que leur forme canonique (répertoriée dans un lexique par exemple) comme ***grand-mère*** qui peut être pluralisée sous la forme ***grand(s)-mères***. Si cette flexion en nombre paraît régulière, certaines EP affichent des particularités (*des* ***dos d'âne****(*s)* mais *des* ***toiles d'araignée(s)*** (Savary, 2008)).

Comparée à ces EP nominales, une EP verbale telle que ***prendre décision*** couvre un large ensemble de réalisations possibles (Ex. 3.1-3.12). Malgré cette importante variabilité, il s'agit bien d'une EP car la substitution lexicale de ses composants par un synonyme y est impossible (Ex. 3.13). En effet, ***prendre décision*** ne sous-entend pas l'action de *prendre* au sens propre comme ce serait le cas avec *prendre (= saisir) un stylo*.

(3.1) *J'ai* ***pris*** *une* ***décision***

(3.2) *Il* ***prendra*** *des* ***décisions*** (flexion nominale et verbale)

(3.3) *J'ai* ***pris*** *ma* ***décision*** (déterminant possessif)

(3.4) *J'ai* ***pris*** *un grand nombre de* ***décisions*** (déterminant complexe)

(3.5) *J'ai* ***pris*** *une* ***décision*** *radicale* (adjectif postposé)

(3.6) *J'ai* ***pris*** *une grande* ***décision*** (adjectif antéposé)

(3.7) *Je regrette la* ***décision*** *que j'ai* ***prise*** (relative)

(3.8) *C'est la* ***décision*** *que j'ai* ***prise*** (clivage)

(3.9) *Ma* ***décision*** *est* ***prise****, je pars* (passive avec auxiliaire)

(3.10) *Ma* ***décision prise****, je suis partie* (passive sans auxiliaire)

(3.11) *Cette* ***décision*** *ne se* ***prend*** *pas à la légère* (passive en SE : voix moyenne)

(3.12) *J'ai* ***repris*** *une* ***décision*** *identique* (variation lexicale par dérivation)

(3.13) **J'ai saisi une décision* (variation lexicale par synonymie)

Compte tenu de la diversité des formes de surface qu'une EP peut adopter *via* ses réalisations observées (ou *tokens*), il est nécessaire d'en déterminer la forme canonique afin d'avoir une référence unique pour chaque EP, autrement dit pour chaque *type*, à la façon d'une clé d'index [1]. Cela revient à en sélectionner d'une part les composants requis et non substituables et, d'autre part, la forme la moins marquée avec maintien du sens. Dans ***prendre*** *une* ***décision***, les composants sont uniquement *prendre* et *décision* car le déterminant est modifiable (***prendre*** *cette* ***décision***). La détermination des composants canoniques n'est pas toujours évidente car une même EP à l'oral peut être transcrite différemment (***autant/au temps pour moi*** ⇒ 'admettre avoir commis une erreur') mais de tels cas restent marginaux. Alternativement, l'omission d'un composant (*il*) est fréquente lors du passage à l'oral dans le cas de l'EP ***(il) y a***.

 La forme canonique telle que nous la définissons est une construction artificielle établie d'après la forme canonique des composants, c'est-à-dire bénéficiant de la flexion la moins marquée possible avec maintien de la lecture idiomatique. L'ensemble des réalisations (*tokens*) possibles d'un *type* d'EP – qui demeure une représentation abstraite – pourra ainsi être unifié et représenté de façon concrète grâce à cette seule forme canonique.

La forme canonique des composants sera fréquemment l'infinitif pour les verbes et le singulier pour les noms mais cela ne constitue pas une règle générale. Par exemple, la forme canonique de ***prendre*** *des* ***vacances*** est `prendre vacances` avec le nom au pluriel sinon l'EP perdrait son sens initial. De même, dans ***on aura tout vu*** ⇒ 'avoir vu quelque chose d'original', la forme future du verbe doit être conservée dans la forme canonique, ce qui donne donc `on aura tout vu`.

Cette forme canonique ne correspond pas obligatoirement à une forme valide dans le cadre de l'énonciation : `prendre décision` est une séquence invalide à moins de s'exprimer en style télégraphique. Néanmoins, son intérêt est de neutraliser la variabilité en fusionnant un ensemble varié de réalisations ce qui sera utile pour notre focalisation sur les variantes.

A titre d'exemple, cette forme canonique [2] `prendre décision` constitue un représentant canonique du type de l'EP instancié par les tokens *prendra décisions* (Ex. 3.2) et *pris décision* (Ex. 3.3).

3.2 Variabilité imprévisible des EP

La variabilité des EP est souvent imprévisible, ce qui peut poser problème pour des applications de TAL. Dans le cas de la traduction automatique par exemple, savoir qu'une séquence de tokens peut admettre une lecture idiomatique uniquement sous certaines conditions permet de résoudre l'ambiguïté entre lecture littérale et idiomatique (section 3.2.1). Or, il est impossible de déterminer *a priori* la nature des transformations autorisées.

1. Cela évite des incohérences telles que ***bébé-éprouvette*** référencé à l'entrée *bébé* et ***bébé éprouvette*** à l'entrée *éprouvette* du même dictionnaire (Mathieu-Colas, 1995).
2. Nous revenons sur cette notion dans la section 7.3.

Par ailleurs, les EP ne sont figées ni dans le temps ni dans l'espace (section 3.2.2) ce qui, dans le premier cas, pose problème pour la traduction d'auteurs classiques par exemple.

3.2.1 Variabilité et maintien de lecture idiomatique

La variabilité d'une EP doit être mise en regard avec ses contraintes, autrement dit sa non-variabilité sur certains plans. Cette (non-)variabilité peut être considérée comme une propriété définitoire d'une EP donnée, comme l'impossibilité de modifier le nombre du nom dans ***prendre acte*** ⇒ 'retenir formellement une information', au risque d'en modifier complètement la signification (Ex. 3.14). Les limites de variabilité d'une EP sont illustrées par les différentes idiosyncrasies précédemment répertoriées : morphologique, syntaxique, lexicale, (ortho)graphique. La comparaison de l'expression libre *prendre stylo* (Ex. 3.15) et de l'EP ***prendre décision*** (Ex. 3.16) met en évidence des contraintes identiques sur la nécessité d'un déterminant. A l'instar d'une expression libre, cette contrainte peut cependant disparaître sous réserve que le nom soit au pluriel (Ex. 3.17). En revanche, l'EP ***prendre acte*** (Ex. 3.14, Ex. 3.18) qui possède un patron syntaxique similaire (verbe *prendre* + nom masculin) n'admet qu'une seule construction : l'absence de déterminant et le nom au singulier. Ainsi, même lorsque des transformations sont possibles – c.-à-d. elles conduisent à des énoncés valides –, elles impliquent parfois une modification significative de sens en conférant une lecture littérale, comme dans (Ex. 3.18) par l'ajout d'un déterminant. Il est donc impossible de se fonder sur le seul critère de patron syntaxique pour prédire les emplois tolérés ou non pour une EP donnée.

(3.14) *J'ai **pris** ∅/#un/#mes **acte**(s) du verdict*

(3.15) *J'ai pris *∅/un/mes stylo(s)*

(3.16) *J'ai **pris** *∅/une/mes **décision(s)***

(3.17) *Depuis qu'il est chef, il **prend décisions** sur décisions*

(3.18) *J'ai pris un acte de naissance à la Mairie*

Quant à l'EP ***apporter des oranges***, tandis que l'exemple 3.19 demeure ambigu, il suffit de modifier le nombre du nom pour entraîner une lecture littérale (Ex. 3.20).

(3.19) *Je t'**apporterai des oranges***

(3.20) *Je t'apporterai une orange*

Par ailleurs, même quand une EP tolère une certaine variabilité lexicale, cela peut concerner uniquement certains composants sans qu'il soit possible de prédire lesquels. Par exemple, ***manger les pissenlits par la racine*** ⇒ 'être mort et enterré' conserve une lecture idiomatique par substitution du verbe par un synonyme (***bouffer les pissenlits par la racine***) tandis que remplacer *pissenlit* par *dent-de-lion* n'autorise plus que la seule lecture littérale.

3.2.2 Variabilité diachronique et diatopique

Comme tout objet linguistique, les EP sont susceptibles d'évoluer dans le temps et l'espace. Un locuteur donné peut donc y être confronté face à des productions orales ou

écrites dont il n'est pas l'auteur. En matière de diachronie, la plupart des EP comportant des *cranberry words* étaient compréhensibles à l'époque où elles se sont popularisées comme dans le cas de ***crier haro sur le baudet*** ⇒ 'exprimer sa révolte envers un individu ou quelque chose' ou ***prendre la poudre d'escampette*** ⇒ 's'enfuir'. Mais, en tombant en désuétude, les mots *haro/escampette* rendent désormais ces EP inintelligibles d'un point de vue compositionnel. Cette désuétude peut même conduire à la disparition totale d'une EP, phénomène qualifié de *nécrologie* par Drouin et Dury (2009) :

> the disappearance of a term, the disappearance of a part of term, a change in grammatical status, and/or the disappearance of a meaning over a given period of time. [3]

S'intéresser à la nécrologie des termes est d'autant plus intéressant que leur appartenance à des domaines spécialisés les rend vulnérables face au phénomène d'obsolescence. Cette obsolescence s'observe également dans la langue générale : l'EP ***petit bleu*** ⇒ 'télégramme' a ainsi disparu, le télégramme étant supplanté par d'autres moyens de communication.

D'autres expressions ont évolué au fil du temps comme l'EP ***être (heureux) comme un poisson dans l'eau*** ⇒ 'être heureux' qui, jusqu'au XVII^{e} siècle, était utilisée sous la forme contraire ***être comme le poisson hors de l'eau*** ⇒ 'être malheureux' (Haßler et Hümmer, 2005). De nouvelles expressions apparaissent continuellement et il est impossible de prévoir quelles nouvelles EP sont susceptibles de voir le jour. Cette imprédictibilité est d'autant plus forte que certaines utilisent des néologismes comme dans ***faire avancer le schmilblick*** ⇒ 'faire avancer un sujet', d'après un jeu où il s'agissait de deviner la nature d'un objet surnommé le *schmilblick*.

Du point de vue diatopique, un référent unique peut être associé à plusieurs signifiants polylexicaux selon les régions : on parle de ***crayon à papier*** en Normandie, de ***crayon de papier*** en Bourgogne, de ***crayon de bois*** en Vendée, de ***crayon gris*** sur la Côte d'Azur et de ***crayon mine*** dans la Marne. La diatopie intervient également entre le français de France métropolitaine et d'autres variantes de français : l'expression ***jeter l'éponge*** ⇒ 'abandonner', qui n'admet aucune variabilité lexicale en métropole, dispose de plusieurs équivalents québécois ***jeter/lancer l'éponge/la serviette***.

A la croisée de la variation diatopique et diastratique, on trouve le "parler de banlieue" avec des expressions telles que ***se capter*** ⇒ 'se rejoindre' ou ***être en pit*** ⇒ 'être seul', adaptée du créole ***être en chien*** (Rey et la Peste, 2007).

Les variabilités diachronique et diatopique recouvrent des modes d'expression partagés par un ensemble de locuteurs. Or, la variabilité peut aussi s'observer au niveau de chaque individu qui, sciemment (jeux de mots pour les EP très fréquentes) ou non (par méconnaissance de l'EP), détourne des EP. Ce processus est qualifié de *défigement* (Haßler et Hümmer, 2005), comme le fait de qualifier un concert de Patti Smith de *seringue sur le gâteau*[4] (vs. ***cerise sur le gâteau*** ⇒ 'l'avantage supplémentaire') en référence à son passé sulfureux. Ce défigement résulte parfois de la méconnaissance de composants de l'expression comme dans (Ex. 3.21) : l'ivraie est une "mauvaise" herbe et l'opposition relève du contraste entre ce qualificatif sous-entendu et l'adjectif *bon* – qu'il ne faut donc

3. la disparition d'un terme, la disparition d'une partie du terme, un changement de statut grammatical, et/ou la disparition d'un sens sur un laps de temps donné [Notre traduction].

4. *Le Monde*, 25/10/2014.

pas omettre comme dans l'exemple 3.22 – dans la séquence *bon grain*. Des EP compositionnelles peuvent aussi subir ce défigement comme le montre l'adaptation non standard d'un proverbe (Ex. 3.24) comparée à la version initiale (Ex. 3.23). Ici, la substitution d'une plante par une autre n'entrave cependant pas la compréhension du message car sa structure et les verbes utilisés permettent d'en deviner la signification (idiome figuratif transparent).

(3.21) ***Séparer le bon grain de l'ivraie*** ⇒ 'séparer le mal et le bien, les gentils et les méchants'

(3.22) *Certains plats* [...] *ont été élaborés afin de ne pas transgresser l'interdit de* ***séparer le grain de l'ivraie****, c'est-à-dire la chair du poisson de ses arêtes.* (`FR-train1.0`[5])

(3.23) ***Passez-moi la rhubarbe, je vous passerai le séné*** ⇒ 'faisons-nous des concessions mutuelles'

(3.24) *Passe-moi la salade, je t'envoie la rhubarbe* (Nicolas Sarkozy, *Journal Télévisé de France 2*, 07/12/2015)

Cette variabilité diatopique, diachronique et diastratique nécessite de spécifier l'état de langue considéré : s'intéresse-t-on uniquement au français de France dit *standard* contemporain vu comme une référence normative, fondée sur un entre-deux entre parler populaire et registre soutenu (Rebourcet, 2008), ou bien cherche-t-on à disposer d'une vision plus large de la variation des EP ? Certains principes tels qu'une recherche d'économie dans le langage, ou des contraintes phonologiques peuvent exister sur plusieurs aires géographiques, donc mettre en œuvre des évolutions similaires quoique à des rythmes pouvant être différents. Du point de vue du TAL, outre la variabilité diachronique, qui peut poser problème pour la traduction d'œuvres non contemporaines, il nous semble pertinent de prendre en compte la variabilité diatopique. Il est effectivement difficile de garantir la "pureté" de la variante sur laquelle on travaille. Un document en français peut ainsi incorporer des éléments émanant de locuteurs francophones d'autres territoires : le journal *Libération*[6] utilise ainsi comme intertitre l'EP `(BE)` ***"brosser les cours"*** ⇒ 'sécher les cours' au sujet d'une grève lycéenne en Belgique.

Prendre en compte la variabilité diatopique permettrait de développer des systèmes ayant la capacité de s'adapter à différentes variantes d'une même langue, en s'appuyant par exemple sur des mesures de similarité sémantique (section 5.2.2). Certaines EP en français de France et du Canada partagent en effet ce type de similarité : `(FR)` ***jeter l'éponge*** = `(CA)` ***lancer l'éponge*** (verbes synonymes), `(FR)` ***huile de coude*** = `(CA)` ***jus de bras*** (noms sémantiquement proches). Des similarités formelles telles que l'emploi de numéraux pourraient également être exploitées : `(FR)` ***se mettre sur son 31*** = `(CA)` ***se mettre sur son 36***. Par ailleurs, s'intéresser aux EP en diachronie permettrait d'étudier les mécanismes de figement mis en œuvre dans la création de nouvelles EP, notamment dans quel ordre s'opèrent les blocages de variabilité. D'autres variétés de français (de Belgique, de Suisse, du Québec, d'Afrique) témoignent d'ailleurs d'anciennes expressions aujourd'hui inusitées en France (p. ex. `(CA)` ***à cause que*** ⇒ '**parce que**') ou peuvent constituer des calques d'EP étrangères (p. ex. ***fin de semaine*** pour ***week-end***), phénomène susceptible de se produire également en France dans les zones frontalières, ou dans des contextes multiculturels.

5. `FR-train1.0` décrit en section 8.1.3.
6. "A Bruxelles, pourquoi aller à l'école si on n'a pas de futur ?", 31/01/2019

Ces mécanismes de création d'EP par calque ou en lien avec leur figement progressif ne sont pas centraux, même s'ils peuvent indirectement apparaître dans nos données. Notre étude aborde en effet la question de la variabilité des EP d'un point de vue synchronique.

3.3 Continuum de variabilité des EP

La variabilité des EP s'inscrit dans un *continuum* entre formes figées et flexibles plutôt que dans une stricte dichotomie. Seuls 10% des noms composés sont totalement figés d'après Gross (1988a). De même, Tutin (2016) a établi une échelle de variabilité pour les 30 EP françaises les plus fréquentes de type Verbe-(Déterminant)-Nom, en comptant le nombre de transformations observées en corpus parmi les 5 propriétés suivantes :

— pluralisation du nom possible : ***prendre*** *une / des* ***décision(s)***
— variabilité du déterminant : ***prendre*** *une / la / cette / deux /etc.* ***décision(s)***
— construction relative autorisée : *je regrette la* ***décision*** *que j'ai* ***prise***
— construction passive autorisée : *ma* ***décision*** *est* ***prise***
— nom modifiable par un adjectif : ***prendre*** *une importante* ***décision***

De cette façon, chaque EP se voit attribuer un niveau de variabilité, le niveau 0 correspondant à la variabilité minimale (c.-à-d. aucune transformation possible) et le niveau 5 à la variabilité maximale (c.-à-d. les 5 transformations sont possibles). Les niveaux intermédiaires indiquent que l'EP tolère entre 1 et 4 transformations, sans que nous en connaissions la nature. Autrement dit, un même niveau de variabilité pour deux EP n'implique pas que les transformations qu'elles tolèrent sont identiques. Des exemples sont attestés pour chaque niveau : niveau 0 (***donner lieu***), niveau 1 (***faire partie***), niveau 2 (***mettre l'accent***), niveau 3 (***prêter attention***), niveau 4 (***rendre visite***), niveau 5 (***jouer rôle***), ce qui semble confirmer l'existence d'un *continuum* de variabilité. En plus des propriétés énoncées par Tutin (2016), la (non-)variabilité s'observe également sur d'autres plans :

— une flexion verbale limitée : ***qu'importe*** / ***qu'importait*** vs. # *qu'importa*
— la nécessité d'un modifieur non lexicalisé avec pour seule contrainte le fait qu'il soit doté d'une sémantique spécifique, comme l'adjectif mélioratif dans (Ex. 3.25) ou péjoratif dans (Ex. 3.26).
— une variabilité (ortho)graphique dans ***saoul/soûl comme un (P/p)olonais*** ⇒ 'complètement soûl' qui accepte deux graphies de l'adjectif et une variabilité de la casse pour la nationalité.

(3.25) ***Filer le*** *grand/parfait/*médiocre/*Ø* ***amour*** ⇒ 'vivre un amour sans histoire'
(3.26) ***Filer un*** *mauvais/vilain/*parfait/*Ø* ***coton*** ⇒ 'faire de mauvaises affaires'

A l'inverse de cette marge de liberté, d'autres EP ne tolèrent aucune variabilité (Ex. 3.27).

(3.27) ***Honni soit qui mal y pense*** ⇒ 'honte à celui qui y voit du mal'

Les niveaux d'EP mis en évidence par Tutin (2016) soulignent que, parmi les EP de patron VERB-(DET)-NOUN appartenant aux catégories constructions à verbe support et idiomes[7], certaines EP sont davantage variables que d'autres. En effet, les 7 EP de variabilité maximale citées par Tutin (2016) sont exclusivement des constructions à verbe

7. Nous reviendrons sur ces catégories dans la partie II.

support, tandis que les 9 EP de variabilité minimale sont des idiomes dans 89% des cas. Il est toutefois difficile de tirer des conclusions sur les niveaux de variabilité intermédiaires : le niveau 2 comporte par exemple 75% de constructions à verbe support, mais rien ne garantit que les deux propriétés satisfaites par les EP de ce niveau soient comparables d'un type d'EP à l'autre.

Malgré les limites de cette représentation, nous rejoignons Tutin (2016) sur l'existence d'un continuum de variabilité des EP. De fait, si un type d'EP tolère davantage de motifs de variabilité qu'un autre type d'EP, alors son profil sera plus étendu. Or, toutes les EP n'ayant pas les mêmes contraintes, chaque profil sera différent. Nous nous attendons à ce que les profils de variabilité obtenus à partir d'un grand nombre de types d'EP verbales de nature très différente ne se réduisent pas à quelques profils facilement isolables, mais constituent plutôt un continuum de profils. C'est sur l'existence de ce continuum que nous nous appuierons pour estimer non seulement si deux types d'EP ont des propriétés identiques mais également à quel point elles sont similaires.

Une approche alternative, proposée par Sag *et al.* (2002), a pour objectif de classer les EP d'après leurs caractéristiques communes et de fournir une explication sur les différences de variabilité entre EP de classes différentes.

3.4 Niveaux de variabilité des EP

Sag *et al.* (2002) analysent la variabilité (qu'ils qualifient de *flexibility* en anglais) des *phrases lexicalisées* – autrement dit des EP – en les classant par ordre de flexibilité croissante :

- Les *expressions figées* (par exemple ***revenons à nos moutons*** ⇒ 'revenons à notre sujet') n'admettent aucune variation morphosyntaxique : impossibilité d'insertion ou de flexion (#*revenez encore à mon mouton*).
- Les *expressions semi-figées* présentent des contraintes strictes quant à l'ordre des composants mais autorisent certaines variations telles que le choix du déterminant ou la flexion temporelle : *je* ***tire / tirerai mon épingle du jeu*** ⇒ '*je me dégage(rai) adroitement d'une situation délicate*'. Sag *et al.* (2002) utilisent le principe de compositionnalité sémantique défini par Nunberg *et al.* (1994) – c.-à-d. le sens global d'une expression est lié à celui des éléments qui la composent – pour distinguer les *idiomes* (section 2.1.3) *sémantiquement décomposables* et *non-décomposables*. Seuls les idiomes non-décomposables (Sag *et al.*, 2002) feraient partie des expressions semi-figées et se caractériseraient par leur opacité sémantique. Cette opacité impliquerait une variabilité limitée, par exemple une impossibilité de passivation (# *son épingle a été tirée du jeu*).
- Les *expressions syntaxiquement flexibles* : ces expressions ne présentent aucune restriction quant à l'ordre des composants ou aux transformations autorisées (p. ex. passivation, modification adjectivale, etc.).
  Figurent notamment ici les constructions à verbe support (section 2.1.7) et les idiomes décomposables par équivalence sémantique : `(EN)` ***spill the beans*** '*répandre les haricots*'⇒'révéler un secret' avec *spill* = *révéler* et *beans* = *secret*(*s*). La distinction entre idiomes décomposables et idiomes non-décomposables est ce-

pendant remise en question : Sheinfux *et al.* (2018) montrent que l'EP `(EN)` ***to kick the bucket*** '*frapper le seau*'⇒'mourir' traditionnellement utilisée comme exemple d'idiome non-décomposable offre en réalité une marge de flexibilité (variation du déterminant, ajout de modifieur, passive) sous réserve d'utiliser un corpus suffisamment grand (20 milliards de tokens). En effet, une recherche similaire sur un corpus de 350 millions de tokens n'avait fourni que 12 occurrences de cette EP dont une seule non canonique : `(EN)` ***kick*** *their respective* ***buckets*** (Riehemann, 2001).

Malgré l'existence de contre-exemples, l'intérêt de la classification de Sag *et al.* (2002) est de considérer la flexibilité (ou variabilité) comme un critère essentiel de classification des EP. Nous rejoignons cette vision, mais notre objectif n'est pas à proprement parler de classer les EP par niveaux de variabilité, mais de rentrer dans le détail des motifs de variabilité autorisés. Nous supposons en effet qu'une prise en compte de ces motifs de variabilité devrait permettre de distinguer les séquences étant des EP de celles qui n'en sont pas, autrement dit de résoudre leur ambiguïté.

Chapitre 4

Focalisation sur l'ambiguïté : EP vs. non-EP

Le fait qu'un motif de variabilité soit toléré ou non selon l'EP participe de l'hétérogénéité des EP. Dans cette thèse, on suppose que l'ensemble de ces motifs, réunis sous la forme de profils de variabilité, peut constituer un indice pour distinguer des EP de structures libres ressemblantes que l'on qualifierait alors de 'non-EP'. L'ambiguïté entre 'EP' et 'non-EP' est un phénomène polymorphe (section 4.1) qui pose la question de la désambiguïsation (section 4.2).

4.1 Différents types d'ambiguïté

Le fait que le langage soit vecteur d'ambiguïté [1] est une source de difficultés supplémentaires pour les applications de TAL. A l'instar de Nasr *et al.* (2015); Scholivet et Ramisch (2017), nous considérons qu'une EP est ambiguë dès lors que ses composants peuvent apparaître au sein d'une même phrase sans former une EP.

Il existe deux types d'ambiguïté : le premier est lié à une *double lecture littérale/idiomatique*, fréquente pour les métaphores comme ***jeter l'éponge*** dans (Ex. 4.1-4.2), et le second à une *co-occurrence fortuite* des composants sans portée de sens significative (Ex. 4.3) (Savary *et al.*, 2019b). Ces co-occurrences fortuites seront signalées au moyen d'un soulignement par tirets, et les lectures littérales par un soulignement sous forme de vagues, suivant la convention adoptée par Savary *et al.* (2019b).

(4.1) *Il s'est encore fâché avec son chef, c'est alors qu'il a* ***jeté l'éponge***

(4.2) *Il a jeté l'éponge dans l'évier*

(4.3) *Il jette les gants usagés et nettoie l'éponge*

Les EP relèvent à la fois du lexique (par exemple dans le cas des mots composés

1. "Cette ambiguïté, loin d'être marginale, est un de ses traits caractéristiques. On peut d'ailleurs voir là le résultat d'un compromis inévitable entre d'un côté une capacité d'expression quasi illimitée, et de l'autre des contraintes liées à la limitation des ressources physiologiques mises en œuvre (taille de la mémoire à long et court-terme, densité de l'espace phonétique, contraintes articulatoires, etc)." (Yvon, 2010)

sans insertions) et de la grammaire (transformations syntaxiques), ce qui rend d'autant plus complexe leur prise en compte. Certaines EP présentent des idiosyncrasies qui les distinguent de formes libres au premier coup d'œil, qu'il s'agisse de l'omission d'un déterminant (***prendre*** ∅ ***acte***) ou de l'association incongrue préposition-adjectif dans (EN) ***by and large***. Mais la plupart des idiosyncrasies ne sont pas aussi flagrantes, ce qui entraîne des difficultés dans l'identification d'EP par un humain et, *a fortiori*, par une machine.

D'une part, sachant par exemple que ***faire face*** ⇒ 'affronter' est une EP car elle est ainsi identifiée dans un lexique, comment savoir à première vue et en toute certitude que cette séquence forme bien une EP dans une phrase donnée ? Il peut en effet tout aussi bien s'agir d'une EP (*comment* ***faire face*** *à une invasion de termites*) que d'une co-occurrence fortuite (*le match qu'il va faire face à la France*). D'autre part, ne considérer que cette seule séquence ***faire face***, reviendrait à ne jamais considérer des séquences différentes en apparence en raison de flexions temporelles / personnelles ou de discontinuités par exemple (*nous* ***faisons*** *(encore)* ***face*** *à une invasion de termites*) mais qui relèvent cependant du même type d'EP. De plus, contrairement aux entités nommées qui présentent des indices surfaciques (majuscules, titres, etc.), la plupart des EP ne bénéficient pas de tels indicateurs.

4.2 Désambiguïsation

Outre des blocages variationnels (p. ex. un nom au pluriel dans *jeter les éponges*) ou le contexte (Ex. 4.1) qui lèvent l'ambiguïté en orientant vers une lecture respectivement littérale ou idiomatique, les humains désambiguïsent naturellement les énoncés en évaluant leur vraisemblance. Dans *il a* ***bu le bouillon de onze heures*** *dix minutes après s'être couché* ⇒ 'il est mort dix minutes après s'être couché', la lecture *bu le bouillon de 11h10* ne serait pas privilégiée. Si seules les EP contenant exclusivement des *cranberry words* ou des tokens étrangers sont susceptibles d'échapper à l'ambiguïté, le risque d'ambiguïté entre lecture littérale et idiomatique demeure faible d'après l'étude menée par Savary *et al.* (2019b) sur des EP verbales dans cinq langues différentes (basque, allemand, grec, polonais et portugais). La proportion de lectures idiomatiques comparée à l'ensemble des emplois (idiomatiques + littéraux) s'élève à des taux compris entre 96% et 98% selon la langue. Des *taux d'idiomaticité* aussi élevés suggèrent que l'on tend à éviter les tournures relevant de ce genre d'ambiguïté et que, par conséquent, le problème de la désambiguïsation entre lecture littérale et idiomatique n'est pas essentiel. Nous risquerions en effet, lors de la classification, d'introduire du bruit face à un problème résolu dans 96 à 98% des cas. De plus, ce genre d'ambiguïté, critique pour la traduction automatique, ne représente pas toujours un problème pour l'analyse syntaxique (en anglais *parsing*[2]). Le risque d'ambiguïté lié aux co-occurrences fortuites paraît en revanche plus important : il survient dans une proportion variant respectivement de 11% à 52% pour les deux EP ***prendre décision*** et ***prendre l'habit*** (servant d'illustration en section 10.3.2), en raison de la fréquence du verbe *prendre* et du déterminant *le*. Il semble donc plus pertinent de considérer une distinction globale entre 'EP' (lectures idiomatiques) et 'non-EP', sans chercher à savoir si ces dernières

2. Cette tâche a pour objectif de représenter sous forme arborée les relations syntaxiques entre les éléments (ou groupements d'éléments) d'une phrase.

sont des lectures littérales – lorsque l'EP tolère la double lecture littérale/idiomatique – ou des co-occurrences fortuites, relevant simplement de la présence simultanée des composants dans une même phrase.

Ce tour d'horizon des caractéristiques des EP souligne que l'ambiguïté est un problème-clé pour les applications de TAL, et qu'une façon de la résoudre est de bénéficier d'informations sur la variabilité des EP. Résoudre l'ambiguïté entre 'EP' et 'non-EP' permettrait alors de leur consacrer des traitements spécifiques afin d'améliorer les performances des applications. Dans la partie II, nous abordons ces traitements en détail, qu'il s'agisse des ressources disponibles pour les mener à bien, des méthodes existantes, de leurs performances et limites respectives.

Deuxième partie

Du pain sur la planche en matière de traitement automatique des EP

De façon générale, notre intérêt pour les EP se justifie par le fait que, à la fois fréquentes et sujettes à des comportements imprévisibles, leur présence dans la langue dégrade les performances de nombre de tâches et d'applications de TAL. Ces défis se situent sur deux niveaux. D'une part, les caractéristiques mêmes des EP en font des objets particuliers dont il est difficile de brosser le portrait de façon simple et universelle (p. ex. valide aussi bien pour les EP nominales que pour les verbales). D'autre part, les applications de TAL sont, par définition, automatisées. Elles tirent donc parti de régularités du langage, alors que c'est justement l'homogénéité qui fait défaut aux EP. Cette partie II témoigne des difficultés concrètes que cela pose aux tâches de parsing ou de traduction automatique, d'où la nécessité de prise en compte des EP. Nous évoquons ces deux tâches en raison de leur complémentarité : elles reflètent en effet des difficultés distinctes pour la prise en compte des EP, en lien avec l'hétérogénéité de ces EP décrite dans la partie I. La question de l'ambiguïté vis-à-vis de lectures littérales affecte en effet la traduction automatique, tandis que les co-occurrences fortuites peuvent poser problème au parseur.

Chapitre 5

Modélisation et traitement automatique de la variabilité des EP

Comme cette thèse se focalise sur l'identification d'EP connues (c.-à-d. vues dans d'autres contextes), nous dépendons de recensements préalables d'EP pour en identifier des variantes en contexte. On peut donc s'interroger sur la pertinence d'exploiter des ressources informatiques répertoriant des EP et des informations sur leur variabilité (section 5.1). Nous nous intéressons ensuite à la tâche de découverte d'EP dont la finalité est l'enrichissement de lexiques existants (section 5.2). Quoique nous cherchions à identifier des EP connues et non à en découvrir de nouvelles, nous supposons que des travaux portant sur la découverte d'EP grâce à leur variabilité sont susceptibles de nous fournir des informations transférables pour l'identification de variantes (section 5.3), d'autant plus que la découverte d'EP est davantage documentée que leur identification.

5.1 Lexique

Losnegaard *et al.* (2016) ont dressé un inventaire [1] des ressources monolingues et multilingues (corpus annotés, lexiques) utiles pour l'étude des EP. On dénombre 9 ressources pour le français, mais comme certaines excluent les EP verbales, nous nous restreignons à un choix ciblé de ressources afin d'en montrer les avantages et limites : dictionnaires (section 5.1.1) et *Lexique-Grammaire* (section 5.1.2).

5.1.1 Dictionnaires

Le DELAC [2] (Dictionnaire Électronique des mots composés) et le Lefff (Lexique des formes fléchies du français) font partie des dictionnaires électroniques utilisables pour le TAL et comportant des EP. En plus d'être disponibles gratuitement, le premier a été constitué manuellement et le second collecté automatiquement puis manuellement validé, ce qui garantit leur qualité.

1. http://multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets
2. http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/delac.html

La construction du DELAC par le LADL (Laboratoire d'Automatique Documentaire et Linguistique) reposait initialement sur la définition des mots composés de Gross (1988a) et s'est ensuite notablement accrue après l'élargissement de cette définition (Silberztein, 1990). Il contient désormais plus de 100 000 mots composés français (90 000 noms, 15 000 constructions *être*+Prép+Nom, 8 000 adverbes, 500 conjonctions). Cette restriction au seul patron *être*+Prép+Nom est particulièrement pénalisante dans notre cas car aucune des 4550 occurrences d'EP verbales du corpus `FR-train1.1` ne respecte ce patron, non pas en raison de sa rareté, mais parce qu'en réalité l'EP se limite souvent au fragment Prép+Nom, comme dans le cas de *(être)* ***en service*** où le verbe peut être omis : *Le Flak est le premier canon officiel allemand* ***en service***.

Le Lefff[3] comporte quant à lui 536 375 entrées correspondant à 110 477 lemmes, dont 3 295 idiomes verbaux (Sagot, 2010). Plus généralement, il inclut 26 311 entrées composées (dont 22 673 noms) (Constant *et al.*, 2012). Cette ressource serait donc essentiellement utilisable pour les idiomes, mais même pour cette catégorie, la couverture n'est pas complète : par exemple l'EP ***jeter l'éponge*** en est absente.

5.1.2 Tables du Lexique-Grammaire

La constitution du Lexique-Grammaire, ainsi nommé par Maurice Gross (Gross, 1975), repose sur l'idée qu'il est plus pertinent de considérer simultanément le lexique et la grammaire que séparément. Cette approche est particulièrement adaptée aux EP car elles se situent sur ces deux plans. Le Lexique-Grammaire se présente sous la forme de tables et indique, pour chaque entrée lexicale (verbes, EP verbales, noms, adverbes, etc.) figurant sur une ligne donnée, si elle accepte ou non les propriétés ou transformations figurant en colonnes. La catégorie "expressions figées" inclut 39 628 entrées (principalement des expressions verbales et adjectivales) classées selon 276 propriétés.

La Table 5.1 illustre un extrait des informations disponibles pour les EP ***faire équipe*** et ***faire joujou***. Les signes + et - indiquent les propriétés tolérées ou interdites. Le groupe nominal sujet est noté N0, le déterminant (absent ici, d'où l'ensemble vide) Det1, le premier complément régi du verbe C1 et le deuxième groupe nominal complément N2. On voit par exemple que ce complément N2 dans ***faire équipe*** ne peut pas être un non-humain[4], tandis que c'est autorisé pour ***faire joujou*** (*Jean* ***fait joujou*** *avec le ballon*). De même la transformation syntaxique sous la forme *N0 et N2 faire Det1 C1* est autorisée pour ***faire équipe*** (*Jean et Marie* ***font équipe***) mais pas pour ***faire joujou*** (**Jean et le ballon font joujou*).

La limite du Lexique-Grammaire est que cette ressource ne répertorie pas l'intégralité des variations possibles ou interdites, comme le montre l'impossibilité de clivage dans (Ex. 5.2) pour l'EP ***faire l'amour*** par rapport à une expression libre (Ex. 5.1). De même, le fait de savoir si un nom peut être modifié est souvent discriminant pour l'attribution d'une lecture littérale comme dans le cas de *faire le bel amour*[5], mais cette information

3. `http://www.labri.fr/perso/clement/lefff`

4. Nhum désigne un substantif humain (p. ex. *Jean, boulanger*) ou animé (p. ex. *chat*), tandis que N-hum représente un substantif non humain (p. ex. *ballon*).

5. Par exemple dans : *Ce n'est donc pas le mobile – nous parlons du téléphone – qui fait le bel amour. La Dépêche*, 02/03/2018.

<ENT><faire>	<ENT> Det1	<ENT> C1	<ENT> Prep2	N2 := Nhum	N2 := N-hum	N0 := Nhum	N0 := N-hum	N0 et N2 faire Det1 C1
faire	∅	équipe	avec	+	-	+	-	+
faire	∅	joujou	avec	+	+	+	-	-

TABLE 5.1 – Extrait du Lexique-grammaire pour les EP ***faire équipe*** et ***faire joujou***.

n'apparaît pas dans le Lexique-Grammaire.

(5.1) *Jean et Marie ont fait le gâteau → voici le beau gâteau qu'ont fait Jean et Marie*

(5.2) *Jean et Marie ont* ***fait l'amour*** *→ *voici le bel amour qu'ont fait Jean et Marie*

En conclusion, même de grande taille, les ressources à notre disposition demeurent parcellaires, à la fois en terme de couverture d'EP et d'informations disponibles sur leur variabilité. De plus, même si le Lexique-Grammaire permet de suggérer de nouvelles entrées ou propriétés, la mise à jour de cette ressource n'est pas garantie : une expression relativement récente telle que ***se taper des barres*** ⇒ 'rire' n'y figure d'ailleurs pas, d'où l'intérêt des techniques de découverte automatique d'EP.

5.2 Découverte

Cette section décrit la tâche de découverte d'EP (section 5.2.1), les méthodes utilisées (section 5.2.2), dont certaines s'appuient sur la variabilité (section 5.2.3). L'évaluation des données produites est présentée en section 5.2.4.

5.2.1 Définition

La tâche de découverte automatique d'EP consiste à tirer parti d'un corpus textuel pour générer une liste d'EP potentielles (dites EP *candidates*). Ces EP candidates peuvent ensuite être validées manuellement afin d'enrichir des ressources lexicales (Constant *et al.*, 2017). Compte-tenu de cette intervention humaine, il est contre-productif de proposer de multiples emplois d'une même EP (p. ex. ***prendre / prendrai / prenions***/ *etc.* ***décision/décisions***) puisqu'un seul emploi attesté suffit (Constant *et al.*, 2017). La découverte d'EP nécessite donc de s'affranchir de la variabilité observée en corpus en fusionnant les variantes de chaque type d'EP, de sorte que les experts humains n'aient à (in)valider qu'un candidat.

L'enrichissement de lexiques par découverte automatique d'EP trouve son origine dans deux difficultés majeures :

— répertorier manuellement l'intégralité des EP existantes à un instant donné constituerait une tâche chronophage,

— les EP évoluent au fil du temps (nouvelles EP, modification de composants), ce qui impose des mises à jour régulières du lexique.

La découverte d'EP est étudiée depuis les années 1980 (Church et Hanks, 1990; Choueka, 1988). Elle se limite parfois à des EP spécifiques (p. ex. noms composés pour Salehi *et al.* (2015) ou constructions verbe-nom pour Fazly *et al.* (2009)) et repose sur différentes approches, certaines d'ordre général (section 5.2.2), d'autres tirant profit d'éléments linguistiques (section 5.2.3). L'existence de différents types de non-compositionnalité explique le nombre et la diversité des méthodes existantes.

5.2.2 Méthodes générales de découverte

Si l'on générait automatiquement des EP par compositionnalité sémantique, on proposerait de façon indifférenciée des EP attestées (***cabine de douche***) ou non (**loge de douche, *couchette de douche*), ce phénomène étant qualifié d'*overgeneration* (Sag *et al.*, 2002). A cela s'ajoute l'impossibilité de prédiction du sens des séquences générées lorsqu'il s'agit d'EP non sémantiquement compositionnelles. D'autres façons de procéder sont donc privilégiées. La découverte d'EP s'appuie fréquemment sur des patrons morphosyntaxiques (p. ex. Verbe-∅-Nom) pour générer des listes de candidats (***porter***$_{\text{VERB}}$ ***plainte***$_{\text{NOUN}}$, ***mettre***$_{\text{VERB}}$ ***fin***$_{\text{NOUN}}$, etc.). Comme le rappellent Constant *et al.* (2017), la découverte d'EP repose aussi sur certaines de leurs propriétés telles que leur co-occurrence statistique ou leur non-compositionnalité sémantique, laquelle impliquant qu'elles bénéficient rarement de traduction littérale :

— **Méthodes statistiques** Ces méthodes cherchent à découvrir des collocations grâce à des mesures d'association statistique, telles que l'information mutuelle I (ou PMI, de l'anglais *Pointwise Mutual Information*), car deux composants c_1 et c_2 d'une EP tendent à co-occurrer davantage que ne le laisse supposer la probabilité d'emploi P de chaque composant considéré individuellement (Church et Hanks, 1990), d'où une valeur I plus élevée qu'entre deux mots quelconques :

$$I(c_1, c_2) = log_2 \frac{P(c_1, c_2)}{P(c_1)P(c_2)} \tag{5.3}$$

Pour des EP constituées de plus de 2 composants, qui représentent un nombre non négligeable d'expressions (26% des types d'EP dans le corpus `FR-train1.1`), des mesures d'association nouvelles ou adaptées de mesures existantes sont également proposées (Dias *et al.*, 2000; Banerjee et Pedersen, 2003).
Le *mwetoolkit* (Ramisch, 2015), est un exemple d'outil multilingue de découverte en accès libre [6] s'appuyant sur des méthodes statistiques. Il se présente sous la forme de scripts Python et permet de découvrir et d'identifier des EP dans un corpus annoté en parties de discours grâce à différentes mesures d'association (PMI, Dice [7], etc.). Dans un premier temps, l'utilisateur définit un patron syntaxique devant être respecté par les EP à extraire (p. ex. Verbe-∅-Nom). Une fois extraites, elles sont

6. `http://mwetoolkit.sourceforge.net`
7.
$$Dice(c_1, c_2) = \frac{2P(c_1, c_2)}{P(c_1) + P(c_2)} \tag{5.4}$$

filtrées selon des critères de fréquence en corpus (suppression des hapax par exemple) puis classées d'après les différentes mesures d'association.

— **Méthodes distributionnelles** Comme l'affirme Firth (1957) : " *You shall know a word by the company its keeps*" [8]. Le principe de sémantique distributionnelle peut ainsi donner lieu à une représentation des mots sous forme vectorielle (*plongement de mots*) : chaque mot d'un texte est alors représenté sous la forme d'un vecteur de nombres établi d'après les mots qui l'entourent (c.-à-d. son contexte). A titre d'exemple, si l'on considère le proverbe ***plus on est de fous, plus on rit***, l'entourage de chaque composant lemmatisé dans une fenêtre donnée (p. ex. limitée à un seul mot avant et après) peut être représenté par une matrice de co-occurrences en fonction de l'ensemble des mots de cet énoncé :

$$\begin{bmatrix} & plus & on & être & de & fou & rire \\ plus & 0 & 2 & 0 & 0 & 1 & 0 \\ on & 2 & 0 & 1 & 0 & 0 & 1 \\ être & 0 & 1 & 0 & 1 & 0 & 0 \\ de & 0 & 0 & 1 & 0 & 1 & 0 \\ fou & 1 & 0 & 0 & 1 & 0 & 0 \\ rire & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Cette matrice, une fois enrichie par d'autres (très nombreux) exemples pour devenir représentative, peut se décomposer en plusieurs vecteurs de nombres : un pour chaque mot en fonction des autres mots du corpus d'exemples. Ce vecteur, dont la taille égale le nombre de mots différents du corpus, est ensuite compressé (technique dite de réduction de dimensionnalité) et prêt à l'emploi pour des mesures de similarité sémantique. La similarité sémantique entre deux mots est alors évaluée d'après la proximité angulaire de leurs vecteurs associés : plus deux vecteurs de mots sont proches, plus ces mots sont sémantiquement similaires. Cette méthode met en évidence des mots dits *associés* : des synonymes, mais aussi des antonymes (*plaignant* et *inculpé* sont susceptibles d'apparaître dans des contextes semblables) ou des hyperonymes (*bâtiment* vs. *maison*). Cette approche donne lieu à deux utilisations diamétralement opposées : soit une recherche de dissimilarité sémantique, soit une recherche de similarité sémantique.

Dans le premier cas, comme le sens d'une EP n'est pas toujours déductible du sens individuel de chacun de ses composants, une faible affinité sémantique devrait s'observer entre les composants d'EP sémantiquement non-compositionnelles et le contexte où elles apparaissent (Salehi *et al.*, 2015; Katz et Giesbrecht, 2006). De façon concrète, cela revient à évaluer la similarité entre le vecteur global de l'EP et celui reconstitué mathématiquement (par addition ou multiplication (Salehi *et al.*, 2015)) d'après les vecteurs de chacun des composants individuels de l'EP : plus cette similarité est élevée, plus l'EP serait jugée comme étant compositionnelle. Il semble délicat de décréter qu'une EP est (non-)compositionnelle sur la base de ce score de similarité, mais on s'attend à des tendances de plus ou moins forte compositionnalité selon les différents types d'EP. A titre d'exemple, l'EP non-compositionnelle ***jeter l'éponge*** pourra apparaître dans des contextes différents de *éponge*, *Spontex*® ou

8. On connaît un mot par son entourage [Notre traduction].

serpillière comme l'illustre la phrase suivante : *Présidentielle en Égypte : un autre candidat* ***jette l'éponge***, alors qu'une EP semi-compositionnelle comme ***vin gris*** devrait apparaître dans des contextes similaires à ceux de *vin*, d'où un score de compositionnalité *a priori* plus élevé pour cette dernière. Cette méthode nécessite néanmoins des corpus volumineux pour l'acquisition d'informations sémantiques fiables. Ce type d'approche est exploité dans une extension du *mwetoolkit* (Cordeiro *et al.*, 2016a) : outre des mesures d'association pour classer les expressions candidates extraites, un score de compositionnalité sémantique permet de mettre en évidence des EP non compositionnelles sémantiquement.

Dans le second cas, la proximité sémantique permet la découverte en corpus de synonymes de termes multi-mots sémantiquement compositionnels [9] . Pour cela, partant de termes recensés dans un lexique, la substitution de composants par des mots sémantiquement proches permet de générer des synonymes, l'objectif étant d'enrichir ce lexique initial par de nouvelles entrées, par exemple ***machine à induction*** / ***générateur à induction*** ou bien encore ***frein à disque*** / ***frein mécanique*** / ***frein aérodynamique***, ces derniers mettant en jeu une variation Nom de Nom / Nom Adj (exemples cités par Hazem et Daille (2014)).

— **Méthodes hybrides** Ces méthodes exploitent des mesures statistiques et la similarité contextuelle pour apprendre de façon supervisée ce qui distingue des EP attestées de combinaisons libres, à l'instar de Lapata et Lascarides (2003) pour les noms composés.

— **Traduction non-littérale** La découverte d'EP peut être facilitée par l'utilisation de corpus bilingues. L'hypothèse sous-jacente est que l'alignement entre deux textes de langues différentes devrait mettre en évidence (i) une convergence au niveau des expressions (libres/polylexicales) compositionnelles, (ii) une divergence au niveau des EP non-compositionnelles. Dans ce dernier cas en effet, comme beaucoup d'EP ne sont pas traduisibles mot-à-mot, toute traduction non-littérale comme ***avoir un chat dans la gorge*** ⇒ 'être enroué' → `(EN)` ***to have a frog in one's throat***, où *chat* ≠ `(EN)` *frog* (*grenouille*), est susceptible de favoriser la découverte d'EP (Tsvetkov et Wintner, 2014). Une extension de cette méthode est la recherche d'un déséquilibre entre le nombre de composants de la langue source et ceux la langue cible (*n-to-m translations*, noté `n:m`). Un cas particulier est le fait qu'une EP multi-mot s'aligne [10] avec un mot isolé, ce qui est nommé le *one-to-many translations* (noté `1:m`). Par exemple, les deux composants pleins [11] ***peuplement forestier*** et ***essence d'ombre*** se traduisent respectivement par un seul mot `(EN)` *crop* ou par une EP formée de trois composants pleins `(EN)` ***shade tolerant species*** (exemples cités par Morin et Daille (2010)).

Cette méthode de découverte d'EP a été adoptée par Caseli *et al.* (2010) sur des corpus bilingues anglais-portugais. Zarrießet Kuhn (2009) exploitent les traductions `1:m` pour la découverte d'EP verbales dans des corpus parallèles alignés allemand-

9. Ces synonymes de termes complexes sont rares compte-tenu de la volonté d'associer de façon non ambiguë une notion à un terme et vice-versa (Hazem et Daille, 2014).

10. L'alignement de corpus bilingues repose sur la recherche de correspondances entre les mots des deux langues.

11. Les composants pleins excluent les prépositions, articles, etc.

anglais [12]. Ils émettent l'hypothèse que l'idiosyncrasie d'une EP est rarement la même d'une langue à l'autre. Les cas de non-alignement dûs à la variation de flexion verbale sont écartés, car non pertinents (p. ex. `(DE)` *verschlimmern* ↔ `(EN)` *is aggravating*). En plus des catégories habituelles d'EP (LVC, idiomes, VPC), des paraphrases sont également découvertes par les traductions `1:m`, or il ne s'agit pas à proprement parler d'EP.
Face au bruit parmi les EP découvertes par les traductions `n:m`, Tsvetkov et Wintner (2012) signalent que tout désalignement n'est pas nécessairement révélateur d'une EP : cela peut être dû à une traduction incomplète (p. ex. le jour de la semaine dans l'alignement exposé sur la Fig. 6.8, page 98) ou bien s'expliquer par une différence imputable à la langue : "the source language lexically realizes notions that are realized morphologically, syntactically or in someother way in the target language" [13] (p. ex. la préposition *en* traduite par *to the* dans la Fig. 6.8).
Tsvetkov et Wintner (2012) se sont penchés sur les cas où un alignement parfait `1:1` échouait dans un corpus parallèle hébreu-anglais. Ces séquences non-alignées, considérées comme de potentielles EP, ont ensuite été classées d'après leur score de mesure d'association calculé dans un corpus monolingue. 99 des 100 premiers candidats ont été confirmés manuellement comme étant des EP de différentes catégories (entités nommées polylexicales, composés Nom-Nom ou Nom-Adj), ce qui équivaut à une précision de 99%.

Il est également possible de tirer parti de la (non-)variabilité des EP pour favoriser leur découverte.

5.2.3 La variabilité au service de la découverte

Certains blocages au niveau de la compositionnalité des EP peuvent servir d'indices pour leur découverte (Constant *et al.*, 2017). Ces blocages se manifestent par la non-variabilité (également nommée *figement*) lexicale et/ou (morpho)syntaxique des EP ainsi que par l'existence de traductions non-littérales.

5.2.3.1 Non-variabilité lexicale

Les EP interdisent généralement le remplacement d'un de leurs composants par un synonyme (Pearce, 2001) : cela produit soit une expression invalide (***prendre** une **décision*** → **attraper une décision*), soit une expression n'admettant qu'une lecture littérale (***poser un lapin*** → *poser un lièvre*). Partant de ce constat, Fazly *et al.* (2009) estiment que la fréquence de combinaisons Verbe-Nom figées devrait être supérieure à celle de leurs variantes générées par substitutions. Une combinaison Verbe-Nom est alors considérée comme figée si la force d'association calculée entre composants est bien plus élevée que celle de la plupart des variantes générées.

12. Un corpus est dit *parallèle* s'il contient un même texte dans au minimum deux langues. La *Pierre de Rosette* en est un célèbre exemple puisque le même décret y est gravé en grec ainsi qu'en égyptien hiéroglyphique et démotique.

13. La langue source réalise lexicalement des notions réalisées morphologiquement, syntaxiquement, etc. dans la langue cible [Notre traduction].

5.2.3.2 Non-variabilité (morpho)syntaxique

En plus de la mesure de figement lexical, Fazly *et al.* (2009) ont défini une mesure de figement morphosyntaxique pour des EP en anglais de patron verbe-nom. Ils mentionnent trois types de variations morphosyntaxiques fréquents dans les séquences littérales mais rares dans les idiomes :

- variabilité du déterminant : impossibilité de variation (p. ex. article non substituable par un possessif) comme dans ***jeter l'éponge*** vs. *jeter une/cette/son/... éponge*, ajout ou suppression impossible (p. ex. ***donner lieu*** vs. *donner le lieu* ou ***jeter l'éponge*** vs. *jeter* ∅ *éponge*). Mais cette variation du déterminant peut être biaisée par le sémantisme du nom. Les parties du corps humain sont ainsi fréquemment associées à un possessif (***prendre ses jambes à son cou*** ⇒ 'se sauver'), sans qu'il soit possible de les remplacer par un autre déterminant (**prendre les jambes à ce cou*).
- passivation : lorsque le nom d'une EP n'a pas de lien sémantique avec le sens global de l'EP, il est souvent impossible de mettre l'accent sur lui, ce qui implique que la passivation est interdite [14] : ***faire l'objet*** ≠ *l'objet a été fait*.
- variabilité morphologique du nom : comme pour la passivation, la perte de sens du nom interdit souvent la flexion en nombre : ***poser un lapin*** ≠ *poser des lapins*.

Pour la découverte d'EP, Fazly *et al.* (2009) utilisent une liste de 28 verbes anglais "basiques" [15] et cherchent des EP candidates sous la forme de co-occurrences Verbe-Nom – où le nom est l'objet du verbe [16] – dans un corpus parsé. Ces candidats sont ensuite filtrés par les mesures de figement lexical et morphosyntaxique combinées sous la forme d'une mesure d'idiomaticité. Cette méthode donne des résultats intéressants car les 20 candidats les mieux classés sont attestés dans un dictionnaire comme étant effectivement des EP.

Bannard (2007) propose une approche plus globale étant donné qu'il ne se restreint pas à une liste prédéfinie de verbes pour l'extraction des combinaisons Verbe-Nom dans un corpus anglais parsé. Il mesure la fréquence de trois types de variation syntaxique : deux sont communs avec Fazly *et al.* (2009) (passivation, variabilité du déterminant c.-à-d. substituable, effaçable ou ajoutable) tandis que le troisième concerne la possibilité de modification du nom (p. ex. *jeter l'éponge sale*) [17]. Contrairement à Fazly *et al.* (2009), le figement morphologique ne fait pas partie de ses critères d'évaluation. Par ailleurs, il prend en compte les spécificités du verbe ou du nom, susceptibles d'avoir une influence sur la capacité de passivation ou d'accepter des modifieurs adjectivaux par exemple, ce qui lui permet *in fine* d'attribuer une valeur de flexibilité syntaxique pour chaque paire Verbe-Nom. L'auteur souligne une meilleure précision de ses résultats par rapport à l'emploi de mesures d'association lexicale.

Grâce à la découverte de 373 EP menée par Zhang *et al.* (2006), la grammaire d'un parseur atteint une couverture de 18,7% sur un échantillon de phrases en anglais riches en

14. Ce n'est toutefois pas systématique comme l'illustre l'EP ***tourner la page*** : *la* ***page*** *a été* ***tournée***.

15. Ces verbes dits basiques font référence à des états ou des actes au cœur de l'expérience humaine (p. ex. *have*, *take*). Ils sont à la fois fréquents, très polysémiques et se combinent fréquemment avec d'autres mots pour former des combinaisons idiomatiques (Fazly *et al.*, 2009) [Notre traduction].

16. Cette condition est satisfaite dans *Jean a* ***cassé sa pipe***$_{objet}$, mais pas dans *sa pipe*$_{sujet}$ *a cassé*.

17. Ce choix de variations s'appuie sur une analyse manuelle en corpus par Riehemann (2001) du comportement des EP idiomatiques de patron VERB-(DET)-NOUN du *Collins Cobuild Dictionary of Idioms*.

EP (donc problématiques pour le parseur) au lieu de 4,3% avant l'ajout de ces nouvelles entrées polylexicales.

5.2.4 Évaluation

5.2.4.1 Méthodes d'évaluation

Pour évaluer la performance de découverte d'EP, on peut, comme signalé par Constant *et al.* (2017), s'appuyer sur (i) le jugement humain qui valide ou non chaque EP figurant sur la liste d'EP découvertes, (ii) la comparaison avec un lexique d'EP pré-existant, (iii) une évaluation extrinsèque qui évalue indirectement la validité des EP découvertes grâce à la mesure du gain de performances de tâches annexes utilisant cette liste d'EP (parsing ou traduction automatique par exemple). Comme le soulignent Ramisch *et al.* (2012), la qualité des systèmes de découverte d'EP peut aussi être évaluée d'après leur efficacité computationnelle ou leur couverture potentielle par comparaison avec des ressources lexicales, certains systèmes restreignant le nombre de mots des séquences extraites ou n'autorisant pas la découverte d'EP discontinues.

5.2.4.2 Métriques d'évaluation : précision, rappel, F-mesure

L'un des modes d'évaluation de la tâche de découverte repose sur la comparaison entre une liste d'EP obtenue automatiquement et une liste de référence obtenue manuellement (lexique). La performance de découverte est alors évaluée de façon standard avec les mesures de précision, rappel et F-mesure. Ces trois mesures ont des valeurs entre 0% (qualité minimale) et 100% (qualité maximale). Toutefois, le rappel est rarement pris en compte car il suppose que la liste de référence est exhaustive, ce qui n'est presque jamais le cas. Pour la découverte d'EP, la mesure de précision est donc privilégiée à celle du rappel – et par conséquent à la F-mesure –, cette dernière étant davantage adaptée à l'identification d'EP. A titre d'exemple, la Table 5.2 compare une liste d'EP candidates découvertes de façon automatique avec un lexique d'EP servant de référence.

La précision représente la proportion d'éléments extraits à juste titre comme étant des 'EP' par le système. Ici, les deux tiers de cette liste extraite automatiquement coïncident avec le lexique, d'où une précision $P = 66\%$. Le rappel représente quant à lui la proportion d'éléments correctement extraits (ici deux) par rapport à tous les éléments qu'il aurait fallu extraire (ici quatre), soit un rappel $R = 50\%$. La F_1-mesure F_1 (désormais simplement notée F) combine le rappel et la précision sous la forme d'une moyenne harmonique :

$$F = \frac{2P * R}{(P + R)} = 57\% \tag{5.5}$$

Nous adoptons désormais comme convention d'écriture de ces trois métriques (P,R et F) l'omission du signe %, c'est-à-dire ici $F = 57$.

Tandis que la découverte d'EP se focalise sur des types d'EP (***prendre décision*** est un type et ***jeter l'éponge*** en est un autre), l'identification constitue une tâche connexe qui porte sur les tokens d'une EP donnée (***jeta/jetterons l'éponge*** sont deux tokens de

Liste d'EP candidates découvertes	Liste d'EP dans un lexique
prendre décision	***prendre décision***
jeter l'éponge	***jeter l'éponge***
mettre en service	***faire référence***
	donner lieu

TABLE 5.2 – Comparaison entre les EP verbales candidates découvertes automatiquement et un lexique de référence. Le candidat *mettre en service* n'est pas, de notre point de vue, une EP selon l'argumentation développée dans la section 5.1.1.

l'EP ***jeter l'éponge***) et pour lesquels on cherche à déterminer si, dans un contexte donné, ils relèvent bien d'EP. L'identification d'EP sera l'objet du chapitre 6.

5.3 Méthodes transférables à l'identification d'EP

Parmi les méthodes employées pour la découverte d'EP, on s'interroge sur le fait que certaines puissent être également exploitées pour leur identification. En fait, l'un des défis de l'identification est de résoudre l'ambiguïté. Toutes les méthodes de découverte qui permettraient de ne jamais y être confrontés seraient de ce fait parfaitement adaptées pour l'identification.

Les mesures d'association statistique répondraient à ce critère uniquement dans le cas de collocations mettant en œuvre des *cranberry words*. Si un nom tel que *prétentaine* apparaît uniquement dans des contextes où il côtoie le verbe *courir*, alors on peut supposer que (i) il s'agit d'une EP, (ii) il faut l'identifier de façon systématique. Dans les autres cas, cette méthode ne nous semble pas pertinente pour résoudre avec certitude des ambiguïtés potentielles.

Les mesures distributionnelles pourraient permettre d'augmenter le stock d'EP connues par recherche de synonymes par exemple : ***manger/bouffer les pissenlits par la racine***. Mais nos travaux préliminaires révèlent que cette méthode s'avère peu fructueuse, non seulement d'un point de vue quantitatif (peu de candidats générés sont valides), mais également qualitatif (certains candidats valides ne sont pas des variantes sémantiques).

La recherche de traductions non littérales nécessite quant à elle des corpus bilingues alignés, or nous ne disposons pas d'une telle ressource annotée en EP. C'est pourquoi la méthode qui semble la plus prometteuse pour l'identification d'EP est l'exploitation des blocages morpho-syntaxiques *via* le concept de profil de variabilité.

Chapitre 6

Identification d'EP

Ce chapitre décrit la tâche d'identification : son principe, les méthodes traditionnellement utilisées, les défis à relever (section 6.1). Cette tâche est abordée conjointement avec deux applications de TAL : la traduction automatique et le parsing (section 6.2). Nous abordons la traduction car elle se heurte à l'un des défis de l'identification : la gestion de l'ambiguïté entre lecture idiomatique et lecture littérale conditionne en effet la validité d'une traduction comme évoqué dans la section 5.2.2. Le parsing nous intéresse également car il est aussi question d'ambiguïté, mais ici entre lecture idiomatique et co-occurrences fortuites. De plus, comme nous disposons de corpus parsés, nous devons nous assurer que les informations syntaxiques disponibles n'introduisent pas de biais pouvant pénaliser une identification ultérieure.

6.1 État de l'art sur la tâche d'identification

Cette section définit le principe de l'identification d'EP (section 6.1.1), les méthodes utilisées pour mener à bien cette identification (section 6.1.2) ainsi que l'évaluation de l'annotation en corpus produite (section 6.1.3).

6.1.1 Définition

La tâche d'identification a pour objectif de mettre automatiquement en évidence les EP présentes dans un texte, en les balisant par exemple, et en précisant éventuellement leur catégorie : par exemple *construction à verbe support, idiome*, etc. pour les EP verbales ou bien *personne, lieu, organisation*, etc. pour la reconnaissance d'entités nommées (Savary *et al.*, 2017). L'identification d'EP se distingue donc de la découverte par sa finalité : l'annotation de corpus et non la génération d'une liste d'EP.

Cette tâche est cruciale pour différentes applications de TAL, entre autres : parsing sémantique, résumé automatique, analyse de sentiments, extraction d'informations, systèmes de dialogue, etc. Les difficultés liées à la mise en œuvre de la tâche d'identification sont multiples puisqu'il faut (i) signaler toutes les EP présentes dans un texte, (ii) ne pas signaler des non-EP. Parmi les exemples (Ex. 6.1-6.3), seules les deux séquences signalées dans le

premier exemple répondent à ce critère, celles des exemples suivants étant respectivement une lecture littérale et une co-occurrence fortuite.

(6.1) *Après une dispute avec son chef, il* ***jette l'éponge*** *et* ***s'en va****.*

(6.2) *Une fois chez lui, il jette l'éponge dans l'évier.*

(6.3) *Ensuite, il jette la serpillière à côté de l'éponge.*

En définitive, l'identification d'EP se heurte aux trois difficultés précédemment évoquées (Constant *et al.*, 2017) :

- l'ambiguïté (section 4.1), qui s'observe également pour les EN (qualifiée d'homonymie) : une même forme de surface telle que ***Saint-Laurent*** peut faire référence à des entités distinctes (personne, fleuve, etc.).
- la variabilité des EP (section 3.2) qui rejoint la question des données non-vues en *machine learning* : des éléments dont la forme de surface diffère des données vues durant l'entraînement du modèle sont généralement plus difficiles à identifier que celles ayant une forme de surface identique (Augenstein *et al.*, 2017). Cette variabilité (qualifiée de synonymie) pose également problème pour la reconnaissance d'EN : ***U.S.*** = ***U.S.A.*** = *America*. Notons cependant que certains travaux de REN (reconnaissance d'entités nommées) excluent les cas de chevauchement (Sang et De Meulder, 2003). Dans le domaine biomédical, Zhou *et al.* (2007) s'appuient sur des thesauri pour prendre en compte la variabilité : synonymie, hyponymie et hyperonymie. Ils qualifient par ailleurs de variantes lexicales les différences portant sur la forme de surface : graphie (espace, trait d'union) et abréviations [1].
- la discontinuité potentielle des composants (section 6.2.1.2), qui pose problème aux systèmes reposant sur un étiquetage séquentiel (voir section 6.1.2.2).

6.1.2 Méthodes d'identification

Certains travaux consacrés à l'identification d'EP se focalisent parfois sur des EP spécifiques. Fazly *et al.* (2009) se sont ainsi intéressés aux EP verbales idiomatiques et Vincze *et al.* (2013) aux EP de catégorie CVS, incluant des nominalisations (***prendre décision*** → ***preneur de décision***). Les travaux de Nissim et Zaninello (2013) portent quant à eux exclusivement sur les EP nominales. La focalisation peut aussi porter sur les EP d'une langue spécifique : français (Pasquer *et al.*, 2018c), anglais (Fazly *et al.*, 2009; Vincze *et al.*, 2013), italien (Nissim et Zaninello, 2013) ou hongrois (Vincze *et al.*, 2013). Enfin, d'autres travaux se limitent à l'identification d'EP continues (Scholivet *et al.*, 2018).

Si la tâche d'identification utilise parfois des méthodes similaires à celles de la découverte d'EP, elle se heurte à des problèmes spécifiques : les mesures d'association telles que la PMI par exemple ne permettent pas de résoudre l'ambiguïté entre les variations correspondant à des EP ou à des non-EP (Nissim et Zaninello, 2013). De façon générale, l'identification d'EP repose sur un ensemble de méthodes, pouvant être combinées dans le cas d'approches hybrides (Constant *et al.*, 2017).

1. Par exemple : *NF-kappa B* = *NFkappa B* = *NFkB*.

6.1.2.1 Règles

On peut par exemple stipuler que toute séquence de lemmes ***poser un lapin*** dans un texte donné doit être étiquetée EP sous réserve que le substantif *lapin* soit au singulier, ou bien privilégier des co-occurrences survenant dans une fenêtre réduite de mots. Des transducteurs à états finis peuvent aussi être mis à profit pour définir des règles sous la forme de contraintes, comme l'exemple ci-dessous cité par Gross (1987), dans lequel les transducteurs couvrent les 12 variantes lexicales possibles d'une EP :

(6.4)

$$Bob \quad a \quad \begin{pmatrix} \textbf{\textit{nourri}} \\ \textbf{\textit{rechauffé}} \end{pmatrix} \quad \begin{pmatrix} \textbf{\textit{un}} & \textbf{\textit{serpent}} \\ \textbf{\textit{une}} & \textbf{\textit{vipère}} \end{pmatrix} \quad \begin{pmatrix} \textbf{\textit{sur}} \\ \textbf{\textit{dans}} \\ \textbf{\textit{en}} \end{pmatrix} \quad \textbf{\textit{son}} \quad \textbf{\textit{sein}}$$

L'approche choisie par Breidt *et al.* (1996) repose quant à elle sur l'utilisation d'expressions régulières pour la codification des contraintes. Pour l'EP ***perdre la tête*** par exemple, cela donne `perdre V: ADV* [:la :tête | :la :boule | :les :pédales]`, ce qui signifie que *perdre* doit être un verbe, pouvant être suivi d'un nombre non spécifié d'adverbes, et que cette EP tolère des variantes lexicales, mais soumises à des restrictions morphologiques matérialisées par les deux points (singulier ou pluriel selon la variante). Des restrictions syntaxiques (passivation interdite, etc.) sont également ajoutées si nécessaire. Des macros à pouvoir généralisateur regroupent des caractéristiques partagées par différents types d'EP. Des règles et macros ont également été exploitées par Li *et al.* (2003) pour l'identification de VPC en anglais à la fois continues et discontinues. Leur évaluation portant sur 3 VPC représentatives en corpus montre des F-mesures élevées, variant entre 96,6 et 97,5 selon la VPC considérée. L'analyse d'erreurs révèle toutefois que certaines des contraintes formulées se révèlent trop fortes (un seul adverbe optionnel au maximum) ou au contraire trop faibles (ambiguïté du verbe *to have* '*avoir*' devant être traité de façon différenciée des autres verbes).

Les travaux de Cordeiro *et al.* (2016b) mentionnent également l'identification de VPC en anglais d'après des règles. Ces règles sont formulées de façon à prendre en compte les EP vues dans un corpus d'entraînement ainsi que de nouvelles EP. Dans le premier cas, les auteurs s'appuient sur le niveau de systématicité d'annotation des séquences vues : dès lors que ce niveau excède 40% pour les séquences continues et 70% pour les discontinues, ces VPC seront systématiquement annotées dans le corpus de test. Dans le second cas, une phase de découverte de séquences VERB-ADP précède leur identification en exploitant des contraintes portant sur les POS et les lemmes des composants, sur d'éventuelles discontinuités et sur les éléments post-posés selon le modèle présenté dans (Ex. 6.5).

(6.5)

$$\text{VERB} \notin \{be, have\} \quad \begin{pmatrix} \emptyset \\ \text{NOUN} \\ \text{PROPN} \end{pmatrix} \quad \begin{pmatrix} \text{ADP} \in \{\textbf{\textit{about}} \\ \textbf{\textit{around}} \\ \textbf{\textit{away}} \\ \textbf{\textit{back}} \\ \textbf{\textit{down}} \\ \textbf{\textit{in}} \\ \textbf{\textit{into}} \\ \textbf{\textit{off}} \\ \textbf{\textit{on}} \\ \textbf{\textit{out}} \\ \textbf{\textit{over}} \\ \textbf{\textit{through}} \\ \textbf{\textit{up}}\} \end{pmatrix} \quad \begin{pmatrix} \text{ADV} \\ \text{ADP} \\ \text{PART} \\ \text{CONJ} \\ \text{PUNCT} \end{pmatrix}$$

Cette méthode présente une F-mesure nettement plus faible (65) que celle obtenue par Li *et al.* (2003) (plus de 31 points d'écart). Les restrictions imposées par les listes de particules ou les POS situées immédiatement après sont notamment pointées du doigt. La meilleure performance de Li *et al.* (2003) s'explique en effet par la finesse des règles définies : chaque particule était décrite par une macro spécifique et une règle prenait également en considération des formulations passives etc. On s'aperçoit donc que les règles peuvent s'avérer particulièrement performantes à condition de couvrir tous les cas particuliers, ce qui requiert par conséquent un corpus de développement suffisamment représentatif (Li *et al.*, 2003), ce dont il est toujours difficile de s'assurer.

L'inconvénient majeur de ces règles est en définitive de nécessiter l'intervention d'experts humains pour la codification ou l'adaptation de contraintes, ce qui risque de rendre plus difficile la généralisation à d'autres langues (ou variantes d'une même langue) surtout si l'on ne souhaite pas restreindre l'identification d'EP verbales à certaines catégories (donc pas uniquement aux VPC, d'ailleurs quasi-inexistantes en français) : il faudrait dans ce cas envisager des règles spécifiques selon les catégories.

6.1.2.2 Classification supervisée pour l'identification d'EP

Une alternative aux règles est d'utiliser des techniques de *machine learning*. La classification dite supervisée [2] s'appuie sur la disponibilité d'exemples dont les étiquettes sont connues (corpus `train`), l'algorithme cherchant alors, grâce à cette connaissance, à classer (c.-à-d. à attribuer une étiquette) automatiquement chaque nouvel exemple figurant dans un corpus dit de `test`. Un corpus additionnel annoté (dit corpus `dev`) est parfois utilisé pour évaluer les performances du modèle et le perfectionner, indépendemment de l'évaluation finale sur le `test`. Nous nous plaçons dans cette thèse dans le cadre de la classification supervisée en raison de la disponibilité de corpus manuellement annotés en EP. Ces corpus nous offrent en effet des conditions privilégiées pour une analyse et une exploitation de la variabilité observée en corpus.

2. En apprentissage non supervisé, on ne fournit au système ni données étiquetées ni les classes attendues. Ce mode d'apprentissage automatique s'appuie sur la capacité des modèles à découvrir des motifs sous-jacents pour procéder à la classification.

Modélisation des données Du point de vue de la modélisation informatique, les méthodes de classification supervisée tirent bénéfice, durant la phase d'entraînement, de n exemples $\{x^{(1)}, x^{(2)},... x^{(n)}\}$, représentables sous la forme d'un vecteur X. Chaque exemple $x^{(i)}$ est décrit par un ensemble de m informations (ou *traits*) $t_j^{(i)}$ et de leurs m valeurs $v_j^{(i)}$ correspondantes, représentable sous la forme d'un vecteur de traits-valeurs (T). Enfin, en raison du caractère supervisé, chaque exemple est associé à une étiquette correspondant à sa classe $\{y^{(1)}, y^{(2)},... y^{(n)}\}$, également représentable sous forme vectorielle (Y). Les méthodes de classification tirent alors parti des traits caractérisant chacun des exemples : les lemmes, les POS, les caractéristiques morphologiques ou syntaxiques, etc.

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ ... \\ x^{(n)} \end{bmatrix} \quad T = \begin{bmatrix} (t_1^{(1)} = v_1^{(1)}, t_2^{(1)} = v_2^{(1)}, ..., t_m^{(1)} = v_m^{(1)}) \\ (t_1^{(2)} = v_1^{(2)}, t_2^{(2)} = v_2^{(2)}, ..., t_m^{(2)} = v_m^{(2)}) \\ ... \\ (t_1^{(n)} = v_1^{(n)}, t_2^{(n)} = v_2^{(n)}, ..., t_m^{(n)} = v_m^{(n)}) \end{bmatrix} \quad Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ ... \\ y^{(n)} \end{bmatrix}$$

Une façon de modéliser les étiquettes $y^{(i)}$ dans le cadre de l'identification d'EP avec des classifieurs séquentiels est de s'appuyer sur l'encodage dit BIO [3] : le premier token de chaque EP sera étiqueté 'B' (pour *begin*) et les tokens suivants seront étiquetés 'I' (*inside*) tandis que les tokens n'appartenant à aucune EP seront étiquetés 'O' (*outside*). Considérons plusieurs exemples annotés de la sorte :

(6.6) ***Il***$_B$ ***faut***$_I$ ***prendre***$_B$ ***acte***$_I$ *du*$_O$ *verdict*$_O$

(6.7) *Il*$_O$ ***prendra***$_B$ ***acte***$_I$ *demain*$_O$ *de*$_O$ *cette*$_O$ *suggestion*$_O$

(6.8) *J'*$_O$ *ai*$_O$ *pris*$_O$ *des*$_O$ *actes*$_O$ *à*$_O$ *la*$_O$ *mairie*$_O$

(6.9) *Cet*$_O$ *acte*$_O$ *écologique*$_O$ *prend*$_O$ *de*$_O$ *l'*$_O$ *importance*$_O$

On peut modéliser les données de façon séquentielle en traitant chaque phrase du corpus d'entraînement comme un exemple différent. Les traits portent alors sur chaque mot $x_i^{(j)}$ de la phrase $x^{(j)}$. On obtient ainsi, pour l'ensemble des exemples X correspondant aux exemples 6.6-6.9, une matrice de traits T – portant par exemple sur la forme de surface et le lemme de chaque token – et un vecteur d'étiquettes Y. Dans le cas de l'étiquetage BIO, les valeurs de Y se limitent *de facto* aux trois valeurs B, I et O. Cette première méthode de modélisation des données est illustrée par les matrices $X^{①}$, $T^{①}$ et $Y^{①}$.

$$X^{①} = \begin{bmatrix} Il & faut & prendre & acte & du & verdict & & \\ Il & prendra & acte & demain & de & cette & suggestion & \\ J' & ai & pris & des & actes & a & la & mairie \\ Cet & acte & ecologique & prend & de & l' & importance & \end{bmatrix}$$

$$T^{①} = \begin{bmatrix} (\text{SURFACE}_0 = il, \text{LEMME}_0 = il) & (\text{SURFACE}_1 = faut, \text{LEMME}_1 = falloir) & ... \\ (\text{SURFACE}_0 = il, \text{LEMME}_0 = il) & (\text{SURFACE}_1 = prendra, \text{LEMME}_1 = prendre) & ... \\ (\text{SURFACE}_0 = j', \text{LEMME}_0 = je) & (\text{SURFACE}_1 = ai, \text{LEMME}_1 = avoir) & ... \\ (\text{SURFACE}_0 = cet, \text{LEMME}_0 = ce) & (\text{SURFACE}_1 = acte, \text{LEMME}_1 = acte) & ... \end{bmatrix}$$

3. D'autres formats existent (BILOU, IO) mais dépassent notre propos (do Amaral *et al.*, 2015).

$$Y^{①} = \begin{bmatrix} B & I & B & I & O & O & & \\ O & B & I & O & O & O & O & \\ O & O & O & O & O & O & O & O \\ O & O & O & O & O & O & O & \end{bmatrix}$$

Dans cette thèse, nous avons choisi une modélisation non séquentielle : les exemples fournis au classifieur représentent des EP attestées ou potentielles (que nous qualifions de *candidates*) extraites au préalable à partir de corpus annotés, et non des phrases. Autrement dit, si une phrase contient plusieurs (non-)EP (comme dans l'exemple 6.6), elles seront traitées comme des exemples différents. Au lieu d'étiquettes BIO, les classes que nous cherchons à attribuer sont $y \in$ {'EP','non-EP'} [4] selon qu'il s'agit d'une EP, ou d'une non-EP (lecture littérale ou co-occurrence fortuite), ce qui constitue une classification binaire. Cette modélisation adaptée aux exemples 6.6-6.9 donnerait :

$$X^{②} = \begin{bmatrix} (\text{SURFACENOM} = n/a, \text{LEMMENOM} = n/a) \\ (\text{SURFACENOM} = acte, \text{LEMMENOM} = acte) \\ (\text{SURFACENOM} = acte, \text{LEMMENOM} = acte) \\ (\text{SURFACENOM} = actes, \text{LEMMENOM} = acte) \\ (\text{SURFACENOM} = acte, \text{LEMMENOM} = acte) \end{bmatrix} \quad Y^{②} = \begin{bmatrix} 'EP' \\ 'EP' \\ 'non-EP' \\ 'EP' \\ 'EP' \end{bmatrix}$$

Dans cette configuration, comme les traits portent sur les EP, il faut préciser à quel composant correspondent les valeurs de chacun des traits, ce qui peut être effectué en créant des traits différents selon les POS (LEMMENOM, LEMMEVERBE, etc.). L'absence de nom dans l'EP ***il faut*** explique la valeur n/a (non applicable) attribuée.

Capacité de généralisation du modèle : validation croisée Faute de disposer de multiples corpus de `test`, la reproductibilité du modèle de classification peut être estimée par validation croisée. Il est en effet essentiel de savoir dans quelle mesure le modèle développé est influencé par les données d'entraînement fournies, autrement dit à quel point il est sensible au surapprentissage. Pour cela, la méthode utilisée dans cette thèse repose sur une partition à plusieurs reprises (10 dans notre cas [5]) du corpus `train` en deux sous-corpus, $\texttt{train}_{90}$ (90% du `train` initial) et $\texttt{train}_{10}$ (les 10% restants), le premier étant utilisé pour l'entraînement et le second pour l'évaluation. Nous pouvons dès lors calculer un score F moyen sur les 10 partitions et l'écart-type associé. Ces partitions peuvent être réalisées de façon à ce que la proportion des classes dans chaque partition soit identique à celle du `train` initial, ce qui est qualifié de *validation croisée stratifiée*. Nous utilisons dans cette thèse la validation croisée simple (section 13.2.4), le corpus `dev` étant réservé à l'optimisation des (hyper-)paramètres du modèle de classification (section 13.3.3.2).

4. Nous n'avons pas utilisé de classification non-binaire exploitant les catégories ($y \in$ {'non-EP', '$\text{IRV}_{1.1}$', '$\text{VID}_{1.1}$', '$\text{LVC}_{1.1}$', '$\text{MVC}_{1.1}$', etc.)}. Des classifications binaires telles que $y \in$ {'non-EP', '$\text{IRV}_{1.1}$'}, $y \in$ {'non-EP', '$\text{VID}_{1.1}$'}, etc. ont cependant été testées sans obtenir de résultats probants.

5. En anglais : *10-fold cross validation*.

Classification classique Nous exposons dans cette section quelques unes des méthodes de classification supervisée les plus répandues : CRF, Naïve Bayes, SVM et arbres de décision.

Les *CRF* (champs aléatoires conditionnels ou *Conditional Random Fields*) sont des modèles d'apprentissage supervisé (Lafferty *et al.*, 2001). Ils prennent en entrée un ensemble de phrases considérées comme des séquences linéaires de mots, chaque mot étant associé à une étiquette comme illustré par les matrices $X^{(1)}$ et $Y^{(1)}$. Les CRF cherchent à représenter la distribution de probabilité des étiquettes d'après les exemples X fournis. Dans le cas des exemples 6.6 à 6.9, la probabilité de l'étiquette I serait ainsi plus élevée lorsque le nom *acte* est au singulier et immédiatement précédé du verbe *prendre* quelque soit sa flexion. La sortie d'un CRF correspond à l'étiquetage BIO de nouvelles phrases. La nature séquentielle des CRF, autrement dit leur faculté à tirer parti de l'information contextuelle [6] locale (p. ex. le lemme et la POS précédant ou suivant la séquence) les rend aptes au traitement d'EP peu variables, notamment celles dont les composants sont contigus, ce qui explique leur emploi fréquent pour la reconnaissance d'entités nommées (Vincze *et al.*, 2011; Katiyar et Cardie, 2018). Scholivet *et al.* (2018) ont développé un système à base de CRF pour l'identification d'EP continues de différentes catégories en français (dont un ensemble limité d'EP verbales telles que ***avoir lieu***). Cette focalisation sur les EP continues leur permet de se contenter de corpus non parsés tout en ayant des performances d'identification d'EP de toutes catégories (nominales, verbales, etc.) proches – quoique légèrement en deçà – de celles obtenues par Le Roux *et al.* (2014) avec des corpus parsés (resp. F = 79,1 vs. 82,4). Quoique l'encodage BIO prédestine les CRF à l'identification d'EP continues, on peut également les employer pour l'identification d'EP discontinues en ajoutant des étiquettes dédiées au signalement des discontinuités (Schneider *et al.*, 2014; Zampieri *et al.*, 2018). Ces CRF peuvent également exploiter des données externes telles que des lexiques d'EP manuellement constitués ou bien automatiquement extraits à partir de corpus (Constant et Tellier, 2012; Schneider *et al.*, 2014; Riedl et Biemann, 2016; Scholivet *et al.*, 2018). Cette méthode permet de lever des ambiguïtés et se distingue de surcroît par sa rapidité d'exécution. L'identification d'EP discontinues grâce à l'association de CRF et des dépendances syntaxiques est possible : le système TRAVERSAL développé par Waszczuk (2018) s'appuie sur le principe des CRF tout en y apportant une différence notable puisqu'il tire parti d'arbres de dépendances et non de séquences. Pour Moreau *et al.* (2018), cette association entraîne un coût computationnel si élevé qu'ils sont contraints à réduire de moitié le corpus d'entraînement. Une telle réduction serait préjudiciable dans notre cas : d'une part un grand nombre d'exemples permet l'acquisition de profils de variabilité plus précis et d'autre part cela restreindrait le vivier d'EP dont on rechercherait ensuite des variantes.

Une autre méthode de classification supervisée, un *classifieur bayésien naïf* (désormais : *Naïve Bayes*), s'appuie sur le théorème de Bayes :

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}$$

6. Nous distinguons la notion de séquence de celle de contexte quoiqu'elles puissent être reliées : le caractère séquentiel provient du fait de prendre en compte des suites de mots, le caractère contextuel relève de la prise en compte, pour un mot donné, des mots qui l'entourent.

et sur une hypothèse forte d'indépendance mutuelle des traits vis-à-vis des classes y :

$$P(X|y) = \prod_{k=1}^{n} P(x^{(k)}|y) \text{ (Rish } et \text{ } al.\text{, 2001)}$$

La classe $y^{(i)} \in Y$ prédite est celle qui maximise la probabilité calculée d'après (i) les probabilités conditionnelles individuelles de chaque paire trait-valeur $P(t^{(k)} = v^{(k)}|y)$, (ii) la probabilité de la classe :

$$P(y^{(i)}) \prod_{k=1}^{n} P(x^{(k)}|y^{(i)}) > P(y^{(j)}) \prod_{k=1}^{n} P(x^{(k)}|y^{(j)}) \quad \forall j \neq i$$

L'hypothèse d'indépendance des traits sur laquelle se fonde Naïve Bayes ne peut, en réalité, être parfaitement respectée par nos données. De nombreuses informations sont effectivement redondantes, comme le genre et le nombre d'un adjectif qui dépendent de ceux du nom modifié. Toutefois, des travaux mettent en évidence que cela ne constitue pas un obstacle rédhibitoire pour l'utilisation de ce classifieur (Rish *et al.*, 2001). Il s'agit en outre d'un modèle de classification de compréhension aisée, qui se contente de peu de données d'entraînement et qui n'est pas sensible à la présence de traits superflus. C'est pourquoi nous l'utiliserons comme une technique de référence dans nos expérimentations préliminaires (chapitre 12) avant d'envisager l'emploi d'autres techniques de classification (chapitre 13).

Les *SVM* (*séparateurs à vaste marge*) sont une autre technique de classification supervisée (Vapnik, 2013). Un SVM recherche le séparateur (l'hyperplan de séparation) optimal – dans un espace dont les dimensions sont définies par les traits – permettant la distinction des exemples des deux classes avec la marge la plus importante (Fig. 6.1, en haut). Cela le rend particulièrement pertinent pour des classifications binaires comme celle opérée dans cette thèse.

Dans le cas où aucune droite ne parvient à séparer les classes de façon optimale (Fig. 6.1, en bas et Fig. 6.2 en haut), un SVM linéaire n'est pas pertinent et il est alors préférable d'utiliser un SVM non-linéaire permettant de rechercher un hyperplan grâce à l'introduction d'une fonction noyau (en anglais *kernel*) de nature polynomiale, sigmoïde, etc. (Fig. 6.2, en bas). L'un des intérêts des SVM est que l'hypothèse d'indépendance des traits y joue un rôle moins central que pour un classifieur Naïve Bayes. Ils sont également capables de généralisation, car moins enclins au surapprentissage. Dans cette thèse, seuls des SVM linéaires ont été utilisés : tout en étant adaptés à l'utilisation d'un grand nombre de traits, ils sont d'exécution plus rapide que des SVM non linéaires et ne nécessitent pas de multiples paramétrages.

Diab et Bhutada (2009) ont recouru à des étiqueteurs séquentiels élaborés à partir de SVM à kernel polynomial de degré 2 pour l'identification d'EP en anglais de patron VERBE-NOM. Des travaux font état de l'utilisation de SVM pour l'identification d'EP sans privilégier l'approche séquentielle, mais dans d'autres langues que le français (magahi pour Kumar *et al.* (2017), anglais pour Boukobza et Rappoport (2009)). Ces derniers se sont restreints à quelques patrons d'EP (VERBE-DET-NOM, VERBE-PREP-NOM et VERBE-NOM-PREP) et ont utilisé un SVM ayant un kernel PUK avec $\omega = 1$ et $\sigma = 1$. A l'instar de notre système VarIDE décrit dans la partie IV – qui gère cependant davantage de patrons –, ils procèdent par extraction de candidats et opèrent ensuite une classification binaire

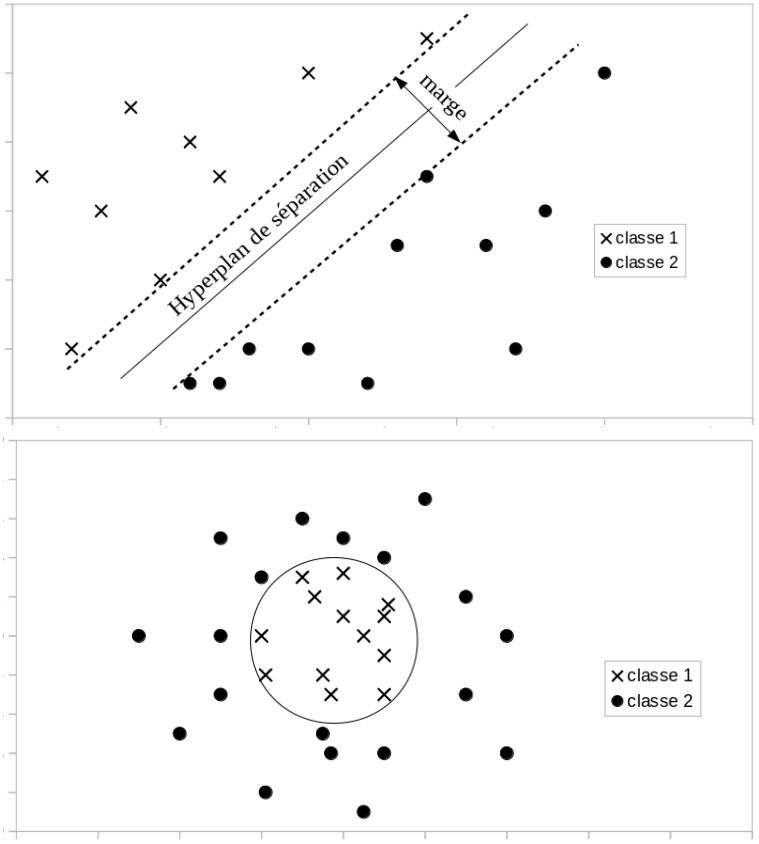


FIGURE 6.1 – Exemples de données linéairement séparables (en haut) et non-linéairement séparables (en bas).

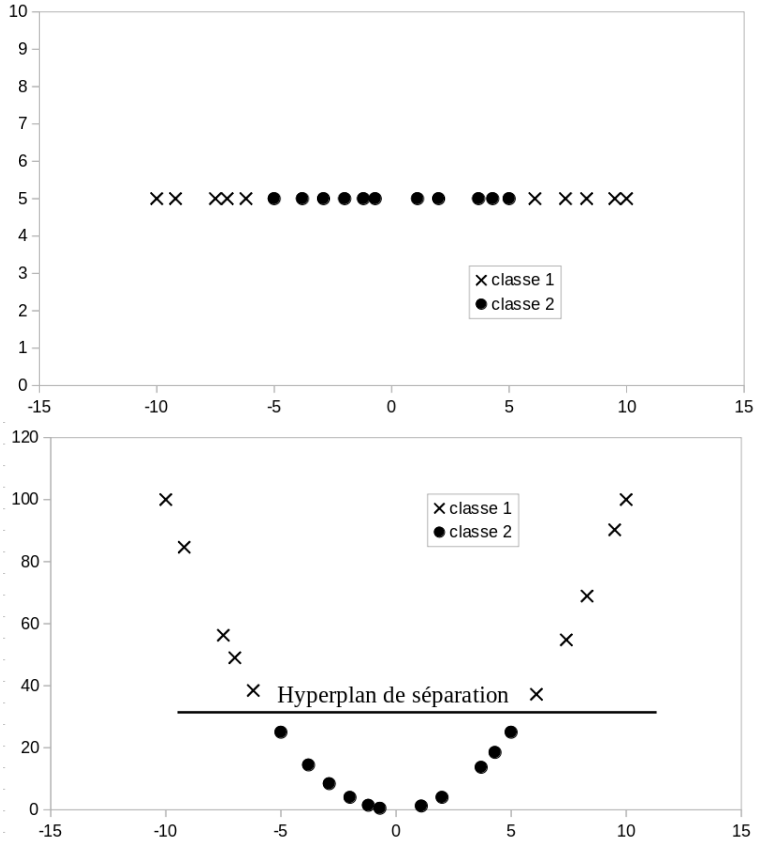


FIGURE 6.2 – Exemple de transformation de données non linéairement séparables (en haut) en données linéairement séparables (en bas) : la dimension de l'ordonnée ne permettant pas de différencier les deux classes est redéfinie en tant que fonction quadratique.

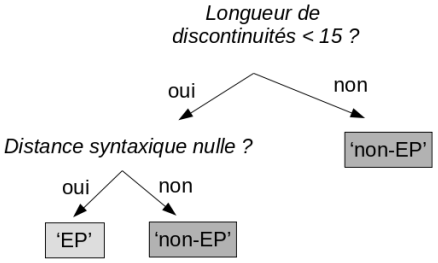


FIGURE 6.3 – Exemple d'arbre de décision : deux tests permettent l'attribution d'une étiquette de classe. Comme chaque arbre se subdivise en deux sous-arbres, il s'agit d'un arbre binaire.

('EP' vs. 'non-EP') sur la base de traits syntaxiques (distance syntaxique, existence de modifieurs, etc.) et surfaciques (longueurs des discontinuités, permutation de composants, pourcentage de composants ayant une flexion différente de celle de la forme de référence, etc.). L'une des différences notables de leur méthodologie vis-à-vis de la nôtre est d'utiliser un classifieur par EP. Ils cherchent en effet à modéliser le profil de chacune d'entre elles, ce qui est rendu possible par le fait d'avoir en moyenne 97 exemples ('EP' + 'non-EP') pour chacune. A titre de comparaison, dans le corpus d'entraînement exploité pour VarIDE, seules 11 EP sur 1339 (0,08%) bénéficient d'une telle couverture.

Citons pour terminer une quatrième technique : les *arbres de décision*. Un arbre de décision est une représentation arborescente d'une série de *tests* suivant des *attributs*, chaque test étant matérialisé sous la forme d'un *nœud de décision* dans l'arbre (Breiman *et al.*, 1984). Comme le montre la Figure 6.3, les *feuilles* de l'arbre fournissent les étiquettes de classe $y \in$ {'EP','non-EP'} d'après la réponse binaire ('oui','non') aux tests concernant les attributs 'Longueur de discontinuités' et 'distance syntaxique' (ces traits seront explicités dans la partie III).

Les arbres de décision sont construits en utilisant le critère de *gain d'information* afin d'identifier les attributs les plus pertinents pour chaque partition associée à la création de nœuds dans l'arbre. Ici, le fait que l'attribut portant sur la distance linéaire soit plus proche de la *racine* de l'arbre signifierait qu'il est plus discriminant pour le choix de la classe que celui portant sur la distance syntaxique. En pratique, la méthode de gain d'information (Kumbhar et Mali, 2016) sépare l'ensemble des observations X d'après les différentes valeurs d'un attribut donné t, puis calcule la somme pondérée des entropies de ces partitions $H(X|t)$, et la compare à l'entropie $H(X)$ de l'ensemble :

$Gain(H|t) = H(X) - H(X|t)$

où $H(X) = -\sum_{y \in Y} p(y) log_2 p(y)$ et $H(X|t) = \sum_{v \in V} p(t = v) H(X|t = v)$,

V représentant l'ensemble des valeurs possibles pour l'attribut t, par exemple $V =$ {'oui','non'} sur la Figure 6.3. Si une partition selon t engendre une baisse notable de la somme d'entropies, alors t a un fort pouvoir discriminant.

L'intérêt majeur des arbres de décision réside dans leur interprétabilité : la classification

s'opère en suivant les branches de la racine aux feuilles, d'après des règles telles que {SI (LongueurDiscontinuités $<$ 15 = oui) ET (DistanceSyntaxiqueNulle = oui) ALORS 'EP'}. Ces arbres permettent en outre la gestion de données catégorielles (chaînes de caractères) et numériques. Leur inconvénient majeur est le risque que des arbres complexes établis d'après le corpus d'entraînement manquent de pouvoir de généralisation vis-à-vis de nouvelles données. De plus, le principe de construction des arbres implique que leurs capacités de discrimination sont celles d'un découpage linéaire de l'espace, puisque chaque test sur un attribut crée un découpage sur une dimension. Notons qu'il est possible, à partir de plusieurs arbres de décision, de construire un classifieur nommé *forêt aléatoire* (Breiman, 2001) (section 13.2.3). L'une des implémentations d'arbre de décision nommée C4.5 a été exploitée par Metin (2018) pour sélectionner les traits statistiques et linguistiques les plus pertinents pour identifier des EP en turc.

Classification neuronale Les modèles de *perceptrons multi-couches* (en anglais *Multi-layer perceptrons*) sont un type de réseau neuronal artificiel, c'est-à-dire une architecture constituée d'un ensemble de neurones. A l'instar des neurones biologiques, les neurones artificiels peuvent être modélisés comme des cellules reliées les unes aux autres et dont le noyau active les signaux de sortie d'après les stimulations en entrée.

On suppose que l'obtention du signal y en sortie du neurone se fait au moyen d'une fonction mathématique dite *d'activation*, par exemple la fonction sigmoïde $f(x) = \frac{1}{1+e^{-\theta^T x}}$. Dans le cadre de l'apprentissage automatique, l'algorithme d'apprentissage ajuste au mieux le vecteur de pondération θ du réseau de neurones sur plusieurs couches, chacune constituée d'un ou de plusieurs neurones, de façon à minimiser l'erreur en sortie par comparaison avec le signal de sortie attendu. La rétro-propagation du gradient est l'algorithme le plus répandu pour minimiser cette erreur : cet algorithme remonte les différentes couches du réseau (c.à.d. des neurones de sortie vers ceux d'entrée), en adaptant les poids synaptiques de chaque couche.

Un autre type de réseau neuronal, les *réseaux de neurones récurrents* (ou RNN), autorisent, sous forme rétro-active, la réutilisation de prédictions précédentes en tant que nouvelles entrées. L'utilisation de réseaux neuronaux semble prometteuse pour l'identification d'EP : ils peuvent par exemple exploiter les plongements de mots pour apprendre à étiqueter un candidat comme étant une EP lorsque la compositionnalité de ses composants n'est pas satisfaite (Constant *et al.*, 2017).

Durant la compétition internationale PARSEME 1.0 portant sur l'identification d'EP verbales, les réseaux de neurones étaient utilisés par 1 seul des 7 systèmes présentés, tandis que pour l'édition suivante PARSEME 1.1 (désormais dénommée ST, de l'anglais *shared task*), 9 des 17 systèmes utilisaient des réseaux neuronaux (BiLSTM [7], CNN [8] ou RNN) (Savary *et al.*, 2017; Ramisch *et al.*, 2018), ce qui traduit un intérêt croissant pour cette technique. Cependant, les résultats obtenus lors de l'édition 1.1 pour le français par les systèmes à base de réseaux neuronaux sont bien inférieurs à ceux du meilleur système : entre 5 et 19 points de $F_{\text{seen-in-train}}$-mesure en moins pour les systèmes ayant participé en mode fermé (c.-à.-d. sans autres ressources que les corpus fournis), et de 5 à 9 points

7. De l'anglais *Bi-directional Long Short-Term Memory*.
8. *Convolutional Neural Network*, en français : réseau de neurones convolutif.

pour les autres. De plus, les ajustements effectués par ces réseaux de neurones sont plus difficiles à interpréter par l'utilisateur, d'où leur qualification fréquente de *boîte noire*. Pour toutes les raisons énoncées et le fait que certains modèles sont plus récents que certaines expériences dans cette thèse, nous n'utiliserons pas de tels classifieurs ici.

6.1.2.3 Classification par similarité

Nous avons développé une nouvelle méthode de classification, dénommée `SIM` qui repose sur l'hypothèse que chaque candidat suffisamment proche, dans l'espace des traits définis, d'une occurrence attestée de la même EP, est également une EP. Il s'agit d'une adaptation d'une méthode classique, *kNN* (k plus proches voisins, en anglais *k-Nearest neighbours*) au cas des EP, *via* la définition d'une mesure de similarité entre exemples. La technique de classification kNN s'appuie sur la classe la plus fréquemment observée pour les k plus proches voisins d'un exemple donné pour lui attribuer cette classe. En revanche, notre méthode `SIM` ne se restreint pas à un nombre spécifié k : elle compare chaque candidat avec chaque token de la même EP. De plus, l'attribution de l'étiquette 'EP'/'non-EP' dépend d'un seuil préalablement défini.

Cela requiert en premier lieu de définir cet ensemble de traits. Il faut en second lieu tenir compte du fait que certains traits peuvent être plus pertinents que d'autres (hypothèse H_5, développée dans la partie IV), d'où le recours à des méthodes de sélection de traits (chapitre 13). La mesure `SIM` s'appuie sur l'une des méthodes de classification supervisée car c'est le modèle de classification qui nous fournit les poids associés à chaque trait-valeur favorisant la prédiction de la classe 'EP', par exemple le trait portant sur des discontinuités ABS_DISCONTSEQ = ADV bénéficie d'un poids de 1,97 dans un modèle de classification (SVM) tandis qu'il n'est que de 0,02 pour ABS_DISCONTSEQ = PUNCT, ce qui signifie que l'insertion d'un adverbe entre les composants est davantage associée à l'étiquette 'EP' que celle d'un signe de ponctuation.

Pour un candidat d'EP donné, décrit par ses valeurs propres pour les traits sélectionnés, nous procédons à une mesure de similarité pondérée vis-à-vis de chaque token de la même EP de référence, grâce aux poids précédemment obtenus. Chaque candidat bénéficie donc d'un score de similarité, qui s'étend de 0 (aucune similarité) à 1 (similarité parfaite), attribué d'après la valeur maximale de similarité observée avec l'un des tokens attestés. Une fois chaque candidat doté de son score de similarité, les classes réelles déduites par comparaison avec l'annotation du corpus permettent de définir un seuil optimal au-delà duquel on attribuera l'étiquette 'EP'. Cette optimisation du seuil est effectuée sur le corpus de `dev`. La classification d'exemples dans le corpus de `test` soumis à l'évaluation s'appuie donc à la fois sur les pondérations des traits (d'après le `train`) et le seuil de similarité (d'après le `dev`). Contrairement aux méthodes standards de classification, on ne tient ici compte que des traits favorisant la classe 'EP', et non d'un compromis entre l'attribution de cette classe et de la classe inverse. Son intérêt est donc sa faculté d'interprétation.

6.1.2.4 Méthodes de REN

Les méthodes employées pour la REN rejoignent celles utilisées pour l'identification d'EP verbales, qu'il s'agisse des techniques employées ou de leur chronologie (Yadav et

Bethard, 2018) : elles reposaient initialement sur la formulation de règles (par exemple toute séquence suivant *Mr.* est un nom de personne), l'exploitation de particularités orthographiques (capitalisation, etc.) et le recours à des lexiques ou des ontologies, ces derniers offrant une précision satisfaisante mais un rappel entaché par les lacunes des ressources, comme dans le cas de médicaments nouveaux ou non autorisés (Segura Bedmar *et al.*, 2013).

Puis des techniques de machine learning ont été utilisées comme l'illustre le recours à ces techniques durant des compétitions d'identification multilingue d'EN en 2002 (espagnol et néerlandais) et 2003 (anglais et allemand) (Tjong Kim Sang, 2002; Sang et De Meulder, 2003) : SVM, CRF, arbres de décision notamment. Lors de l'édition de 2003, les réseaux de neurones font leur apparition pour cette tâche : le système de Hammerton (2003) est l'unique système exploitant cette technique parmi les 16 systèmes participants. Les architectures neuronales – essentiellement récurrentes – sont désormais de plus en plus employées. Elles tirent profit d'informations portant notamment sur les mots de l'EN (considérés en intégralité ou sur la base de caractères ou d'affixes), les parties de discours et peuvent aussi exploiter des word embeddings. Yadav et Bethard (2018) ont appliqué ces techniques neuronales aux corpus des éditions de 2002 et 2003, donc sur quatre langues (anglais, néerlandais, allemand, espagnol) : ils obtiennent un gain de F-mesure pour la REN par rapport à des systèmes non neuronaux allant jusqu'à 3 points pour l'espagnol (Yadav et Bethard, 2018), ce qui justifie l'intérêt qu'elles suscitent.

L'identification d'EN partage certaines similarités avec celle des EP qu'il s'agisse de sa finalité (l'annotation de corpus), des méthodes utilisées, voire des expressions concernées, nombre d'EN étant également des EP (donc des ENP). Cependant, la variabilité des EN est sous-tendue par des mécanismes spécifiques qui ne sont pas toujours transférables aux EP verbales (p. ex. l'emploi d'abréviations comme dans ***Valéry Giscard d'Estaing*** = ***VGE***) et elle outrepasse parfois le cadre de la polylexicalité, une variante d'ENP pouvant apparaître sous forme monolexicale (***Valéry Giscard d'Estaing*** = *Giscard*).

6.1.2.5 Bilan

On s'aperçoit de la diversité des méthodes permettant l'identification d'EP, d'autant plus qu'il existe des approches hybrides, combinant par exemple des méthodes statistiques et des mesures d'association (Schneider *et al.*, 2014; Scholivet *et al.*, 2018; Constant et Tellier, 2012). La tâche d'identification peut aussi être couplée avec celle de parsing, ce que nous détaillons dans la section 6.2.1.2. De façon générale, bien qu'une intervention humaine soit nécessaire pour établir des règles ou annoter manuellement de grandes quantités de données fournies en entrée de modèles de classification, le machine learning supervisé semble plus adapté à notre objectif d'extension multilingue que l'édiction de règles. Une connaissance des spécificités des EP dans chaque langue pourra ainsi être inférée par l'analyse d'un grand nombre d'exemples sans besoin de recourir à des experts de la langue en question, comme le fait qu'une séquence Verbe-∅-Nom ait de fortes probabilités (en français) d'être une EP (p. ex. ***faire appel, avoir lieu, prendre connaissance***, etc.). Notons toutefois que l'application stricte de cette règle conduirait à une piètre précision d'identification d'EP dans le corpus `FR-train1.1` ($P = 23$), à moins d'imposer une contrainte sur la fréquence des types d'EP considérés : l'exclusion des hapax permettrait par exemple

de doubler la précision ($P = 53$). De fait, l'absence de déterminant n'est qu'un critère parmi d'autres pour l'identification d'EP (p. ex. invariabilité en nombre). C'est pourquoi le fait de définir des ensembles de traits – dans le cadre de notre hypothèse H_2 de profil multidimensionnel des EP – octroie davantage de souplesse pour modéliser le caractère 'EP' vs. 'non-EP' que des règles, par définition rigides donc plus enclines à dégrader les performances lorsque l'on s'écarte du modèle.

Les techniques à base de réseaux de neurones semblent davantage tournées vers la recherche de performance que vers la compréhension des phénomènes en raison de l'effet *boîte noire*. Or, la variabilité des EP est un phénomène multidimensionnel et il nous semble intéressant d'acquérir en premier lieu cette connaissance sur leur fonctionnement et de chercher à la modéliser, quoique nous n'écartions pas l'idée de faire appel aux réseaux de neurones dans l'avenir, ne serait-ce que pour confronter les performances avec et sans réseaux de neurones. Dans cette thèse, nous avons privilégié des méthodes de classification classiques, davantage interprétables. Cette recherche d'interprétabilité est en effet cruciale à nos yeux pour mieux appréhender les spécificités des EP.

6.1.3 Évaluation

L'évaluation de la tâche d'identification repose sur la comparaison entre un même corpus annoté automatiquement et manuellement en EP. A l'instar de la découverte (section 5.2.4.2), la performance d'identification est alors évaluée grâce aux mesures de précision, rappel et F-mesure, mais la question de l'importance des tokens dans l'évaluation se pose également. Dans le cadre de la ST, la F-mesure peut ainsi être évaluée sur l'intégralité des composants (*F-per-EP*) ou en considérant les composants de façon séparée. Dans ce cas, si l'EP à identifier est ***jeter l'éponge*** et que seuls le verbe et le nom ont été automatiquement annotés, mais pas le déterminant, la F-mesure *F-per-token* permet de calculer un score au pro-rata des composants correctement annotés. Cependant, aucune des mesures n'est complètement satisfaisante : l'oubli d'un déterminant est autant pénalisé par la mesure *per-token* que l'absence du verbe *jeter*, alors qu'elle nous semble moins préjudiciable. A l'inverse, la mesure *per-EP* ne valorise pas l'identification des composants principaux (ici le verbe et le nom).

Cette mesure expose par ailleurs à l'écueil souligné par Savary *et al.* (2017) : un token correctement identifié peut cependant être erronément assigné à d'autres EP. Ainsi, dans l'exemple 6.10, tous les tokens de ***il est question*** (signalés en gras selon notre convention pour les composants) sont bien identifiés, mais attribués par le système à deux autres types d'EP comme le montrent les indices numériques (***il en est*** et ***poser question***).

(6.10) ***Faut-***$_1$ ***il***$_1$ *redouter les éoliennes géantes qu'****il***$_2$ ***est***$_2$ ***question***$_3$ *de poser*$_3$ *en*$_2$ *mer ?*

La Table 6.1 illustre trois annotations possibles de la phrase 6.10 : le premier système aurait des valeurs *R-per-EP* et *P-per-EP* maximales tandis qu'elles vaudraient 0 pour le deuxième système. Le troisième système, dont l'annotation est illustrée par l'exemple 6.10 aurait un rappel de 50 et une précision de 33. Or, le deuxième système a partiellement identifié chacune des EP (3 composants alignés), d'où l'intérêt de la mesure *per-token* qui lui attribue $R_{\text{per-token}} = \frac{3}{5} * 100 = 60$ et $P_{\text{per-token}} = \frac{3}{5} * 100 = 60$. Le troisième système, n'ayant

au maximum que quatre composants alignés avec la référence, obtient alors : $R_{\text{per-token}} = \frac{4}{5} * 100 = 80$ et $P_{\text{per-token}} = \frac{4}{7} * 100 = 57$.

Position du token	Token	Annotation de référence	Annotation du système 1	Annotation du système 2	Annotation du système 3
1	***faut***	a	1	1	1
2	***il***	a	1	2	1
8	***il***	b	2	1	2
9	***est***	b	2	2	2
10	***question***	b	2	2	3
12	*poser*	_	_	_	3
13	*en*	_	_	_	2

TABLE 6.1 – Comparaison de l'identification automatique d'EP dans la phrase 6.10 par trois systèmes par rapport à une annotation de référence. Les indices numériques et alphabétiques des colonnes d'annotation signalent les composants d'une même EP. L'annotation du troisième système est illustrée par l'exemple 6.10.

Chaque type de mesure (P, R ou F) est également utilisé pour évaluer la performance des systèmes d'identification dans différentes configurations :

- selon que les EP du `test` sont continues ou non,
- selon que les EP du `test` sont multi-tokens ou non,
- selon que les EP du `test` ont été vues dans le corpus d'entraînement (*seen-in-train*) ou non (*unseen-in-train*),
- selon que les EP *seen-in-train* du `test` sont en apparence identiques (*identical-to-train*) ou non (*variant-in-train*) au corpus d'entraînement.
- selon que l'on considère chaque langue individuellement ou bien l'ensemble des langues proposées (*macro-average*). A la différence d'une micro-moyenne, cette macro-moyenne garantit l'indépendance du calcul de performance vis-à-vis de la quantité de données de `test` disponibles selon les langues.

A l'instar de l'évaluation de la découverte, un autre mode d'évaluation, dit *évaluation extrinsèque*, repose sur l'impact d'une identification d'EP préalable vis-à-vis d'une tâche ultérieure, par exemple la traduction automatique ou le parsing. La comparaison des performances de ces tâches ultérieures avec et sans identification préalable d'EP permet alors de quantifier cet impact.

6.2 Lien entre l'identification et des tâches annexes

La traduction et le parsing ne sont pas des tâches portant spécifiquement sur les EP, mais elles mettent en lumière l'impact négatif lié à certaines particularités de EP (section 6.2.1.1), comme la question de l'ambiguïté (lecture idiomatique vs. littérale ou co-occurrence fortuite), prégnante en matière de parsing (section 6.2.1) ou de traduction automatique (section 6.2.2). Une identification des EP permettrait donc d'améliorer les performances de ces tâches. Notons cependant que, si l'ambiguïté entre lecture idiomatique et littérale pose généralement problème pour la traduction automatique, le parseur ne rencontrerait aucune difficulté pour analyser l'énoncé *il* ***jette l'éponge*** *et divorce aus-*

sitôt car sa structure syntaxique est identique à celle de n'importe quelle structure libre (p. ex. *il jette l'éponge et essuie aussitôt*). Dans ce cas, la nécessité d'identifier l'EP dépend des motivations finales de l'utilisateur : s'il est ensuite question de traduction automatique, une identification des EP avant de procéder à la traduction peut permettre de lever certaines ambiguïtés.

On peut dès lors s'interroger sur la façon dont la tâche d'identification d'EP peut s'articuler avec celles de parsing ou de traduction : c'est ce que Constant *et al.* (2017) qualifient d'*orchestration*, autrement dit le fait de choisir si l'identification doit précéder, être concomitante ou succéder aux tâches applicatives. De ce fait, l'identification d'EP peut être vue soit comme un moyen d'optimiser des tâches ultérieures (ce qui en permet une évaluation indirecte), soit comme une finalité.

6.2.1 Parsing

6.2.1.1 Motivation

Le parsing syntaxique permet de mettre en évidence les fonctions syntaxiques (p. ex. sujet, objet) et les relations de dépendance syntaxique qu'entretiennent les éléments d'une phrase (p. ex. adjectif relié au nom qu'il modifie). Considérons par exemple la phrase *La femme jeta l'éponge verte*, la phase de prétraitement inclut sa tokénization (Ex. 6.11), sa lemmatisation (Ex. 6.12) puis son étiquetage en parties de discours (Ex. 6.13). Le parsing en dépendances génère ensuite une représentation arborée (Fig. 6.4) qui permet de comprendre que le substantif *femme* tient la fonction de sujet nominal (*nominal subject* ou NSUBJ), que l'objet est le nom *éponge* par ailleurs modifié par l'adjectif *verte* (AMOD signifiant *adjectival modifier*).

(6.11) La – femme – jeta – l' – éponge – verte – .

(6.12) Le – femme – jeter – le – éponge – vert – .

(6.13) DET – NOUN – VERB – DET – NOUN – ADJ – PUNCT

Le parsing est important car il permet un accès à la structure de la phrase, laquelle peut influer sur son sens. Reconnaître le sujet et l'objet d'une phrase permet de distinguer *le garçon mord le chien* de *le chien mord le garçon*. De plus, en ce qui concerne plus spécifiquement les EP, reconnaître l'adjectif *verte* comme modifieur du nom *éponge* fait basculer la lecture de l'expression *jeter l'éponge* vers son acceptation littérale.

Le parsing peut également être réalisé par un parseur en constituants, à partir de grammaires formelles établies manuellement ou extraites de corpus (CFG, TAG, HPSG, CCG, etc.). Parmentier et Waszczuk (2019) dressent un panorama de certaines de ces méthodes en relation avec la représentation et le parsing des EP. L'emploi de grammaires LTAG pour la modélisation de l'ambiguïté entre lectures idiomatiques et littérales est par ailleurs mentionné par Lichte et Kallmeyer (2016). L'analyse syntaxique superficielle (en anglais *shallow parsing* ou *chunking*) identifie les constituants de la phrase comme les groupes nominaux (en anglais *nominal phrase* ou NP) ou verbaux (en anglais *verbal phrase* ou VP) mais sans préciser leur fonction (p. ex. sujet) comme illustré en (Ex. 6.14), par opposition au *deep parsing* de la Figure 6.4. Nous précisons à ce propos que cette thèse s'appuie exclusivement sur le parsing en dépendances.

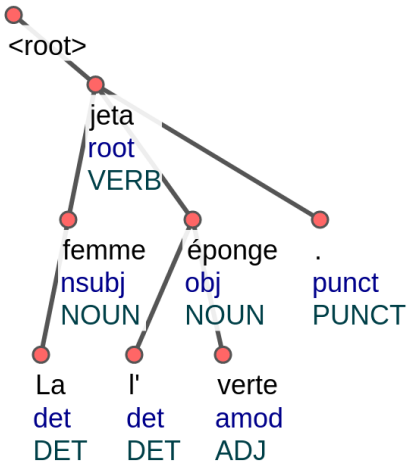


FIGURE 6.4 – Exemple de phrase parsée via UDPipe (Straka et Straková, 2017).

(6.14) [La femme]$_{NP}$ [jeta]$_{VP}$ [l'éponge verte]$_{NP}$

Des EP ayant une idiosyncrasie syntaxique particulièrement marquée (p. ex. ***by and large***) risquent de faire échouer le parsing, à moins d'avoir intégré au modèle l'existence d'une telle idiosyncrasie (Constant *et al.*, 2017). La non-compositionnalité syntaxique de certaines EP (p. ex. le substantif ***laissez-passer*** formé à partir de deux verbes) peut aussi engendrer des erreurs de parsing. Citons également l'exemple ***Il pleut des cordes*** *de Paris à Marseille* ⇒ 'il pleut très fort' (Fig. 6.5, à gauche) dans lequel *Paris* est erronément rattaché par un analyseur syntaxique automatique en dépendances, en l'occurrence UDPipe, au nom *cordes* – comme s'il s'agissait de *cordes de nylon* par exemple – alors qu'il aurait dû être rattaché au verbe comme dans ***Il pleut*** *de Paris à Marseille* (Fig. 6.5, à droite). Il suffirait ici de savoir que l'EP interdit la modification du nom pour en obtenir une représentation correcte. Ce défaut de détection d'EP est problématique car Baldwin *et al.* (2004) a observé qu'il engendrait 8% d'erreurs de parsing.

6.2.1.2 Orchestration

Parsing → identification En procédant dans l'ordre Parsing → identification (c.-à-d. parsing préalable à la tâche d'identification), des discontinuités apparentes peuvent être neutralisées car des composants très éloignés sur le plan linéaire de la phrase peuvent entretenir une relation syntaxique directe. C'est le cas pour le verbe *prendre* et le nom *décision* dans l'exemple 6.15 (illustré par l'arbre Fig. 6.6), malgré une discontinuité de longueur égale à 15 éléments, ponctuation incluse. Notons que ces discontinuités peuvent être multiples : elles sont au nombre de deux dans (Ex. 6.16).

Certaines expressions peuvent également se chevaucher comme dans (Ex. 6.17) où une locution adverbiale s'insère entre ***prendre*** et ***décision***. D'autres partagent un même composant : dans (Ex. 6.18), le verbe *prendre* intervient dans les deux EP ***prendre bain*** et ***prendre douche***.

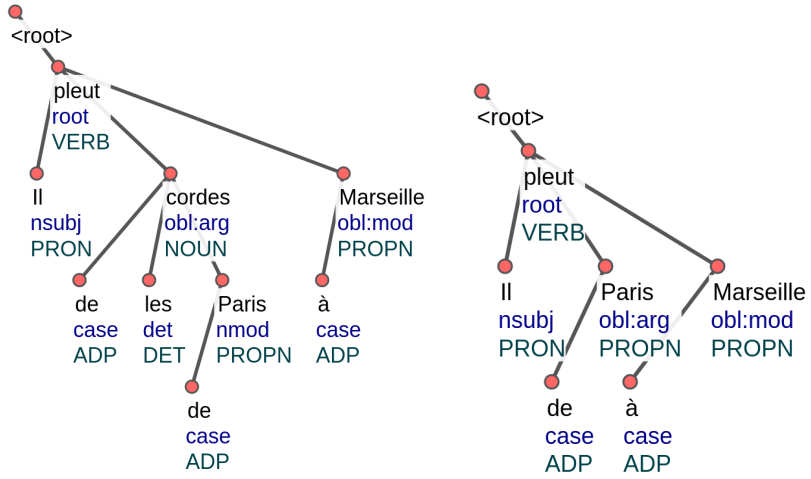


FIGURE 6.5 – Représentation sous la forme d'arbre syntaxique de la phrase ***Il pleut (des cordes)*** *de Paris à Marseille*, obtenue à partir de UDPipe (Straka et Straková, 2017)

(6.15) *Il* ***prend*** *également, et cette fois-ci seul et sans vote de l'assemblée, des* ***décisions*** *réglementaires.*

(6.16) *J'aimerais* ***en avoir*** *vraiment* ***le cœur*** *bien* ***net***. ⇒ 'J'aimerais savoir à quoi m'en tenir'

(6.17) *Il a* $\textbf{\textit{pris}}_1$ $\textbf{\textit{de}}_2$ $\textbf{\textit{toute}}_2$ $\textbf{\textit{évidence}}_2$ *la meilleure* $\textbf{\textit{décision}}_1$

(6.18) *Il a* $\textbf{\textit{pris}}_{1\text{-}2}$ *une* $\textbf{\textit{douche}}_1$ *puis un* $\textbf{\textit{bain}}_2$

Malgré certains cas d'idiosyncrasie syntaxique (p. ex. ***by and large***), bon nombre d'EP ont un fonctionnement syntaxique standard (p. ex. ***prendre*** *une* ***décision***/ *un fruit*), ce qui permet au parseur de les gérer de façon appropriée.

Cette orchestration a été choisie dans les travaux de Fazly *et al.* (2009) ainsi que par certains systèmes ayant participé à la ST (Moreau *et al.*, 2018; Pasquer *et al.*, 2018a). Disposer de corpus parsés permet de cibler la recherche de composants qui co-occurrent dans une même phrase tout en étant reliés syntaxiquement. Dans le cas de l'EP ***jouer rôle***, on s'attend par exemple à l'existence d'une connexion syntaxique entre le verbe *jouer* et le nom *rôle*. Cette connexion, matérialisée par une flèche est bien présente dans l'exemple de la Figure 6.7-a – qui est une EP – mais pas dans la figure 6.7-b – qui n'en est pas une. L'existence de connexions révélées par le parsing a été employée dans les travaux de Savary et Cordeiro (2017); Pasquer *et al.* (2018c).

Identification → parsing Ne pas reconnaître des mots complexes engendre des erreurs de parsing d'autant plus préjudiciables que ces expressions sont très fréquentes. Or il existe des ressources d'EP, certes incomplètes, mais qui recensent des EP figées telles que les locutions adverbiales (p. ex. ***en fait***), conjonctives (***au cas où***) ou prépositionnelles (***au détriment de***) (Ramisch, 2017). Fort de cette connaissance, le parseur peut appliquer un traitement spécifique, nommé *retokénization*, qui consiste à remplacer toutes les séquences du lexique d'EP relevées en corpus par un nouveau token (p. ex. ***en fait*** → ***en_fait***). Dans le cas d'EP tolérant une flexion morphologique, cette variabilité est neutralisée si le

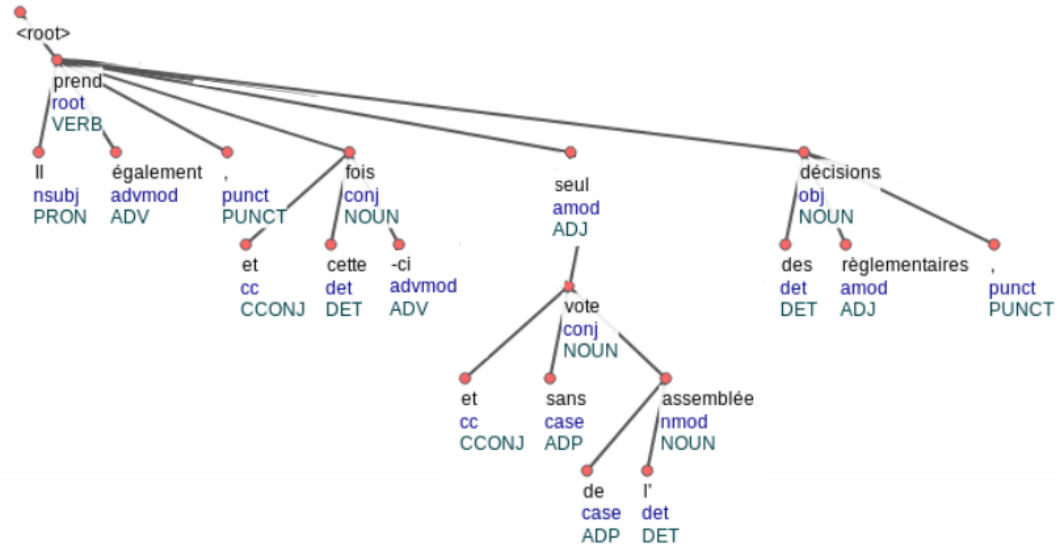
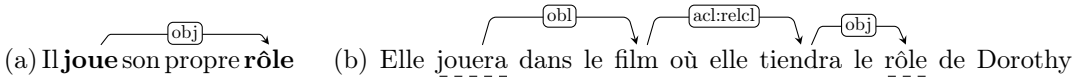


FIGURE 6.6 – Représentation sous la forme d'arbre syntaxique de dépendances de la phrase *Il* ***prend*** *également, et cette fois-ci seul et sans vote de l'assemblée, des* ***décisions*** *réglementaires.* Représentation obtenue à partir de UDPipe (Straka et Straková, 2017)



(a) Il **joue** son propre **rôle** (b) Elle jouera dans le film où elle tiendra le rôle de Dorothy

FIGURE 6.7 – Connexion syntaxique entre le verbe *jouer* et le nom *rôle* dans (a) mais pas dans (b)

texte et l'entrée du lexique sont lemmatisés (***grand(s)-mère(s)*** → ***grand-mère***).

La retokénization n'est réellement envisageable que pour les EP continues, sinon il faut modifier l'ordre des composants : *Il viendra* ***à*** *très* ***court terme*** → *Il viendra* ***à_court_terme*** *très* (Constant *et al.*, 2017). A condition de disposer de ressources d'EP, le processus de retokénization facilite la tâche de parsing et s'accompagne d'un gain de performances pour des EP fortement idiosyncratiques (p. ex. sémantiquement non-compositionnelles) : comparé à l'absence de retokénization, ce gain est de l'ordre de 20% pour des EP en anglais de patron ADJ-NOM contiguës (p. ex. `(EN)` ***dead end*** '*morte extrémité*'⇒'impasse') identifiées avant un shallow parsing (Korkontzelos et Manandhar, 2010). En revanche, la retokénization d'EP de catégorie CVS en turc s'est révélée préjudiciable pour le parsing ultérieur (Eryiğit *et al.*, 2011) (cité par Constant *et al.* (2017)). Cela est dû à la faible fréquence alors associée à chaque EP retokénizée. Le verbe `(TR)` *etmek* ('*faire*') fait ainsi partie de 88% des tokens d'EP du corpus turc, mais la retokénization a pour effet d'occulter cette prépondérance.

Il est également possible de conserver plusieurs alternatives de parsing afin de gérer les cas d'ambiguïté ***en fait*** → *en fait,* ***en_fait*** (Constant *et al.*, 2017). Le parsing final conserve la meilleure option en écartant les propositions agrammaticales : dans la phrase *Des concerts, il en fait un par an*, seule la première option (non-idiomatique) est valide car

la présence du pronom *il* est incompatible avec une phrase averbale.

Parsing ↔ identification Si, lors du parcours d'un texte, une séquence de mots donnée s'avère présente dans un lexique d'EP, procéder à une retokénization systématique est particulièrement tentant. Prenons l'exemple de la locution conjonctive ***bien que***, elle serait retokénizée sous la forme ***bien_ que***. Pourtant, toute séquence *bien que* n'est pas une EP : il peut s'agir de l'adverbe *bien* suivi par le relatif *que* (Ex. 6.20) et non d'une EP (Ex. 6.19) (Nasr *et al.*, 2015).

(6.19) *Je mange* ***bien que*** *je n'aie pas faim* (exemple cité par Nasr *et al.* (2015))

(6.20) *Je pense bien que je n'ai pas faim* (exemple cité par Nasr *et al.* (2015))

A la lecture de ces exemples, la différence de modes (subjonctif/indicatif) permet de discriminer respectivement une lecture idiomatique / co-occurrence fortuite. Toutefois, cet indice n'est pas exploitable durant la phase de tokénization si bien que Nasr *et al.* (2015) ne procèdent pas à la retokénization systématique sous la forme ***bien_ que***. Pour lever l'ambiguïté de cette séquence, au lieu d'attendre que le parsing ait eu lieu pour écarter les constructions agrammaticales comme dans l'orchestration Identification → parsing, ils s'appuient sur un lexique spécifiant la sous-catégorisation de chaque verbe. Concrètement, ce lexique indique si un verbe donné accepte un objet introduit par la conjonction de subordination *que*, ce qui est bien le cas pour *penser* (*je pense que j'ai faim*) mais pas pour *manger* (**je mange que j'ai faim*). Le parsing concomitant à l'identification permet donc de résoudre l'ambiguïté.

Nivre et Nilsson (2004) ont ainsi mis en évidence une réduction de 5% d'erreurs de parsing bénéficiant non seulement aux EP elles-mêmes mais aussi aux structures syntaxiques environnantes. De bons résultats avec cette orchestration ont également été relevés chez Constant et Nivre (2016); Green *et al.* (2013) : ils ont démontré que le parsing était particulièrement efficace pour l'identification d'EP continues en français avec 50% de gain par rapport à l'exploitation de n-grammes.

Face aux différentes orchestrations répertoriées, celle privilégiée pour cette thèse est : parsing → identification, avec une restriction opérée sur la tâche d'identification puisqu'elle se limitera à l'identification de variantes d'EP connues.

6.2.2 Traduction

6.2.2.1 Motivation

La traduction représente un défi pour le TAL notamment en raison de la présence d'EP. En effet, leur non-compositionnalité interdit généralement la traduction mot-à-mot, comme en témoignent les traductions automatiques erronées des EP `(EN)` ***kick the bucket*** ⇒ 'mourir' et ***casser sa pipe*** ⇒ 'mourir' par le service de traduction en ligne de *Google*[9] (Ex. 6.21-6.22). De plus, pour qu'une traduction soit valide dans la langue cible, il ne suffit pas qu'elle soit correcte, elle doit également être ressentie comme naturelle par les locuteurs

9. `https://translate.google.fr`, consulté le 10 avril 2019.

natifs [10]. Étant donné la proportion habituelle d'EP dans une langue, il est possible qu'une traduction comprenant une trop faible ou trop forte proportion d'EP sonne faux aux oreilles de ces locuteurs natifs.

(6.21) (EN) *The old guy* ***kicked the bucket*** *at the age of 102.* → (FR) # *Le vieil homme a frappé le seau à l'âge de 102 ans.*

(6.22) (FR) *Le vieil homme a* ***cassé sa pipe*** *à l'âge de 102 ans.* → (EN) # *The old man broke his pipe at the age of 102.*

De même, la désambigüisation entre lecture idiomatique ou littérale est parfois nécessaire afin de garantir la validité de la traduction :

(6.23) *Il* ***met*** *ses priorités* ***sur la table*** → (EN) *He raises his priorities.*

(6.24) *Il met ses livres sur la table* → (EN) *He puts his books on the table.*

Gérer l'ambiguïté de certains énoncés ou cas de défigement s'avère parfois nécessaire pour fournir une traduction fidèle aux choix stylistiques de l'auteur. Il est d'ailleurs une ambiguïté qui ne doit pas être résolue : c'est l'ambiguïté stylistique puisqu'elle est souhaitée par le locuteur (double sens). C'est le cas dans le titre d'article de presse (Ex. 6.25) dans lequel ***retourner au charbon*** ⇒ 'reprendre une activité désagréable' fait référence à une visite aux salariés d'un groupe sidérurgique dans un contexte social tendu.

(6.25) *La candidate Ségolène Royal* ***retourne au charbon*** *en Lorraine* (*Libération*, 21/05/2008, exemple cité par Privat (2008))

Les applications de TAL prenant en considération le sens d'un texte (p. ex. traduction, résumé automatique) doivent appliquer un traitement particulier pour les EP. La traduction automatique d'une EP se heurte à trois problèmes majeurs : (i) la nécessité de disposer d'une traduction déjà existante pour cette EP, (ii) l'impossibilité de traduction mot à mot, car toutes les EP ne sont pas compositionnelles sémantiquement, (iii) l'ambiguïté. Ne pas reconnaître une EP conduit à plusieurs écueils : la non-acceptabilité (Ex. 6.26), l'agrammaticalité lorsque des composants discontinus ne sont pas reconnus (Ex. 6.27) ainsi que la génération de traductions peu naturelles comme dans l'exemple 6.28, pour lequel *these* ***ready-to-explore*** *itineraries* serait jugé plus naturel que la traduction automatique du service de traduction de *Google* (Hilma, 2011).

(6.26) (EN) ***It's raining cats and dogs*** → (FR) *? ? Il pleut des chats et des chiens*

(6.27) (EN) *John* ***picked*** *the book* ***up*** → (FR) **John prit le livre jusqu'à* (Constant *et al.*, 2017)

(6.28) (FR) *Ces itinéraires* ***prêts-à-parcourir*** → (EN) *These routes ready to go* (traduction du service *Google Translate*, citée par Hilma (2011)).

La volonté de jouer avec les mots conduit également au détournement d'EP. Or, le transfert parfait de défigement d'une langue vers une autre en conservant la nature du défigement (phonétique dans l'exemple ***seringue sur le gâteau*** précédemment cité) n'est pas toujours possible (Mejri, 2009) :

(6.29) (EN) *Don't give pain* ***a leg to stand on*** '*ne laissez pas à la douleur une jambe sur laquelle se tenir*'⇒'n'aidez pas la douleur'/ (FR) *Faites* ***fuir*** *la douleur...* ***à toutes jambes***. (Publicité Dr Scholl's citée par Quillard (2001))

10. Le service de traduction de *Google* établit ce distinguo lorsqu'il propose de signaler les traductions erronées comme étant "incorrectes" ou bien "choquantes".

La traduction automatique de textes d'une langue source vers une langue cible peut s'inspirer de la traduction humaine : le texte source est d'abord *analysé* pour fournir une représentation intermédiaire et abstraite du sens, puis un nouveau texte est *généré* dans la langue cible tout en préservant cette représentation initiale (Ashraf et Ahmad, 2015; Ramisch, 2017).

Mais cette prise en compte d'un sens abstrait n'est pas systématique comme en attestent les trois courants majoritaires en traduction automatique (approche statistique, fondée sur des règles ou des exemples) mentionnés par Constant *et al.* (2017). L'approche statistique repose sur des techniques de machine learning pour construire des modèles de prédiction à partir de corpus bilingues. La traduction automatique neuronale, également associée à cette approche statistique (Constant *et al.*, 2017), sera décrite dans la section 6.2.2.2. L'approche fondée sur des règles, dite *rule-based machine translation*, s'appuie quant à elle sur des lexiques et met à profit les informations linguistiques (contraintes syntaxiques et sémantiques) propres aux langues sources et cibles. Enfin, la méthode nommée *example-based machine translation* exploite des exemples de traductions existantes pour extraire des correspondances de traduction et des connaissances linguistiques par analogie entre ces exemples (Hutchins, 2005).

Il existe des corpus parallèles donnant accès à des équivalences de traduction comme le corpus *EuroParl* qui contient les transcriptions multilingues des débats ayant eu lieu au sein du Parlement européen entre les années 2007 et 2011 [11]. Toutefois, peu de corpus parallèles disposent d'annotations en EP (Ramisch, 2017) : il en existe en anglais-hongrois et anglais-italien (Vincze, 2012; Monti *et al.*, 2015). Par ailleurs, certains corpus se focalisent parfois sur des catégories spécifiques à l'une des deux langues, par exemple les VPC anglaises telles que ***give up*** ⇒ 'abandonner' dans un corpus anglais-français pour Ramisch *et al.* (2013).

De façon générale, les corpus bilingues servent de support pour la recherche d'alignements (c.-à-d. de correspondance entre un élément ou une séquence d'éléments) entre deux textes de langue différente. La Figure 6.8 illustre l'alignement entre la version française d'un extrait du *Gentleman-cambrioleur* et sa traduction anglaise [12]. Après alignement des deux versions, on remarque que, d'une part, la version anglaise traduite est plus concise que la version française. D'autre part, si certaines séquences s'alignent sans difficulté (`(FR)` *avant huit jours* → `(EN)` *within eight days*), l'ordre des mots est parfois inversé (`(FR)` *les expédier* → `(EN)` *ship them*) et les EP ***faute de, procéder à déménagement*** sont remplacées par des mots isolés (*otherwise*, *remove*).

On remarque également que la segmentation en phrases peut différer entre deux textes, comme ici le point de la version française qui est remplacé par un point-virgule dans la version anglaise, ce qui peut compliquer la procédure d'alignement lorsque des éléments sont répartis sur plusieurs phrases dans l'un des deux textes.

Tant le texte source que le texte cible sont susceptibles de nécessiter une prise en charge des EP, qu'une EP soit traduite par une autre EP (***port payé*** → `(EN)` ***charges prepaid*** et non *#charges paid*), qu'une EP se traduise par un mot isolé (***procéder à déménagement*** → `(EN)` *remove*) ou l'inverse. Toutefois, c'est sur l'identification d'EP

11. `http://www.statmt.org/europarl`
12. Leblanc, Maurice. *The Extraordinary Adventures of Arsene Lupin, Gentleman-Burglar* : Project Gutenberg. Consulté le 17 avril 2019 sur `http://www.gutenberg.org/ebooks/6133`.

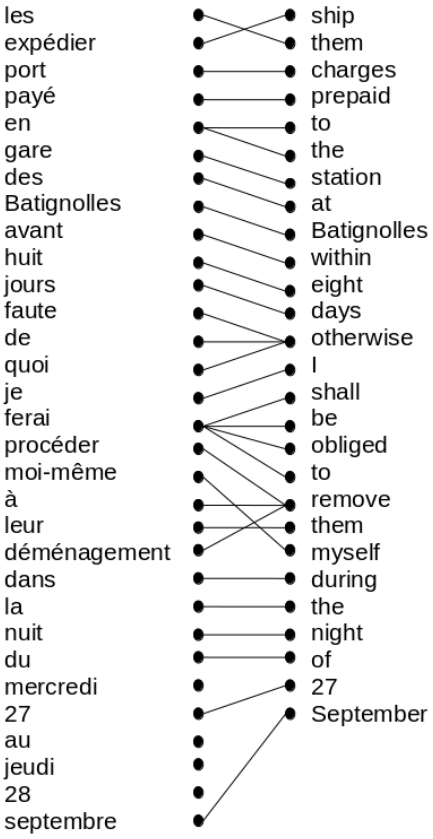


FIGURE 6.8 – Alignement entre la version originale française d'un extrait de roman (à gauche) et sa traduction anglaise (à droite).

dans le texte source que nous choisissons de nous focaliser, autrement dit sur la première étape de traduction automatique : celle qui porte sur l'analyse du texte source. D'ailleurs, peu de travaux portent sur prise en compte des EP lors de la phase de génération de textes (Ramisch, 2017). A l'instar du parsing, l'orchestration de l'identification avec la traduction peut s'effectuer de différentes façons.

6.2.2.2 Orchestration

Identification → traduction Si l'identification précède la traduction, on peut alors concaténer les composants des EP identifiées et en chercher des équivalents (Carpuat et Diab, 2010). Ramisch (2017) dresse un bilan des techniques d'identification antérieures à la traduction : soit par des règles, soit par étiquetage séquentiel, cette dernière méthode étant davantage adaptée pour les EP continues. Or les EP verbales sont fréquemment discontinues (*il a **pris** hier cette très importante **décision***). L'identification par règles repose sur un lexique pré-existant d'EP qui sont recherchées à l'identique en corpus (c.-à-d. sans flexion ni modification de l'ordre des composants), ce qui expose à des erreurs liées à l'ambiguïté des EP (*je pense bien qu'il risque de neiger* vs. *je viens **bien qu'**il risque de neiger*). Les lectures littérales doivent également être désambiguïsées. Dans ce but, l'utilisation d'un modèle ou système d'identification d'EP (p. ex. un CRF) peut permettre, en s'appuyant sur des données contextuelles, de signaler les EP avant de procéder à leur traduction (Ramisch, 2017).

Traduction ↔ identification La traduction automatique statistique (*statistical machine translation* ou SMT) nécessite des corpus parallèles bilingues volumineux. Pour chaque séquence de n mots dans la langue source, la SMT propose la séquence traduite la plus probable grâce à la connaissance extraite du corpus bilingue. On peut distinguer deux manières de gérer les EP dans le cadre de la SMT. La première consiste à ne pas traiter spécifiquement les EP : elles sont capturées de façon indirecte comme n'importe quelle séquence de n-grammes, à la manière d'une identification passive car non explicitement mise en œuvre. Cette approche convient essentiellement aux EP continues. La seconde méthode inclut une modélisation des EP de façon explicite au sein même du modèle de SMT (Carpuat et Diab, 2010). Par ailleurs, certains prétraitements et post-traitements sont parfois appliqués pour la gestion des EP. Ainsi, pour accroître le vocabulaire parallèle disponible, Stymne *et al.* (2013); Cap *et al.* (2014) scindent les mots composés allemands en mots simples, une stratégie qui leur permet ensuite, par combinaison des traductions des unités simples, de générer de nouveaux mots composés allemands.

La traduction automatique neuronale (en anglais *Neural Machine Translation* ou NMT) ne s'appuie pas sur des séquences de mots contrairement à la SMT. Elle procède en effet par traduction de l'intégralité des phrases en tirant profit de l'architecture récursive pour affiner le modèle de traduction. Le système GNMT (*Google* NMT) développé par Wu *et al.* (2016) atteint une qualité de traduction équivalente à celle de traducteurs humains pour différentes paires de langues (`EN` ↔ `FR`, `EN` ↔ `ES` et `EN` ↔ `CN`). De plus, selon les langues, le nombre d'erreurs de traduction diminue de 60% à 87% par rapport à une traduction de type SMT (Durrani *et al.*, 2014), ce qui témoigne de l'intérêt de la NMT. Ces méthodes de traduction automatique étant relativement récentes, des travaux portant sur leur prise

en compte des EP n'ont pas – à notre connaissance – encore vu le jour.

Cette thèse ne porte pas sur la traduction, donc il n'est pas pertinent de se positionner sur le choix d'une orchestration particulière.

Chapitre 7

Identification de variantes d'EP

Ce chapitre définit le cadre général de la tâche d'identification de variantes, qu'il s'agisse des défis associés (section 7.1) ou du traitement des variantes dans la littérature (EN, termes ou EP verbales) dans la section 7.2, ce qui donne lieu à la proposition d'une définition de la notion de variante (section 7.3). Une esquisse de la façon dont la variabilité peut aider à l'identification d'EP est présentée en section 7.4. La partie III, qui porte exclusivement sur la variabilité, fournira davantage de détails sur cette méthode.

7.1 Motivation

L'identification de variantes d'EP se présente comme une sous-tâche de l'identification d'EP car son objectif est d'identifier uniquement des EP connues, c'est-à-dire répertoriées dans un lexique, préalablement vues dans un autre corpus (nommé *corpus d'entraînement*) ou faisant partie d'une liste résultant d'une étape préliminaire de découverte [1]. Concrètement, il s'agit de repérer les composants d'EP connues quelque soit la forme sous laquelle elles apparaissent, autrement dit la focalisation porte ici sur la prise en compte de la variabilité des EP.

Le caractère zipfien de la distribution des EP en corpus (cf. Section 2.2) implique qu'un nombre restreint d'EP est très fréquent tandis qu'un grand nombre d'entre elles est rarement utilisé. Ces EP très fréquentes sont donc susceptibles d'apparaître dans n'importe quel corpus, et par conséquent de faire partie de lexiques ou de corpus annotés. Ces EP réservent néanmoins des difficultés liées à leur variabilité puisqu'une même forme canonique peut être associée à un nombre de réalisations plus ou moins important :

(7.1) `(LA)` ***alter ego*** → ***alter ego***.

(7.2) ***grand-mère*** → ***grand-mères***, ***grands-mères***.

(7.3) ***prendre décision*** → ***prendre*** *une* ***décision***, ***prendrons****-nous cette* ***décision****?, les* ***décisions prises*** *hier, cette* ***décision*** *importante mais difficile que tu as* ***prise****, etc.*

En effet, d'après Jacquemin (2001) qui s'est penché sur les EP nominales en français et en anglais, 28% des occurrences en corpus sont des variantes de formes canoniques

1. La rareté de ressources de bonne qualité est préjudiciable pour l'identification d'EP qui repose sur de telles ressources : il suffit qu'une non-EP y figure à tort pour générer des annotations erronées.

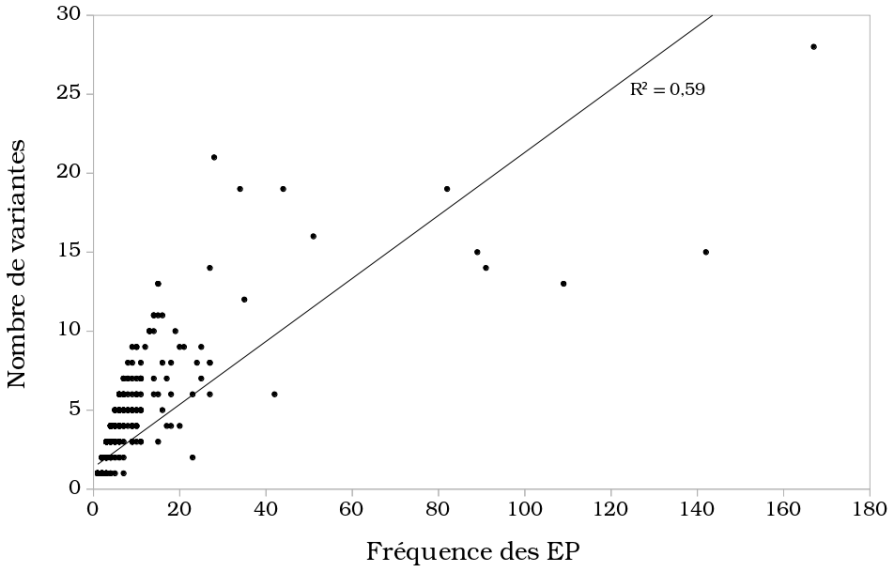


FIGURE 7.1 – Nombre de variantes observées dans le corpus `FR-train1.1` en fonction de la fréquence des types d'EP.

répertoriées dans des lexiques. Comme le montre l'exemple 7.3 par rapport aux exemples 7.1-7.2, les EP verbales sont encore plus susceptibles d'apparaître sous forme de variantes puisqu'elles ne se limitent pas à la flexion en genre ou en nombre. C'est d'ailleurs ce que nous observons dans le corpus français `FR-train1.1` où les 30 EP verbales les plus fréquentes (de 17 à 167 occurrences en corpus) se présentent sous 2 à 28 formes différentes si nous nous intéressons à la séquence de tokens allant du premier au dernier composant de chaque EP (en incluant les discontinuités), avec une moyenne de 11 formes par EP. A titre d'exemple, les 11 occurrences de l'EP ***mettre fin*** apparaissent ainsi sous 6 formes différentes : ***met fin***, ***mettait fin***, ***mettant fin***, ***mettre fin***, ***mit*** *donc* ***fin*** et ***mit fin***. Par ailleurs, comme le montre la Figure 7.1, la corrélation entre le nombre d'occurrences d'un type d'EP et le nombre de formes différentes observées est modérée (coefficient de détermination de Pearson $R^2 = 0{,}589$). C'est cette propension des EP verbales à apparaître sous une grande diversité de formes qui justifie notre focalisation sur leur variabilité.

Supposons que le corpus d'entraînement contienne l'EP `jeter l'éponge` sous la forme ***jette l'éponge*** (Ex. 7.4), la tâche d'identification de variantes aura pour objectif d'identifier dans un autre corpus les différentes réalisations possible de cette EP, et ce quelque soit leur forme de surface comme illustré dans (Ex. 7.5-7.6). Quant à l'EP ***s'en aller*** dans (Ex. 7.5), elle ne devra pas être annotée puisqu'elle ne figure pas parmi les EP vues. Les lectures littérales (Ex. 7.7) et les co-occurrences fortuites devront également être écartées (Ex. 7.8).

(7.4) *Il aurait fallu que je* ***jette l'éponge*** *avant d'être ruiné.*

(7.5) *Après une dispute avec son chef, il* ***jette l'éponge*** *et s'en va.*

(7.6) *Après 1000 euros de réparations, tu* ***jetteras l'éponge****.*

(7.7) *Une fois chez lui, il jette l'éponge dans l'évier.*

(7.8) *Ensuite, il jette la serpillière à côté de l'éponge.*

Plusieurs travaux portent sur la gestion de la variation d'EP, qu'il s'agisse d'entités nommées, de termes ou d'EP verbales.

7.2 EP de référence vs. variantes : *l'œuf ou la poule ?*

Lorsque l'on se penche sur la question des variantes d'EP vis-à-vis d'une expression de référence, il est parfois difficile de savoir ce qui doit être qualifié de référence ou de variantes en découlant. Cette problématique se rencontre pour des EP de différents types : entités nommées, termes et EP verbales. Cette section aborde la définition de variantes et de leur référence, souvent qualifiée de *forme canonique*. Cette section illustre à quel point cette notion est difficile à circonscrire, certaines définitions se distinguant de celle mentionnée en section 3.1, d'où notre proposition d'en établir une approximation afin d'automatiser les analyses et les expériences.

7.2.1 Variantes d'entités nommées (EN)

Il existe une tâche nommée *named entity normalization* (NEN) dont l'objectif est d'associer les variantes d'EN présentes dans un texte à leur forme canonique, ce qui permet ensuite de relier ces entités (tâche dite d'*entity linking*) à des entrées dans des bases de données, aux pages *Wikipedia* les concernant par exemple. Liu *et al.* (2012) ont appliqué cette approche aux tweets car leur caractère informel génère de nombreuses variantes d'EN (en moyenne 3,3 variantes par EN), par exemple le nom de la chanteuse ***Anneke Gronloh*** apparaît sous les formes ***Mw.,Gronloh, Anneke Kronloh, Mevrouw G***. Les méthodes d'obtention d'une forme canonique ne sont généralement pas pertinentes pour les variantes d'EP verbales : Liu *et al.* (2012) choisissent la mention la plus longue comme forme canonique, or la séquence ***prendre décision*** – qui est la moins marquée – est plus courte que ***prendraient décisions***. On note cependant que la normalisation numérique d'EN (et de termes biomédicaux) pour limiter la dispersion de données par Tsai *et al.* (2006) [2] pourrait s'appliquer à certaines EP verbales, par exemple ***ne pas casser trois (quatre) pattes à un canard*** ⇒ 'n'avoir rien d'extraordinaire'.

Pour favoriser la normalisation d'EN, Khalid *et al.* (2008) soulignent les nombreux atouts de *Wikipedia* :

— il s'agit d'une ressource à la fois de grande taille et largement constituée d'EN,
— la (quasi-)synonymie peut être gérée grâce aux redirections, comme dans le cas de la section vide ***King of pop*** qui pointe vers l'entrée ***Michael Jackson***,
— les cas d'homonymie sont répertoriés (par exemple 5 individus nommés ***George Bush***) et l'URL spécifique à chaque entrée leur octroie en outre une distinction aisée.

2. Les auteurs mentionnent le cas des gènes nommés IL2, IL3, IL4 et IL5. Si seuls IL2, IL3 et IL4 sont dans le corpus d'entraînement, mais pas IL5, ce dernier ne pourra pas être identifié. La normalisation numérique consiste alors à replacer tous les chiffres par la valeur 1, les différents gènes étant alors normalisés sous la seule forme IL1.

Cucerzan (2007) considère qu'une forme de surface est normalisée si elle apparaît (avec ou sans redirection) dans un lien *Wikipedia*, ce que l'on observe pour les synonymes *U.S.*, *U.S.A.* et *America*, tous redirigés vers la page *United_States*, qui constitue donc leur forme normalisée. Dans le cas contraire, c'est la forme la plus fréquemment utilisée comme lien hypertexte qui est considérée comme normalisée. D'autres tiennent également compte du nombre de liens entrants pour chaque homonyme afin d'en identifier le plus populaire, et c'est ce critère de popularité qui permet dès lors de déclarer quelle est la forme normalisée (Khalid *et al.*, 2008). Or, la popularité ne constitue probablement pas un indice fiable en raison de son caractère évolutif, ce qui constitue un biais important (Hachey *et al.*, 2013). Par ailleurs, la normalisation doit parfois n'être effectuée que sur une partie de l'EN, comme dans le cas où deux formes génitives sont enchâssées mais que seule la seconde doit rester à cette forme (Piskorski *et al.*, 2017). La question de la normalisation se révèle donc une tâche non triviale.

7.2.2 Variantes de termes

Définition et enjeux Jacquemin (2001) rappelle qu'une prise en compte de la variabilité des termes spécialisés est essentielle pour des applications telles que l'extraction d'informations. Il faut en effet s'abstraire de formes de surface similaires mais non pertinentes (co-occurrences fortuites) tout en conservant celles qui sont différentes (en apparence) de la forme référencée en lexique et cependant pertinentes. Pour qu'une séquence soit considérée comme la variante d'un terme, elle doit réunir les trois critères suivants (Daille, 2017) :

1. Une variante est une forme attestée en corpus,
2. Une variante se définit par rapport à une forme attestée d'un terme dans une ressource lexicale,
3. Une variante peut être (quasi-)synonyme du terme (*variante dénonimative* p. ex. *produit de la forêt, produit d'origine forestière*) ou instaurer une distance sémantique avec le terme (*variante conceptuelle*) comme dans *produit alimentaire forestier*.

Si la différence entre le terme de référence et ses variantes n'est pas toujours évidente, la fréquence d'apparition du premier est généralement supérieure en corpus (Daille, 2017). Quoique chaque variante considérée individuellement soit peu fréquente, l'ensemble des variantes constitue le tiers d'un corpus médical anglais (Jacquemin, 2001). Plus précisément, 9% sont des variantes syntaxiques avec insertion ou permutation de composants (p. ex. ***protéine (d'origine) végétale***), 6,5% des variantes morphosyntaxiques correspondant à une modification de la structure syntaxique et de la forme morphologique (p. ex. ***acidité du sang / sanguine***) et 22% des variantes sémantiques (p. ex. ***maladie génétique / héréditaire***). Ces modes de variation ne sont pas exclusifs : `(EN)` ***disease** is **familial*** est une variante syntactico-sémantique de `(EN)` ***genetic disease*** (Savary et Jacquemin, 2003).

A partir d'un terme de référence qui serait *produit forestier*, une autre classification des variantes s'établit de façon hiérarchique : les *variantes conceptuelles* acceptent des *variantes dénominatives*, qui à leur tour acceptent des *variantes linguistiques* (Daille, 2017) :

— variantes conceptuelles : préfixation (*sous-produit forestier*), dérivation (*production forestière*), ajout d'un adjectif (*produit alimentaire forestier*). Un accès vers le sémantisme du terme est parfois possible d'après des indices de surface tels qu'une suf-

fixation en *-raie* indiquant un lieu (*fraiseraie*, *oliveraie*, etc.). Les variantes conceptuelles incluent également des antonymes (*produit non forestier*),
- variantes dénominatives (synonymie) : *produit de la forêt, produit d'origine forestière*,
- variantes linguistiques (même concept) : variantes graphiques (*pro(-)duit forestier*), flexion (*produit(s) forestier(s)*), coordination (*produit agricole ou forestier*) ou énumération (*produit agricoles, hialeutiques et forestiers*).

Les variantes conceptuelles, dénominatives et linguistiques sont transposables aux EP verbales, comme l'illustrent quelques exemples :
- Variantes conceptuelles : ***(re)faire appel***, ***plaider (non) coupable***, ***lancer/recevoir*** *un* ***appel***,
- Variantes dénominatives : ***mener/conduire*** *une* ***étude***,
- Variantes linguistiques : ***faire parti(e)*** (*sic*), et discontinuités par coordination (***jouira*** *de tous les* ***droits*** *et* ***obligations***).

Le choix du type de variantes prises en compte dépend essentiellement des applications visées (Daille, 2017) : la présence de coordination ne nécessite ainsi pas d'identification préalable si l'on souhaite procéder à une traduction automatique. Elle est en revanche préférable pour l'indexation contrôlée.

Découverte et identification de variantes de termes nominaux Comme le souligne Daille (2017), la difficulté est double : les variantes sont polymorphes mais chaque forme n'apparaît que rarement en corpus. Daille (2017) présente plusieurs méthodes pour la découverte de termes nominaux. On peut ainsi s'appuyer (i) sur la recherche de n-grammes, voire de skip-grammes [3] afin d'autoriser une seule insertion (pour limiter le bruit), (ii) sur des méthodes de TAL telles que la distance d'édition [4] pour détecter par exemple des variantes graphiques ou la dérivation de variantes conceptuelles, (iii) sur des règles définissant des opérations élémentaires au niveau morphologique et syntaxique. La règle la plus productive pour le français est la conversion NOUN_1 ADJ_1 → NOUN PREP NOUN_1 ADJ_1 comme dans *parc éolien* → *installation de parcs éoliens*. Des variantes dénominatives peuvent également être générées par substitution d'un composant d'un terme complexe par un synonyme.

En matière d'identification de termes nominaux, l'outil FASTER [5] Jacquemin (2001) s'appuie sur une liste pré-établie de termes en détectant les variantes de ces termes en corpus. Les variations sont prises en compte grâce à un ensemble de règles morpho-syntaxiques et syntaxico-sémantiques. Il existe par exemple une règle qui stipule que l'adjectif `(EN)` *genetic* dispose de plusieurs synonymes (`(EN)` *familial*, *hereditary*, *inherited*, etc.).

Daille (2017) souligne l'intérêt d'étudier le phénomène de variabilité des termes dans un cadre multilingue et de bénéficier ainsi de traits linguistiquement motivés pour l'emploi de techniques de machine learning. Nous rejoignons cet intérêt pour la compréhension et la

3. A la différence des n-grammes qui capturent exclusivement des séquences continues, les skip-grammes – sans lien avec l'architecture du même nom de word2vec – tolèrent une discontinuité de k mots. Si $k = 0$, cela revient donc à un n-gramme.

4. Cette distance correspond au nombre d'opérations à effectuer au niveau des caractères (substitution, suppression ou insertion) pour passer d'une séquence à une autre.

5. `https://perso.limsi.fr/jacquemi/FASTR`

description de la variabilité. Toutes deux sous-tendent d'ailleurs notre tâche d'identification de variantes d'EP verbales.

7.3 Variantes d'EP verbales : notre définition

Quoique cette thèse porte exclusivement sur les variantes d'EP verbales, elle rejoint les travaux sur la variation d'entités nommées ou de termes sur une question fondamentale : que doit-on considérer (ou pas) comme une variante de la forme de référence ? Il faut pour cela non seulement circonscrire les contours de ce que l'on nomme *variante* mais également déterminer ce que l'on considère comme étant cette EP de référence. Cette recherche d'une normalisation des variantes rappelle la tâche de NEN, et la définition de variantes de termes peut également s'appliquer aux variantes d'EP verbales, qu'il s'agisse du point N° 1 (forme attestée en corpus, p. ex. *je* ***jetterai l' éponge***), N° 2 (définition par rapport à une forme attestée dans une ressource lexicale, p. ex. ***jetterai l'éponge*** ← ***jeter éponge***), N° 3 (quasi-synonymie : ***(re)jeter l'éponge***, ***assurer*** *la* ***(co-)présidence***, etc.).

Variante C'est dans le cadre de l'édition 1.1 de la compétition PARSEME sur l'identification d'EP verbales en corpus (Ramisch *et al.*, 2018) que nous avons proposé notre système d'identification de variantes VarIDE (Pasquer *et al.*, 2018a). Les variantes y étaient en effet bien représentées : la moitié des EP de `FR-test1.1` figurait, mais pas obligatoirement sous la même forme, dans le corpus `FR-train1.1`.

Notons à ce propos que, lors de l'édition 1.1, seules les réalisations d'EP dont les séquences de tokens entre le premier et le dernier élément lexicalisé différaient de celles du corpus d'entraînement (p. ex. ***jette l'éponge*** vs. ***jetteras l'éponge*** ou ***jette*** *encore* ***l'éponge*** vs. ***jette*** *très rarement* ***l'éponge***) étaient qualifiées de *variantes* (*variant-of-train*), et les autres de *identical-to-train*.

Or, une même forme surfacique peut occulter des différences : dans (Ex. 7.4), le verbe *jette* est au présent du subjonctif et à la 1ère personne du singulier alors que dans (Ex. 7.5), il est au présent de l'indicatif et à la 3ème personne du singulier. De telles différences peuvent être pertinentes, car un énoncé tel que *je/il vide son sac* ne pourra bénéficier d'une lecture idiomatique que si le sujet du verbe est à la même personne que le possessif.


C'est pourquoi nous choisissons de qualifier de *variante* toute réalisation attestée en corpus disposant des mêmes composants qu'une EP connue – servant de référence – sans considérer la similarité surfacique de leurs composants ou d'éventuelles discontinuités, ce qui correspond aux EP dites *seen-in-train* durant la ST.

A titre d'exemple, supposons que l'on ait vu l'EP ***jettes l'éponge*** dans le corpus d'entraînement, et ***jettes l'éponge***, ***jettes l'éponge***, ***jettera l'éponge*** dans le corpus de test, cela représente pour nous trois variantes tandis que seule la dernière serait considérée comme variante dans la terminologie PARSEME. La Table 7.1 revient sur les différences entre notre approche et celles de Jacquemin (2001) et Daille (2017).

Forme lemmatisée normalisée ≈ forme canonique La tâche d'identification de variantes repose sur les variantes d'EP connues. Si cette connaissance résulte d'EP annotées

Impact de la variation	Type de variation	Exemples	(a)	(b)	(c)	(d)	(e)
Composants	Ajout/suppression	***séparer le (bon) grain de l'ivraie***	✓	✓			
	Permutation	***prendre*** *une* ***décision*** ***décision prise***	✓	✓	✓	✓	✓
	Préfixation en *r-/re-/ré-* ou *de-/dé-*	***(re)prendre*** *une* ***décision***	✓	✓		✓	
	Synonymie	***mener/conduire*** *une* ***étude***	✓	✓			
	Antonymie	***lancer/recevoir*** *un* ***appel***		✓			
	Dérivation	***prendre*** *une* ***décision*** *prise de décision*	✓	✓			
	Graphie	***faire parti(e)*** (*sic*)	✓	✓			
	Flexion (différence visible)	***prendre*** *une* ***décision*** *il* ***prend*** *des* ***décisions***	✓	✓	✓	✓	✓
	Flexion (différence non visible)	*Il faut qu'il* ***mène***$_{\text{Présent.Subj.3.sing}}$ *une* ***étude*** *Je* ***mène***$_{\text{Présent.Ind.1.sing}}$ *une* ***étude***	✓	✓		✓	✓
Insertions	Ajout/suppression	***prendre*** *(d'ici la fin du mois) une* ***décision***	✓	✓	✓	✓	✓
	Adjectif	***prendre*** *une (importante)* ***décision***	✓	✓	✓	✓	✓
	Coordination	***prendre*** *un arrêté ou une* ***décision***	✓	✓	✓	✓	✓
	Énumération	***prendre*** *une ordonnance, un décret,* *un arrêté, un règlement, une* ***décision***	✓	✓	✓	✓	✓
Dépendances	Adjectif	***prendre*** *une* ***décision*** *(importante)*	✓	✓	✓	✓	✓

TABLE 7.1 – Différences entre les conceptions d'une *variante* dans l'état de l'art telle que définie par (a) (Jacquemin, 2001), (b) (Daille, 2017), (c) pour les *variant-of train* durant la compétition PARSEME 2018 et celles que nous avons utilisées lors de l'identification de variantes d'EP verbales de patron VERB-(DET)-NOUN dans (d) (Pasquer *et al.*, 2018c) et (e) pour tout patron par notre système VarIDE (Pasquer *et al.*, 2018a).

en corpus, le pré-requis est d'extraire de ce corpus toutes les réalisations observées. Il faut alors se distancier de cette observation concrète pour mettre à jour l'EP correspondante (p. ex. les 11 exemples de la page 49 relèvent de l'EP ***prendre décision***), de façon à réunir l'ensemble des réalisations propres à une EP sous une même entrée. Dans l'extrait de roman présenté p. 45, on aurait par exemple une entrée pour ***il y a*** avec cette seule réalisation, une autre pour ***me contenterai*** avec deux réalisations etc. Si cet extrait contenait également les tokens ***y-a-t-il, il y avait, contentez-vous***, elles ne devraient pas donner lieu à de nouvelles entrées car ***y-a-t-il*** et ***il y avait*** sont des variantes – avec des composants dans des ordres d'apparition parfois différents – du même type d'EP que ***il y a***, de même que ***contentez-vous*** vis-à-vis de ***me contenterai***. Pour regrouper toutes les réalisations sous une même entrée, à la façon d'une forme canonique, un choix naturel est la forme la moins marquée, donc avec une préférence pour des verbes à une forme finie, la voie active, l'absence d'extraction, de relatives, etc. (Savary *et al.*, 2018).

Certaines approches définissent comme forme canonique les emplois majoritaires, quoique cela ne soit pas complètement satisfaisant (Riehemann, 2001) :

> The definition of 'canonical form' is a bit tricky. It is essentially the form in which the idiom occurs most frequently in the data, modulo inflection of the head (if there is one). For example the canonical form of the idiom **spill the beans** is **spill/spills/spilled/spilling the beans**, while **spilling** many **beans**, **spilled the** royal **beans**, and **the beans** were **spilled** are non-canonical occurrences. However, defining 'canonical form' this way would be circular, as this form would by definition be the most frequent one. [6]

Dans l'optique de fusionner les variantes d'une même EP, cette notion de 'forme canonique' au sens de 'forme la plus fréquente' ne semble pas appropriée, car certains types d'EP n'apparaissent qu'une seule fois en corpus et il est alors impossible de savoir si ces hapax sont, ou non, une manifestation de LA forme canonique. Le corpus `FR-train1.1` ne contient par exemple qu'une seule occurrence du type illustré dans l'exemple 7.9, or ni la forme plurielle du nom *piège*, ni l'emploi d'une relative ne nous semblent les plus fréquents pour cette EP.

(7.9) *Les* ***pièges*** *qu'on lui* ***tend*** *menacent son existence.* (`FR-train1.1`)

Cette appréciation fondée sur la fréquence dépend en outre du corpus d'où le risque d'avoir des formes canoniques différentes selon le corpus utilisé – étant donné la difficulté de garantir sa 'représentativité' –, ce qui représente un biais important.

Faute d'obtenir la forme canonique de chaque EP de façon automatique, c'est sur la forme canonique de ses composants, autrement dit leur forme lemmatisée, que nous nous appuierons. La lemmatisation neutralise la flexion morphologique mais demeure toutefois une approximation : dans ***sucrer les fraises***, lemmatisé sous la forme *sucrer le fraise*, on perd l'information sur l'idiosyncrasie morphologique du nom. Outre cette lemmatisation des composants, il faut également s'affranchir de l'ordre des composants. En effet, cet

6. Définir une 'forme canonique' est un peu compliqué. Il s'agit essentiellement de la forme sous laquelle un idiome apparaît le plus fréquemment dans les données, moyennant la flexion de la tête (s'il y en a une). Par exemple, la forme canonique de l'idiome **spill the beans** est **spill/spills/spilled/spilling the beans**, tandis que **spilling** many **beans**, **spilled the** royal **beans**, et **the beans** were **spilled** sont des occurrences non-canoniques. Toutefois, définir une 'forme canonique' de cette façon serait circulaire, car cette forme serait par définition la plus fréquente [Notre traduction].

ordre est fréquemment variable pour les EP verbales ne serait-ce qu'en raison de phrases interrogatives (***y-a-t-il*** vs. ***il y a***).

Dans ce but, nous qualifions de *forme lemmatisée normalisée* (dorénavant *LemmNorm*) la séquence constituée des composants lemmatisés l_i triés par ordre lexicographique et que nous présentons sous la forme $\langle l_1 ; l_2 ; \ldots ; l_n \rangle$.

Pour les exemples mentionnés, deux entrées seraient donc créées selon cette procédure : ⟨avoir ;il ;y⟩ et ⟨contenter ;se⟩, ce qui permet de regrouper toutes les variantes d'une EP sous une entrée unique à la façon d'une *forme canonique*. Un défi supplémentaire est lié à la fusion inappropriée de variantes pouvant se produire lors de l'étape *LemmNorm*. En effet, une même *LemmNorm* telle que ⟨appel ;faire⟩ peut correspondre à différentes EP : ***faire appel*** *d'une décision* ⇒ 'contester' vs. ***faire appel*** *à quelqu'un* ⇒ 'solliciter' ou bien supprimer une distinction morphologique discriminante : ***faire la course*** vs. ***faire les courses***, quoique de tels cas restent rares. Cette *LemmNorm* n'est pas à proprement parler une *forme canonique* car l'ordre généré et la non-prise en compte des flexions conduisent à des créations non-naturelles (contrairement à une entrée de lexique par exemple), mais ici l'unique but recherché est la fusion de variantes sous une même entrée.

7.4 Esquisse de méthode : la variabilité au service de l'identification de variantes

Identifier une EP préalablement référencée revient à déterminer si la présence simultanée de ses composants dans une phrase donnée signale bien une EP. Or cette seule co-occurrence des composants n'est pas suffisante comme le prouvent les lectures littérales et les co-occurrences fortuites. C'est pour cette raison que, à l'instar de Nissim et Zaninello (2013) pour les EP nominales, nous considérons que la connaissance de l'idiosyncrasie d'une EP joue un rôle déterminant dans l'identification, et qu'il doit exister des "familles" d'idiosyncrasie, ce que nous chercherons à modéliser grâce à la notion de profil de variabilité. Prenons l'exemple de l'expression ***il y avoir***, les 167 tokens observés dans le corpus `FR-train1.1` sont tous fléchis à la 3ème personne du singulier, comme pour l'EP ***il en est***[7]. On en déduit qu'une séquence fléchie à une autre personne a moins de chances d'être une EP (*ils y ont droit, ils en sont contents*). L'idiosyncrasie des EP peut donc être apprise d'après des corpus annotés en EP, sous réserve que ces corpus soient suffisamment volumineux. En effet, une flexion non observée en corpus n'implique pas nécessairement qu'elle est interdite par l'EP.

Rosén *et al.* (2016) ont établi une liste de 23 corpus annotés à la fois en syntaxe (p. ex. dépendances syntaxiques) et en EP (nominales, verbales, prépositionnelles, adjectives, proverbes, etc.) pour différentes langues. Pour le français, seul le *French Treebank* – composé d'environ 20 000 phrases extraites du journal *Le Monde* – (Abeillé *et al.*, 2003) apparaît mais, d'une part les EP verbales se limitent aux idiomes, d'autre part ils sont rarement discontinus (quelques dizaines de cas, par exemple *ne* ***fait*** *pas* ***fi*** *de*) (Abeillé et Clément, 2003). En comparaison, sur les 17 000 phrases d'un corpus PARSEME à notre

7. Par exemple dans : ***Il en est*** *de même pour le master recherche.*

disposition [8], nous dénombrons 412 idiomes discontinus. En conclusion, faute de couverture suffisante à la fois des catégories d'EP verbales (absence d'annotation des constructions à verbe support dans le *French Treebank* par exemple) et du phénomène de variabilité dans cette ressource, il nous semble plus pertinent de nous restreindre aux corpus mis à disposition dans le cadre des compétitions PARSEME (Savary *et al.*, 2017; Ramisch *et al.*, 2018).

8. Ce corpus, noté `FR-train1.1`, est décrit dans la section 8.1.3.

Chapitre 8

Corpus utilisés

Pour mener à bien la tâche d'identification de variantes d'EP, il nous faut bénéficier d'exemples d'emplois de ces mêmes EP afin qu'elles nous servent de référence. Nous avons choisi pour cela de nous appuyer sur des corpus annotés en EP provenant des deux éditions de la compétition PARSEME (Savary *et al.*, 2017; Ramisch *et al.*, 2018). Dans la section 8.1 consacrée à ces corpus, nous présentons la méthodologie d'annotation suivant les choix de catégorisation d'EP opérés dans le cadre de PARSEME – et que nous avons ici adoptés – ainsi que les informations relatives à leur constitution et à leurs statistiques. Un autre corpus, de taille très supérieure mais non annoté en EP, a également été exploité afin de permettre l'évaluation de notre système sur un corpus externe (section 8.2).

8.1 Corpus PARSEME

La tâche d'identification de variantes d'EP consiste à reconnaître en contexte des EP connues. Disposer de cette liste d'EP connues représente donc un pré-requis, cette liste pouvant être extraite à partir d'un corpus annoté en EP. Or un tel corpus doit être de bonne qualité au risque de compromettre la qualité d'identification.

8.1.1 Processus d'annotation

La seule façon de garantir la qualité d'annotation de corpus est de l'effectuer manuellement, ce qui pose plusieurs problèmes. Il s'agit en effet d'une opération chronophage, ce qui peut favoriser l'annotation de corpus de taille réduite et/ou le recours à plusieurs annotateurs, devant donc respecter une même procédure d'annotation. Une telle tâche d'annotation de corpus a été effectuée dans le cadre de la campagne multilingue PARSEME (Savary *et al.*, 2017; Ramisch *et al.*, 2018). La partie française du corpus de l'édition 1.0 (Candito *et al.*, 2017) résulte de la fusion des corpus Séquoia (Candito et Seddah, 2012) et la partie française de *Universal Dependencies* (UD) (Nivre *et al.*, 2016) :

> [L]a partie française du corpus UD [...] comprend 16 447 phrases françaises extraites au hasard de *Google News, Blogger, Wikipedia* et des avis de consommateurs ; [...] le corpus Sequoia [...] comprend 3 099 phrases issues de l'*Est*

> *Républicain*, de rapports de l'*Agence Européenne du Médicament*, de *Wikipedia* et d'*Europarl*. Pour la campagne PARSEME, les 500 premières EP ont été réservées comme corpus d'évaluation des systèmes participants (TEST). Le restant du corpus a été considéré comme corpus d'entraînement pour les systèmes (TRAIN). (Candito *et al.*, 2017)

Quant au corpus français de l'édition 1.1 :

> The French corpus contains the Sequoia corpus (Candito et Seddah, 2012) converted to UD, the GDS French UD treebank, the French part of the ParTUT corpus, and part of the Parallel UD (PUD) corpus. [1]

Chacun des corpus TRAIN, DEV et TEST de l'édition 1.1 provient des trois corpus Sequoia, GDS et ParTUT, mais nous ne bénéficions pas d'informations détaillées sur leur constitution, hormis le fait que le corpus de TEST bénéficie en outre des 500 premières phrases du corpus PUD.

Les autres langues s'appuient sur des corpus UD, des corpus nationaux ou la compilation de textes provenant d'Internet. Il en résulte une grande diversité dans la nature des textes : littéraires, juridiques, journalistiques, encyclopédiques (*Wikipédia*) ou blogs (Ramisch *et al.*, 2018). Nous ferons référence aux corpus de cette seconde édition sous la forme `xx-train1.1, xx-dev1.1` et `xx-test1.1`, xx devant être substitué par l'abréviation relative à chaque langue [2].

Ces corpus offrent l'intérêt de fournir des informations détaillées établies manuellement [3] et/ou de façon automatique avec des outils tels que UDPipe, qu'il s'agisse d'étiquetage morphosyntaxique ou des relations de dépendances syntaxiques [4] au format UD [5]. Cette annotation en EP a été effectuée grâce à une plate-forme dédiée sur Internet (Fig. 8.1) : les composants de chaque EP sont signalés (p. ex. ***joué rôle***) et une étiquette de catégorie – parmi celles détaillées dans la section 8.1.2 – leur est attribuée (p. ex. 'LVC.full'). Outre ces informations, les corpus comportaient initialement, pour chaque mot, sa forme fléchie (p. ex. *éponges*), sa forme lemmatisée (p. ex. *éponge*), sa POS (p. ex. NOUN), l'élément régissant du mot (p. ex. le verbe *jeter*), le type de dépendances syntaxiques (relation sujet/objet par exemple).

A la différence d'un corpus textuel, ces fichiers ne permettent pas toujours de bénéficier d'informations contextuelles sur plusieurs phrases : certaines phrases des corpus PARSEME [6] n'ont parfois – pour le français tout du moins – aucune relation avec les phrases les précédant, ni avec celles qui les suivent [7].

1. Le corpus français comporte le corpus Sequoia avec les jeux d'étiquettes convertis au format UD, le GDS French UD treebank, la partie française du corpus ParTUT, et une partie du corpus parallèle UD (PUD). [Notre traduction]

2. Par exemple : FR pour le français. Les autres langues proposées sont : arabe (AR), bulgare (BG), allemand (DE), grec (EL), anglais (EN), espagnol (ES), basque (EU), farsi (FA), hindi (HI), hébreu (HE), croate (HR), hongrois (HU), italien (IT), lituanien (LT), polonais (PL), portugais du Brésil (PT), roumain (RO), slovène (SL) et turc (TR).

3. Par exemple pour le français, le hongrois et le slovène.

4. Sauf pour le lituanien.

5. `https://universaldependencies.org/format`

6. Ce sont plus précisément celles issues du corpus UD-GSD (*Google Stanford Dependencies*) dont le choix, peu usuel, a été d'annoter des phrases au hasard.

7. Ce corpus fournit un échantillon représentatif d'emplois langagiers (Sinclair, 1996) sans souscrire à

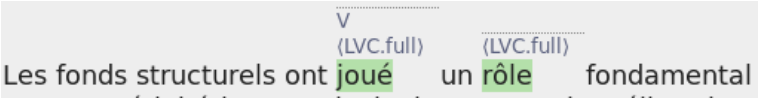


FIGURE 8.1 – Exemple d'annotation sur la plate-forme FLAT : l'annotateur humain sélectionne les composants de chaque EP (ici en surbrillance) et leur attribue une étiquette en fonction de leur catégorie (LVC.full, ID, etc.).

En raison de la taille des corpus à annoter, plusieurs annotateurs se sont relayés et bien que se conformant au guide d'annotation, l'interprétation du figement de certains composants reste parfois subjective, par exemple la prise en compte de la négation dans *ça* ***(ne) paie pas de mine*** ou l'annotation de l'article dans ***poser un lapin*** vs. *ne pas* ***poser*** *de* ***lapin***. C'est pourquoi il est intéressant d'évaluer la cohérence d'annotation grâce la mesure d'accord inter-annotateur. Cette évaluation – à laquelle nous avons participé pour le français en tant que second annotateur – a été effectuée sur un échantillon du corpus (cf. section 10.2.2). Notons que, quoique de taille réduite, ce corpus est sans équivalent dans l'espace francophone.

8.1.2 Catégories d'EP utilisées

Concernant la classification des EP, nous nous appuyons sur la terminologie utilisée lors de la seconde campagne d'annotation de corpus PARSEME dont le projet éponyme signifie *PARSing and Multi-word Expressions* (Savary *et al.*, 2017). L'intérêt majeur de cette terminologie est sa capacité de généralisation puisqu'elle résulte d'un compromis impliquant au minimum les 19 langues participant au projet et, de plus, elle se focalise sur notre objet d'études : les EP verbales, c'est-à-dire comportant au minimum un verbe et fonctionnant comme un verbe (ce qui exclut des EP à valeur de substantif comme ***laissez-passer***). En tenant uniquement compte des catégories utiles pour le français, les EP y sont classées selon quatre catégories. De façon succinte – car la proposition d'une catégorisation n'est pas l'objet de notre contribution – on distingue :

— Les *constructions à verbe support* (de l'anglais *light verb constructions* ou *LVC*) elles-même subdivisées en 2 catégories : les *LVC.full* et les *LVC.cause*. Dans les constructions de catégorie LVC.full, le verbe ne doit pas être aspectuel (*commencer, continuer, cesser, finir*). De plus, contrairement à la définition classique des CVS [8], on ne se limite pas aux verbes fréquemment vides de sens tels que *avoir* dans ***avoir accès***. En effet, un verbe est considéré *light* si le nom se suffit à lui-même comme porteur du sens comme dans ***apporter témoignage, livrer bataille*** ou ***courir risque***, et cela bien que les verbes *apporter, livrer* ou *courir* ne soient pas eux-mêmes seulement légers. Les LVC incluent également les LVC.cause dans lesquelles le verbe ajoute un sens causatif au nom comme dans ***donner*** *le* ***droit*** ou ***entraîner***

la définition de Rastier (2005) qui décrit un corpus comme étant un "regroupement structuré de textes intégraux" [Notre soulignement].

8. Nous réservons cet acronyme à la définition traditionnelle des constructions à verbe support présentée en section 2.1.7.

la ***fatigue****.*

— Les *idiomes verbaux*, désormais abrégés VID (de l'anglais *verbal idiom*). Leur sens n'est pas compositionnel et ils possèdent souvent une double lecture littérale / idiomatique (***tourner la page***) nécessitant une désambiguïsation contextuelle. Dans la terminologie PARSEME, cette catégorie ne coïncide pas systématiquement avec celle traditionnellement admise. Par exemple ***faire partie*** est ici considéré comme un VID et non une LVC car il n'autorise pas l'omission du verbe, contrairement à ***prendre décision*** qui est un exemple de LVC : *Pierre* ***prend*** *une* ***décision*** $\rightarrow$ *la décision de Pierre* vs. *Pierre* ***fait partie*** *du groupe* $\rightarrow$ # *la partie de Pierre.*

— Les *verbes intrinsèquement réflexifs*, désormais abrégés IRV (de l'anglais *Inherently Reflexive Verbs*), satisfont l'une des trois conditions suivantes :
  — ils possèdent un sens complètement différent de celui du verbe seul (c.-à-d. hors pronom), comme ***se rendre*** $\Rightarrow$ 'aller' $\neq$ *rendre*,
  — ils n'existent pas sans *se* : ***se prélasser***/ **prélasser*,
  — ils possèdent un cadre de sous-catégorisation spécifique (***se confesser*** *de X* / *confesser* X) tout en conservant le même sens.

— Les *constructions multi-verbes* (en anglais *multi-verb constructions* ou MVC) constituées de deux verbes adjacents dont l'un gouverne l'autre comme ***vouloir dire*** $\Rightarrow$ 'signifier' ou ***laisser tomber*** $\Rightarrow$ 'abandonner'.

Pour plus de détails, on pourra se référer au guide d'annotation PARSEME [9].

8.1.3 Corpus PARSEME du français

8.1.3.1 Description des sous-corpus

Dans le cadre du projet PARSEME, deux compétitions portant sur l'identification automatique d'EP dans plusieurs langues de familles différentes (romanes, germaniques, slaves, etc.) se sont déroulées en 2017 (18 langues) et 2018 (20 langues) (Savary *et al.*, 2017; Ramisch *et al.*, 2018). Pour chaque édition, le corpus PARSEME se divise en sous-corpus : le corpus dit d'*entraînement* (`FR-train`), le corpus dit de *développement* (`FR-dev`) (uniquement pour l'édition 1.1) et le corpus dit de *test* (`FR-test`).

Le corpus `FR-train` donne accès à des EP manuellement annotées, ce qui permet d'extraire des connaissances sur leurs propriétés et d'en tirer ensuite profit pour l'identification d'EP dans un corpus différent (`FR-dev` ou `FR-test`). La fonction de chacun de ces corpus correspond à celle qui leur est traditionnellement attribuée. En effet, le corpus `FR-train` permet d'*entraîner* les systèmes d'identification automatique sur des données connues. Ensuite, durant la phase de *développement*, le corpus `FR-dev` permet d'évaluer en conditions réelles les performances du système développé afin de l'optimiser. Enfin, puisque la compétition a pour objectif de mettre au point des systèmes d'identification et de les évaluer (les *tester*), cette évaluation est réalisée par comparaison entre les annotations automatiquement produites sur `FR-test` et les annotations manuelles servant de référence (non disponibles durant la phase de compétition).

Les corpus des deux éditions étant différents, chaque corpus sera associé à l'édition correspondante, respectivement 1.0 (édition 2017) et 1.1 (édition 2018), p. ex. `FR-train1.1`

9. `https://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1`

pour le corpus français d'entraînement de la seconde édition.

8.1.3.2 Comparaison entre les deux éditions

Si les données brutes pour les corpus `FR-train1.1`, `FR-dev1.1` et `FR-test1.1` de l'édition 1.1 sont de nature similaire à celles de l'édition 1.0, les catégories d'EP diffèrent entre les deux éditions. Les catégories présentées en section 8.1.2 – LVC [10], VID, IRV et MVC – relèvent en effet de l'édition 1.1. Pour l'édition antérieure 1.0, les catégories étaient définies de façon légèrement différente (Candito *et al.*, 2017; Ramisch *et al.*, 2018) :

— *Constructions à verbes support* ($LVC_{1.0}$) : cette catégorie correspond aux LVC.full de l'édition 1.1,
— *Expressions verbales idiomatiques* ($ID_{1.0}$) : cette catégorie inclut les VID de l'édition 1.1 ainsi que certaines LVC.cause qui ne bénéficiaient pas encore d'un traitement spécifique lors de l'édition antérieure,
— *Verbes intrinsèquement pronominaux* ($IReflV_{1.0}$) : cette catégorie correspond aux IRV de la seconde édition,
— *Autres* (en anglais *other* ou $OTH_{1.0}$) : cette catégorie – rarement utilisée – couvre les cas d'EP comportant des verbes coordonnés (***aller et venir***) ou séparés par un trait d'union (***court-circuiter***). Pour le français, ces expressions sont devenues des VID dans la campagne 1.1.

Le corpus de la seconde édition représente la forme la plus aboutie des choix de catégorisation d'EP verbales et aurait idéalement été mise à profit pour nos analyses. Cependant, sa construction s'est opérée de façon simultanée au déroulement de cette thèse, d'où les deux versions 1.0 et 1.1 qui figurent ici. Les correspondances approximatives entre les catégories des deux éditions sont résumées dans la Table 8.1. Les statistiques portant notamment sur la distribution de ces catégories en corpus figurent dans les Tables 8.2 [11] et 8.3. Les trois principales catégories de chaque édition ($LVC_{1.0}$, $ID_{1.0}$, $IReflV_{1.0}$ et $LVC_{1.1}$, $VID_{1.1}$, $IRV_{1.1}$) couvrent une proportion relativement équilibrée d'occurrences. On remarque par ailleurs qu'environ 2% des tokens du corpus font partie de ces EP, ce qui laisse supposer que la majorité des EP ne sont pas verbales si l'on se réfère au pourcentage de 40% d'EP relevé par Gross et Senellart (1998).

Édition 1.1	Édition 1.0
$LVC.full_{1.1}$	$LVC_{1.0}$
$LVC.cause_{1.1}$	$ID_{1.0}$
$VID_{1.1}$	$ID_{1.0}$, $OTH_{1.0}$
$IRV_{1.1}$	$IReflV_{1.0}$
$MVC_{1.1}$	$ID_{1.0}$

Table 8.1 – Correspondances approximatives entre les catégories des éditions 1.0 et 1.1

Les corpus exploités s'inscrivent dans le cadre d'une étude en synchronie. Ce sont essentiellement les versions françaises des différents corpus que nous avons utilisées, tout

10. LVC.full et LVC.cause.
11. Sur les 4462 tokens d'EP annotés, il en reste 4441 occurrences après suppression de phrases doublonnées.

	# Phrases	# Tokens	#LVC$_{1.0}$		# ID$_{1.0}$		# IReflV$_{1.0}$		# OTH$_{1.0}$		# Total
			occ.	%	occ.	%	occ.	%	occ.	%	
FR-train1.0	17 880	450 221	1 633	36,60	1 786	40,03	1 313	29,43	1	0,02	4 462
FR-test1.0	1 667	35 784	271	54,20	119	23,80	105	21	5	1	500
Total	19 547	486 005	1 633	32,91	1 905	38,39	1 418	28,58	6	0,12	4 962

TABLE 8.2 – Statistiques du corpus français de l'édition PARSEME 1.0 : nombre de phrases, de tokens et distribution des EP par catégorie.

	# Phrases	# Tokens	# LVC$_{1.1}$				# VID$_{1.1}$		# IRV$_{1.1}$		# MVC$_{1.1}$		Total
			cause		full								
			occ.	%	occ.	%	occ.	%	occ.	%	occ.	%	
FR-train1.1	17225	432389	1470	32,31	68	1,49	1746	38,37	1247	27,41	19	0,42	4 550
FR-dev1.1	2236	56254	252	40,06	15	2,38	207	32,91	154	24,48	1	0,16	629
FR-test1.1	1606	39489	160	26,76	14	2,81	212	42,57	108	21,69	4	0,08	498
Total	21067	528132	1882	33,15	97	1,71	2165	38,14	1509	26,58	24	0,42	5 677

TABLE 8.3 – Statistiques du corpus français de l'édition PARSEME 1.1 : nombre de phrases, de tokens et distribution des EP par catégorie.

en cherchant à rendre possible une extension multilingue. Notons d'ailleurs que, malgré les nombreux travaux sur les EP dans la littérature, la plupart ne concernent pas le français. Or, par comparaison avec l'anglais, il s'agit d'une langue morphologiquement plus riche avec la présence d'un genre (*violet(tte)* vs. (EN) *violet*) et de nombreuses désinences verbales (p. ex. *chant-e / es / ons / ez / ent* vs. (EN) *sing(s)*) dont témoignent par ailleurs les 104 modèles de conjugaison de Delaunay et Laurent (2012). Nous soulignons par ailleurs que cette thèse se place dans le cadre de la syntaxe de dépendance : notre orchestration des tâches est une identification de variantes post-parsing (cf. Section 6.2.1.2) car nous exploitons l'analyse syntaxique opérée sur les corpus.

Outre les corpus PARSEME, nous avons également fait appel à un autre corpus, similaire au niveau du format, mais dépourvu d'annotation en EP, et dont l'avantage majeur est d'être de taille nettement supérieure.

8.2 Corpus CoNLL17 et WebSample

Le corpus de la compétition Conll 2017, désormais désigné sous le nom CoNLL17, provient de l'exploration d'Internet et du site *Wikipedia* [12]. Il est disponible en ligne en accès libre [13] et couvre 45 langues [14]. Avec ses 306 431 406 phrases et ses 5 242 235 570 mots, la partie française est 10 000 fois plus volumineuse que le corpus FR-corpus1.1. Comme tout l'étiquetage de corpus (morphologie, dépendances syntaxiques) a été obtenu automatiquement grâce à UDPipe [15], nous avons jugé pertinent d'estimer manuellement sa qualité

12. Définitions fournies par http://www.wikipedia.fr
13. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989
14. Ce corpus est bénéficie aussi de plongements de mots mais nous n'y avons pas fait appel dans le cadre de cette thèse.
15. http://ufal.mff.cuni.cz/udpipe

	candidats de FR-train1.1			candidats de WebSample		
Cat.	**# types**	**# occ.**	**%**	**# types**	**# occ.**	**%**
$VID_{1.1}$	180	1404	39,2	30	1900	40,9
$IRV_{1.1}$	144	1136	31,7	29	1508	32.5
$LVC_{1.1}$	263	1024	28,6	27	1213	26.1
$MVC_{1.1}$	4	18	0,5	4	24	0,5
Total	591	3582	100	90	4645	100

TABLE 8.4 – WebSample par rapport à FR-train1.1 (EP vues au moins deux fois)

d'après un échantillon de 4645 phrases, nommé WebSample. WebSample a été élaboré de façon à fournir un échantillon représentatif du corpus FR-train1.1 à partir du corpus CoNLL17. Concrètement, WebSample respecte la distribution observée dans FR-train1.1, qu'il s'agisse de la fréquence des EP ou de leur catégorie. Dans ce but, nous avons d'abord vérifié la proportion des 4 catégories ($VID_{1.1}$, $LVC_{1.1}$, $IRV_{1.1}$ et $MVC_{1.1}$) parmi les EP vues au moins deux fois dans FR-train1.1. 90 types d'EP [16] sont ensuite sélectionnés de telle façon que ces proportions soient respectées. Pour les $MVC_{1.1}$, aucune sélection n'est effectuée en raison de leur rareté, nous nous contentons de conserver les 4 uniques types du corpus FR-train1.1. Cette sélection de types d'EP tient également compte de leur niveau de fréquences dans FR-train1.1. Pour éviter de biaiser WebSample en privilégiant des types d'EP très ou peu fréquents, nous recherchons 1/3 d'EP fréquentes, 1/3 de fréquence médiane et 1/3 d'EP rares. Nous faisons également en sorte de constituer un corpus équilibré au niveau des deux sources du corpus CoNLL17 : 50% provenant de *Wikipédia* et 50% d'exploration du Web. Finalement, ces 90 types servent d'ensemble de référence pour une méthode d'extraction de candidats [17], qui, une fois appliquée au corpus CoNLL17, fournit 4645 candidats dont la distribution par catégorie figure dans la Table 8.4. Après vérification manuelle, les 4618 candidats valides sont manuellement annotés selon le guide d'annotation PARSEME par 4 experts : les exemples satisfaisant les conditions pour être annotés comme EP couvrent 68,7% de l'ensemble. Ce corpus [18] est l'une de nos contributions vis-à-vis de l'enrichissement des ressources existantes pour le traitement automatique des EP.

Nous avons remarqué que certaines phrases étaient incomplètes (p. ex. *"La a eu lieu le"*), peut-être en raison d'un prétraitement dont nous ignorons la nature. Après examen de WebSample, il apparaît que seules 27 phrases (0,6%) ne sont pas exploitables, en raison d'incorrections, d'erreurs de POS ou de lemmes, de phrases dans d'autres langues que le français ou de contexte insuffisant pour la tâche. Malgré ce faible taux de rejet, rien ne permet de déterminer si un corpus tel que CoNLL17 permettra l'obtention de profils de variabilité fiables. On suppose cependant que la finesse de profils de variabilité d'EP est *a priori* fortement dépendante du volume d'informations disponibles. C'est pourquoi la question de la variabilité des EP en corpus sera abordée dans la partie III.

16. cf. Annexe A
17. ExtractCands2, présentée en section 13.1.
18. disponible à l'adresse : https://gitlab.com/cpasquer/websample

Troisième partie

Où la variabilité des EP *a voix au chapitre*

Cette partie explore, dans le chapitre 9, la pertinence de l'hypothèse principale H_1 (mentionnée en Introduction, p. 27) sur l'existence d'un profil de variabilité grâce à une étude de la variabilité des EP en corpus :

H1 Chaque EP a un profil de variabilité (c.-à-d. un ensemble de transformations autorisées) qui lui est propre et qui est attesté par un éventail plus ou moins étendu de réalisations (*tokens*). A l'inverse, des séquences ressemblantes mais ne relevant pas du statut d'EP ne respectent pas ce profil. Par exemple, l'EP ***jeter l'éponge*** ne conserve pas son statut d'EP en cas de modification du nom par un adjectif (*jeter l'éponge verte*), alors que cette transformation n'a aucune incidence pour la séquence *jeter la serpillière (verte)*.

Les chapitres 10 et 11 se focaliseront respectivement sur l'hypothèse H_2 (voir Introduction, p. 27) ainsi que sur la sous-hypothèse H_{2a} qui en découle :

H2 On peut modéliser ce profil de variabilité comme un phénomène multidimensionnel d'après les informations morphosyntaxiques fournies par des corpus annotés, par exemple le type de discontinuités, l'existence de modifieurs ou bien encore la variation interne des composants.

H2a La variabilité est quantifiable car le profil de variabilité peut donner lieu à une mesure de l'étendue de cette variabilité. On s'attend à observer des différences significatives entre catégories d'EP (constructions à verbe support plus variables que les idiomes) voire entre EP d'une même catégorie (l'$IRV_{1.1}$ *je me/il se prélasse* est plus variable que **je me/il se déroule*).

Chapitre 9

Variabilité observée en corpus

Pour appréhender l'importance et les caractéristiques du phénomène de variabilité, nous nous appuyons sur le corpus `FR-train1.0` annoté en EP [1]. Cela permet d'identifier les variations majoritaires pour, ensuite, orienter la modélisation, en termes de traits de description de la variabilité qui serviront à l'identification automatique. Cette étude de la variabilité, menée sur le corpus `FR-train1.0`, a fait l'objet d'un article à TALN-RECITAL (Pasquer, 2017) [2]. On s'intéresse tout autant aux composants lexicalisés [3] des EP – en se focalisant sur les EP de type VERB – (DET) – NOUN – qu'aux autres éléments, qu'ils soient situés entre ces composants et/ou qu'ils leur soient syntaxiquement liés, afin d'obtenir un aperçu de la variabilité à la fois d'un point de vue morphologique (section 9.1) et syntaxique, en exploitant les discontinuités (section 9.2) et les relations de dépendances syntaxiques (section 9.3).

9.1 Variabilité morphologique des composants d'EP de patron VERB-(DET)-NOUN

Les EP de patron VERB – (DET) – NOUN représentent 96% des tokens de $\text{LVC}_{1.0}$ et 44% des tokens d'$\text{ID}_{1.0}$. Ce patron couvre donc une proportion importante des occurrences observées en corpus, mais exclut totalement les $\text{IReflV}_{1.0}$ dont le patron majoritaire est PRON – VERB (p. ex. ***se trouver***). Nous nous intéressons ici à la variation morphologique des composants principaux des EP de patron VERB – (DET) – NOUN, c'est-à-dire à la variation verbale et nominale. La morphologie du déterminant dépend en effet de celle du nom.

9.1.1 Variation verbale

La variation verbale s'observe à deux niveaux : la flexion personnelle, temporelle et/ou de mode d'une part, et la dérivation par préfixation d'autre part :

1. Seule cette version du corpus était disponible au moment de l'étude.
2. La section 9.3 figurant dans cette thèse constitue un ajout par rapport à cet article.
3. Désormais qualifiés simplement de *composants*.

— *Flexion personnelle, temporelle et de mode.* Certains verbes tolèrent une flexion limitée, souvent aux 3èmes personnes (*il(s)* ***se déroule(nt)***/#*je me déroule*), voire uniquement à la 3ème personne du singulier (***il pleut***). Parmi les tokens d'$ID_{1.0}$, 29% (518 tokens) relèvent de tournures impersonnelles (***il y a, il faut, il s'agit, il pleut des cordes, sera-t-il question,...***). Par ailleurs, quelques EP se distinguent par une flexion temporelle et/ou modale limitée (***qu'importe/qu'importait***[4]/#*qu'importa*) ou interdite (***honni soit qui mal y pense*** vs. #*honni fût qui mal y pensât*). Certaines EP opèrent également une restriction du mode : impératif (***arrête(z) ton (votre) char*** ⇒ 'arrête(z) de dire n'importe quoi') ou subjonctif par exemple (***quelque soit/ fût***).

— *Préfixation du verbe.* 7 EP du corpus comportent un verbe préfixé avec *r(e)-* (p. ex. ***(re)donner raison***, ***(re)faire appel***). S'il s'agit ici de variantes d'EP non préfixées, d'autres EP nécessitent au contraire ce préfixe (***revenir à la raison*** ≠ venir à la raison). La question de la double lecture littérale / idiomatique se pose également dans ***relever la tête*** puisqu'au sens littéral *lever* et *relever* sont des variantes alors que le préfixe est obligatoire dans l'EP. Prêter attention à ce phénomène pourrait permettre, à partir d'un lexique préétabli d'EP, d'en identifier des variantes pour des verbes *a priori* très productifs comme *faire* ou *donner*. Notons cependant que la préfixation peut engendrer des EP ayant une signification bien distincte : ***mettre en cause*** ⇒ 'impliquer' vs. ***remettre en cause*** ⇒ 'réévaluer'.

9.1.2 Variation nominale

L'étude de la variation flexionnelle du nom a été restreinte aux EP de patron VERB − (DET) − NOUN apparaissant au moins 10 fois en corpus (Table 9.1). Cet examen porte sur 12 types d'$ID_{1.0}$ (356 tokens) et 17 de $LVC_{1.0}$ (274 tokens) qui représentent chacun 20% des tokens de leur catégorie dans le corpus.

Pour la plupart des EP, l'examen du corpus suffit à mettre en évidence la capacité de flexion en nombre du nom. Mais, en raison de la taille réduite du corpus, il faut parfois recourir à des ressources complémentaires pour confirmer l'invariabilité. Ainsi, 12 des 17 LVC présentent une variabilité de la flexion en nombre du nom dans le corpus, et 3 autres après consultation d'un moteur de recherche sur Internet (Ex. 9.1 à 9.3), tandis que les 2 LVC restantes (***faire appel***, ***avoir*** *(le)* ***droit***) apparaissent morphologiquement rigides.

(9.1) (a) *Elle* ***fait*** *également une* ***apparition*** *au festival* (`FR-train1.0`)
(b) *Cet artiste* [...] *a déjà* ***fait*** *des* ***apparitions*** *au Caveau de la Gare* (Web)

(9.2) (a) *l'occasion de lui* ***rendre*** *un dernier* ***hommage*** (`FR-train1.0`)
(b) [...] *et leur* ***rend*** *les derniers* ***hommages*** (Web)

(9.3) (a) *J'****ai*** *plutôt* ***tendance*** *à penser qu'il s'agit d'un coup monté* (`FR-train1.0`)
(b) *J'****ai*** *des* ***tendances*** *à être claustrophobe* (Web)

De la même façon, un examen des 11 $ID_{1.0}$ apparemment figés en corpus met en évidence une flexibilité nominale de l'$ID_{1.0}$ ***prendre nom*** sur Internet[5]. En conclusion, seuls 10% des tokens de catégorie $ID_{1.0}$ sont morphologiquement variables au niveau du nom,

4. Dans *L'Étranger* (Camus) par exemple : ***Qu'importait*** *si, accusé de meurtre, il était exécuté.*
5. Par exemple : *Deux villes qui ont* ***pris*** *les* ***noms*** *des tribus indiennes vivant sur le site.*

12 types d'$ID_{1.0}$		17 types de $LVC_{1.0}$	
avoir lieu	95 occ.	***jouer rôle***	33 occ.
faire partie	88 occ.	***faire appel***	32 occ.
faire l'objet	27 occ.	***avoir besoin***	28 occ.
mettre fin	23 occ.	***avoir droit***	17 occ.
faire face	20 occ.	***jouer match***	16 occ.
porter nom	19 occ.	***faire apparition***	15 occ.
prendre part	17 occ.	***signer contrat***	15 occ.
tenir compte	17 occ.	***disputer match***	15 occ.
mettre un terme	13 occ.	***poser question***	14 occ.
prendre nom	13 occ.	***faire référence***	13 occ.
donner lieu	12 occ.	***rendre hommage***	12 occ.
donner naissance	10 occ.	***lancer appel***	11 occ.
		avoir chance	11 occ.
		avoir tendance	11 occ.
		prendre mesure	11 occ.
		marquer but	10 occ.
		avoir effet	10 occ.
Total	356 occ.	Total	274 occ.

TABLE 9.1 – EP de patron VERB – (DET) – NOUN les plus fréquentes dans `FR-train1.0`

alors que cette proportion atteint 82% pour les $LVC_{1.0}$. Malgré le recours à Internet pour confirmer/infirmer l'invariabilité, la représentativité du corpus n'est pas remise en question car la majorité des types d'EP (15/17 pour les $LVC_{1.0}$ et 11/12 pour les $ID_{1.0}$) présente une variabilité en corpus conforme à ce que nous observons à plus grande échelle sur Internet.

La variabilité des EP ne se restreint pas à la variabilité de leurs constituants principaux (ici le verbe et le nom) : il faut aussi prendre en compte la variabilité externe, autrement dit l'existence de discontinuités en raison de l'insertion d'éléments entre les composants. Ces discontinuités constituent en effet un problème majeur pour certains outils de TAL, notamment ceux reposant sur un étiquetage séquentiel.

9.2 Discontinuités

Si les composants représentent les constituants essentiels des EP, les autre éléments sont également instructifs, notamment lorsqu'ils interviennent entre les composants.

En pratique, l'ensemble des éléments compris entre le premier et le dernier composant (discontinuités incluses) constitue ce que nous appelons *fenêtre d'annotation*. Celle-ci est matérialisée par des crochets (Ex. 9.4-9.5). Ces discontinuités, qui peuvent être contiguës (Ex. 9.4) ou non (Ex. 9.5), dénotent une plage de variabilité.

(9.4) *C'est ce que* [***mettent*** *à mon avis* ***en lumière***] *quelques propositions d'amendement.* (`FR-train1.0`)

(9.5) *Le président du tribunal* [***procède*** *alors* ***à*** *la* ***lecture***] *des chefs d'accusation.* (`FR-train1.0`)

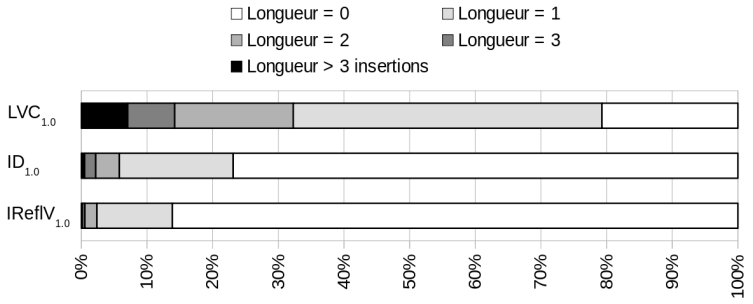


FIGURE 9.1 – Fréquence de la longueur des discontinuités par catégorie d'EP dans `FR-train1.0`

9.2.1 Discontinuités par catégories d'EP

Sur les 4441 tokens d'EP du corpus `FR-train1.0`, 62% (2772 tokens) sont dépourvus de discontinuités, et 25% (1097 tokens) comportent un unique élément inséré. Autrement dit, les discontinuités de longueurs 0 et 1 couvrent à elles-seules 87% du corpus. D'après la distribution du nombre de discontinuités par catégorie d'EP, la Figure 9.1 montre que, par ordre de variabilité croissante, on obtient : les $\text{IReflV}_{1.0}$, puis les $\text{ID}_{1.0}$ et enfin les $\text{LVC}_{1.0}$. En effet, seuls 21% des tokens de $\text{LVC}_{1.0}$ sont continus, tandis que cette proportion atteint 86% pour les $\text{IReflV}_{1.0}$. Dans le cas des discontinuités de longueur 1 (Table 9.2), les différences de discontinuités (sous forme de POS) sont également significatives :

- $\text{IReflV}_{1.0}$: dans 79% des cas, il s'agit de l'auxiliaire *être* qui reflète une variation temporelle (p. ex. *l'Empire* ***s'est***$_{\text{AUX}}$ ***écroulé***).
- $\text{LVC}_{1.0}$: dans 66% des cas, cette insertion correspond à un déterminant non lexicalisé p. ex. ***pris*** la_{DET} ***décision***)
- $\text{ID}_{1.0}$: les deux insertions uniques les plus fréquentes sont des adverbes (44%) et des déterminants (28%).

Parmi ces insertions uniques, certaines sont régulières, qu'il s'agisse de la variation temporelle des $\text{IReflV}_{1.0}$ ou de l'insertion d'adverbes (Tutin, 2016). En dehors de ces insertions uniques, nous trouvons, parmi les 12 discontinuités les plus fréquentes (Fig. 9.2) [6], 5 séquences de POS à deux éléments (Ex. 9.6-9.8), dont certaines sont fortement corrélées à la catégorie d'EP : 96% des discontinuités DET-ADJ correspondent à des $\text{LVC}_{1.0}$ et 80% des AUX-ADV à des $\text{IReflV}_{1.0}$.

(9.6) *La Bulgarie a* ***joué*** un_{DET} $grand_{\text{ADJ}}$ ***rôle***. (`FR-train1.0`)

(9.7) *Il* ***donne*** $alors_{\text{ADV}}$ sa_{DET} ***démission***. (`FR-train1.0`)

(9.8) *Des seigneurs protestants qui* ***s'***$étaient_{\text{AUX}}$ $déjà_{\text{ADV}}$ ***emparés*** *des terres.* (`FR-train1.0`)

(9.9) *La Banque populaire de Chine doit encore* ***déployer*** $beaucoup_{\text{ADV}}$ d'_{ADP} ***efforts***.(`FR-train1.0`)

(9.10) *C'est le meilleur* ***souvenir*** que_{PRON} $nous_{\text{PRON}}$ ***gardons***. (`FR-train1.0`)

6. Nous fixons arbitrairement un seuil de fréquence à 10 tokens minimum attestés dans au moins l'une des catégories d'EP.

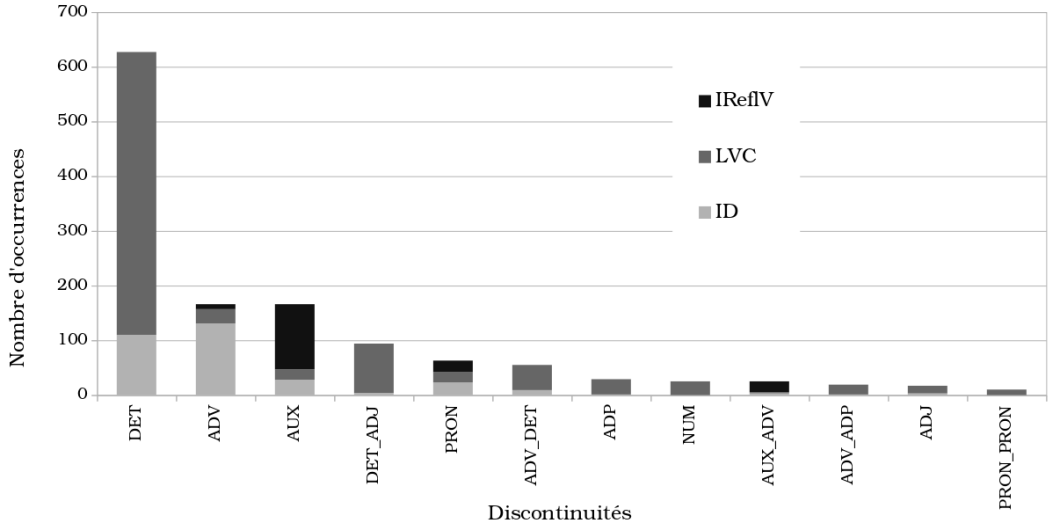


FIGURE 9.2 – Discontinuités (sous forme de séquences de POS) les plus fréquentes dans `FR-train1.0`

Ces discontinuités fournissent également un aperçu de la variabilité syntaxique : l'insertion d'un auxiliaire entre un nom et un verbe peut ainsi suggérer une passivation (p. ex. *la* ***décision*** est_{AUX} ***prise***).

9.2.2 Variation syntaxique *via* les discontinuités

Les discontinuités nous informent sur le degré de variabilité syntaxique d'une EP : plus une EP en tolère (à la fois en quantité et en variété), plus on s'attend à ce qu'elle soit flexible. Comme évoqué précédemment (page 126), les constructions à verbe support présentent une longueur de discontinuité supérieure à celle des idiomes verbaux, eux-mêmes ayant une longueur de discontinuité supérieure à celle des verbes intrinsèquement réflexifs. Nous observons une tendance similaire dans le corpus `FR-train1.1` à propos de la diversité de patrons d'insertions (évaluée d'après leurs parties de discours) : 199 patrons pour les 1470 LVC.full$_{1.1}$ (soit 13,5%), 81 patrons pour les 1746 VID$_{1.1}$ (soit 4,6%) et 13 patrons pour les 1247 IRV$_{1.1}$ (soit 1%).

Par exemple, les deux phrases (Ex. 9.11-9.12) ont un patron syntaxique similaire *prendre* + NOUN, mais la seconde est moins variable que la première, comme en témoigne l'insertion interdite d'un déterminant possessif ou de la séquence DET-ADJ, alors que toutes deux sont autorisées dans la première phrase. Seule l'insertion adverbiale (*encore*) est tolérée par les deux EP.

(9.11) *il* ***prend*** $encore_{\text{ADV}}$ / son_{DET} / un_{DET} $grand_{\text{ADJ}}$ ***essor*** (`FR-train1.0` : LVC$_{1.0}$)

(9.12) *il* ***prend*** $encore_{\text{ADV}}$ / *sa_{DET} / *une_{DET} $grande_{\text{ADJ}}$ ***conscience*** (`FR-train1.0` : ID$_{1.0}$)

L'étude des discontinuités devrait donc permettre de tirer des conclusions sur la variabilité des différentes catégories d'EP.

9.2.2.1 Relative et clivage

Les discontinuités ne permettent pas toujours d'identifier clairement le type de modification syntaxique : la séquence d'insertions PRON − PRON peut correspondre soit à un clivage (*c'est* [...] *que / qui*) comme dans l'exemple 9.10, soit à une relative (Ex. 9.13). Une classification manuelle est alors nécessaire pour distinguer ces deux cas.

(9.13) *C'est également un excellent tireur à trois points,* ***exercice*** *qu*$_{\text{PRON}}$*' il*$_{\text{PRON}}$ ***pratique*** *volontiers.* (FR-train1.0)

D'ailleurs, comme le signale Laporte (1988) au sujet du clivage, "l'extraction d'un nom est généralement interdite lorsque le déterminant est figé, et permise lorsque le déterminant et les modifieurs sont libres". De même, la présence simultanée d'un même substantif à l'intérieur et à l'extérieur d'une EP (Ex. 9.14) nous renseigne sur sa capacité à fonctionner sémantiquement de façon autonome et laisse supposer une EP de catégorie $\text{LVC}_{1.0}$.

	$\text{ID}_{1.0}$	$\text{ID}_{1.0}$	$\text{LVC}_{1.0}$	$\text{LVC}_{1.0}$	$\text{IReflV}_{1.0}$	$\text{IReflV}_{1.0}$
	occ.	%	occ.	%	occ.	%
ADJ	3	0,97	15	2,35	1	0,67
ADV	136	**44,16**	27	4,23		
REL			3	0,47		
DETERMINANT :						
article	88	**28,57**	424	**66,46**		
démonstratif			17	2,66		
indéfini	5	1,62	27	4,23		
numéral			12	1,88		
possessif	23	7,47	49	7.68		
être	6	1,95	19	2,98	119	**79,33**
avoir	13	4,22				
Modal : *devoir, pouvoir, sembler*	10	3,25				
NOM			2	0,31		
PREPOSITION	1	0,32	30	4,70		
PRONOM :	2	0,64				
démonstratif	2	0,64				
personnel	5	1,62				
en	4	1,30			16	**10,67**
y					13	8,67
le	4	1,30			1	0,67
passif SE			13	2,04		
t euphonique (p. ex. ***y a-t-il***)	8	2,60				
TOTAL	308	100	638	100	150	100

TABLE 9.2 – Nature des discontinuités de longueur 1. Les pourcentages les plus significatifs sont mis en gras.

(9.14) *Les deux pilotes déclassés* ***firent appel****, appel jugé en janvier 2007* (FR-train1.0)

9.2.2.2 Passive

Les passivations potentielles sont détectées grâce à des composants dans l'ordre Nom-Verbe. On distingue trois formes de passives :

— passives avec AUX : identifiées par la présence du verbe *être* non annoté dans la fenêtre d'annotation, comme dans *le* ***ton*** *a d'ailleurs été*$_{\text{AUX}}$ ***donné*** ($\text{LVC}_{1.0}$),
— passives sans AUX : identifiées par la terminaison du verbe, comme dans *le* ***rôle joué*** ($\text{LVC}_{1.0}$) ou ***compte-tenu*** ($\text{ID}_{1.0}$),
— passives en SE : identifiées par l'insertion de ce pronom, comme dans *la* ***finale*** *se* ***joue*** ($\text{LVC}_{1.0}$).

9.2.2.3 Fréquence des transformations syntaxiques

Outre ces trois formes de passives, les occurrences trouvées ont fait l'objet d'une vérification manuelle de façon à dresser un aperçu statistique (Table 9.3) de la relativation (Ex. 9.15) et du clivage (Ex. 9.16). Les $\text{ID}_{1.0}$ apparaissent moins variables que les $\text{LVC}_{1.0}$. Les $\text{ID}_{1.0}$ ont également davantage tendance à lexicaliser les déterminants (20 fois plus souvent) (Ex. 9.17) ou les négations (10 fois plus) (Ex. 9.18).

(9.15) ***tâche*** *qu'il a* ***accomplie*** ($\text{LVC}_{1.0}$), ***question*** *que nous devons nous* ***poser*** ($\text{ID}_{1.0}$)

(9.16) *c'est du bon* ***boulot*** *que vous m'avez* ***fait*** *(LVC$_{1.0}$)*

(9.17) ***prit la fuite*** ($\text{LVC}_{1.0}$), ***tentent le tout pour le tout*** ($\text{ID}_{1.0}$)

(9.18) ***ne pas tarir*** ($\text{ID}_{1.0}$)

Variation	$\text{LVC}_{1.0}$ occ.	$\text{LVC}_{1.0}$ %	$\text{ID}_{1.0}$ occ.	$\text{ID}_{1.0}$ %
passive avec AUX	59	4,3	5	0,3
passive sans AUX	135	9,9	5	0,3
passive en SE	45	3,3	0	0
TOTAL passives	239	**17,6**	10	**0,6**
relative	48	**3,5**	1	**0,06**
clivage	2	0,1	0	0
$\geq$ 1 det lexicalisé	13	1	346	19,5
négation lexicalisée	1	0,1	20	1,1
TOTAL	1357	100	1777	100

TABLE 9.3 – Quelques aspects de la variabilité et de figement des $\text{LVC}_{1.0}$ vs. $\text{ID}_{1.0}$ dans `FR-train1.0`

Il ressort de la Table 9.3 que les passives sont bien plus fréquentes que les relatives et le clivage : 5 fois plus fréquentes chez les $\text{LVC}_{1.0}$ et 100 fois plus chez les $\text{ID}_{1.0}$. Par ailleurs, les passives représentent une part non négligeable des occurrences de $\text{LVC}_{1.0}$ (17,6%) contre seulement 0,6% des $\text{ID}_{1.0}$. Les discontinuités reflètent une mise à distance des composants du point de vue linéaire de la phrase. Cette distance linéaire est ainsi égale à 4 éléments dans l'exemple 9.16. On s'attend à ce qu'une distance linéaire faible (resp. élevée) soit majoritairement associée à des EP (resp. des non-EP), ce qui en fait un indicateur potentiellement pertinent. Mais, comme signalé en section 6.2.1.2 (p. 92), des composants

linéairement éloignés peuvent cependant être syntaxiquement proches si l'on se penche sur les relations de dépendances qu'ils entretiennent : dans l'exemple 11.2, le nom et le verbe sont ainsi reliés, le nom *décisions* étant l'objet direct du verbe *prendre*. C'est pourquoi nous introduisons la notion de *distance syntaxique* pour représenter cet éloignement des composants dans l'arbre de dépendances.

9.3 Distance syntaxique

Nous définissions la *distance syntaxique* comme suit : lorsque deux éléments sont connectés dans l'arbre de dépendances, cette distance syntaxique vaut 0 et nous parlons dans ce cas de *dépendance directe*. Outre la distance syntaxique nulle signifiant une connexion directe entre composants, nous prenons également en considération les cas de *dépendance indirecte*, dans lesquels cette distance syntaxique est non nulle. De fait, chaque élément inséré dans la chaîne de dépendances incrémente la distance syntaxique. Lorsque cette distance vaut 1, cela signifie que les éléments sont séparés par un autre élément dans l'arbre de dépendances, ce que l'on observe pour les exemples 9.19-9.20, dont les arbres syntaxiques respectifs sont représentés sur la Figure 9.3. On rencontre cette quasi-connexion dans le cas d'énumérations, de coordinations, de relatives, de passives en SE ou avec auxiliaire ou de déterminants complexes.

(9.19) *Les 12 équipes engagées* ***disputèrent*** *un total de 42* ***matchs****.* (`FR-train1.0`)

(9.20) *Le* ***risque*** *est infinitésimal, mais il ne doit pas être* ***couru****.* (`FR-train1.0`)

(9.21) *Les points de vue sur Coye ayant jadis existé ont malheureusement disparu.* (`FR-train1.0`)

(9.22) [...] *elle a* ***réalisé*** *bon nombre de* ***dessins*** *et d'****habillages*** *graphiques.*(`FR-train1.0`)

Deux cas de figure illustrent en outre cette distance unitaire :

— soit l'un des éléments est le grand-parent de l'autre dans l'arbre de dépendances (ce que nous qualifions de *dépendance en série*) comme dans la Fig. 9.3 (à gauche), où ***disputèrent*** est le parent de *total*, lui-même parent de ***matches***,

— soit ils ont un même parent (*dépendance en parallèle*) comme dans la Fig. 9.3 (à droite) où le parent commun de *risque* et *couru* est *infinitésimal*, cette dernière configuration étant, d'après notre étude en corpus, majoritairement associée à des non-EP. Bien que l'exemple situé à droite résulte d'une erreur de parsing automatique dans nos données, il permet d'illustrer le principe de distance syntaxique unitaire dans la configuration dite parallèle.

La connexion syntaxique entre composants d'une EP revêt une importance certaine pour les distinguer de co-occurrences fortuites qui disposent rarement d'une telle connexion, comme l'illustre la Fig. 9.4 (à droite) associée à l'exemple 9.21. A l'inverse, une dépendance indirecte n'est pas nécessairement synonyme de non-EP, comme dans le cas de l'énumération de la Fig. 9.4 (à gauche, associée à l'exemple 9.22).

On remarque d'ailleurs que la différence entre connexion en série et en parallèle semble jouer un rôle, les secondes n'étant jamais, dans le corpus `FR-train1.1`, associées à des EP lorsque la distance syntaxique vaut 2. Un grand nombre d'EP (77% dans `FR-train1.1`)

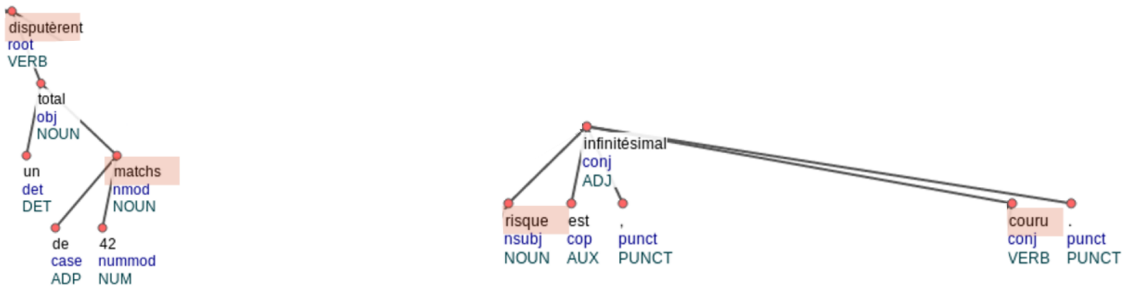


FIGURE 9.3 – Distance syntaxique d'EP avec une valeur de 1 entre éléments surlignés : en série (à gauche) ou en parallèle (à droite).

FIGURE 9.4 – Distance syntaxique avec une valeur de 2 entre éléments surlignés : en série (à gauche) pour une EP ou en parallèle (à droite) pour une co-occurrence fortuite.

est constitué de deux composants[7], essentiellement un verbe et un nom (***prendre***$_{\text{VERB}}$ ***décision***$_{\text{NOUN}}$) pour les $\text{LVC}_{1.0\text{-}1.1}$ et les $\text{ID}_{1.0}$/$\text{VID}_{1.1}$ – le nom étant majoritairement l'objet du verbe – ou un verbe et un pronom (***se***$_{\text{PRON}}$ ***dérouler***$_{\text{VERB}}$) pour les $\text{IReflV}_{1.0}$/$\text{IRV}_{1.1}$. Sur les 3513 tokens d'EP à deux composants, 98% (3433 occ.) ont une distance syntaxique nulle. De plus, sur les 2578 tokens d'EP de `FR-train1.1` comportant un seul verbe et un seul nom, mais pas nécessairement limités à deux composants (p. ex. ***ne***$_{Adv}$ ***pas***$_{\text{ADV}}$ ***en***$_{\text{PRON}}$ ***croire***$_{\text{VERB}}$ ***ses***$_{\text{DET}}$ ***oreilles***$_{\text{NOUN}}$ ⇒ 'être surpris par des propos'), 97% (2503 occ.) ont une distance syntaxique nulle entre le verbe et le nom. Cette distance nulle s'avère pertinente dans la mesure où une recherche en corpus de la co-occurrence de composants d'EP incluant au maximum 2 insertions (autrement dit une distance linéaire faible) offre une précision notablement réduite par rapport à l'utilisation du seul critère de distance syntaxique nulle : $P = 46$ (4103 EP et 4814 non-EP) vs. $P = 70$ (4212 EP et 1811 non-EP) dans le corpus `FR-train1.0`.

Cette étude en corpus permet d'identifier plusieurs indicateurs *a priori* utiles pour modéliser la variabilité des EP :

— La variabilité morphologique, essentiellement au niveau du nom pour les EP de patron VERB-DET-NOUN,
— La nature des discontinuités, ainsi que la catégorie des EP qui leur est souvent corrélée,
— La distance linéaire, évaluée d'après la longueur des discontinuités,
— La distance syntaxique, appréciée d'après le nombre d'éléments interposés dans l'arbre de dépendances syntaxiques.

Ces informations constituent un premier aperçu (non exhaustif) de traits utiles. Nous y ajouterons par exemple la nature des relations de dépendances syntaxiques (section 11.1.1), qui permettent notamment de considérer des modifications en dehors de la fenêtre d'annotation (donc non couvertes par les discontinuités), comme l'adjectif dans : [***prendre*** *une importante*$_{\text{AMOD}}$ ***décision***] vs. [***prendre*** *une* ***décision***] *importante*$_{\text{AMOD}}$.

7. Ce qui permet un calcul simple de la distance syntaxique car nous n'avons pas défini de méthode de calcul pour un nombre d'éléments supérieur à 2.

Chapitre 10

Modélisation de la variabilité

Nous proposons de modéliser le profil de variabilité comme un ensemble de traits morpho-syntaxiques (section 10.1). La validité de cette approche est discutée dans la section 10.2. Le profil de variabilité peut être représenté sous forme graphique (section 10.3) ou bien retranscrit sous la forme d'une combinaison de traits (section 10.4).

10.1 Profil de variabilité multidimensionnel

Parmi les informations fournies par les corpus `train` annotés en EP, les dépendances syntaxiques sont particulièrement intéressantes car elles permettent de savoir à quel point deux éléments sont syntaxiquement reliés. Pour les EP de patron VERB-NOUN, on s'attend ainsi à avoir fréquemment le nom relié au verbe en tant qu'objet direct. Tout écart à ce niveau rend moins probable le fait qu'il s'agisse réellement d'une EP, mais là encore les EP ne sont pas égales, certaines autorisant une distance syntaxique non nulle (***prendre** une dizaine de **décisions***), d'autres non (*jeter une dizaine d'éponges*). La modélisation de la variabilité des EP peut donc bénéficier des liens de dépendances syntaxiques ainsi que de la nature de ces liens entre les éléments en tenant compte de l'étiquette associée à la relation de dépendances. Comme ces relations peuvent être matérialisées par des flèches, leur direction permet de distinguer les *dépendances sortantes* des *dépendances entrantes*, comme dans la Figure 10.1.

Une distance syntaxique nulle est cependant un indice faible pour distinguer des EP (Ex. 10.5) de lectures littérales (Ex. 10.7). Dans ce cas, il convient de se pencher également sur l'existence de modifieurs du nom : l'adjectif *verte* dans (Ex. 10.7) suggère une lecture littérale. De là découle le fait de considérer un ensemble de traits pour distinguer les EP

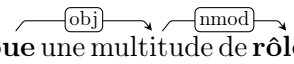


FIGURE 10.1 – Exemple de dépendance entrante (*jouer*) et sortante (*rôles*) à partir du nom *multitude*.

Annotation	Lecture	Traits	
		Lien syntaxique	Modifieur du Nom
EP	Idiomatique	+ (Ex. 10.5)	-(Ex. 10.5)
non-EP	Littérale	+ (Ex. 10.7, Ex. 10.6) - (Ex. 10.9, Ex. 10.8)	+ (Ex. 10.7, Ex. 10.9) - (Ex. 10.6, Ex. 10.8)
	Co-occurrence fortuite	+ (Ex. 10.4, Ex. 10.3) - (Ex. 10.2, Ex. 10.1)	+ (Ex. 10.2, Ex. 10.4) - (Ex. 10.1, Ex. 10.3)

TABLE 10.1 – Profil (sommaire) de variabilité de l'EP ***jeter l'éponge*** vu comme une combinaison de traits.

des non-EP (lectures littérales et co-occurrences fortuites). A titre d'exemple, d'après les deux seules propriétés (ou *traits*) qui seraient la présence d'une connexion syntaxique et d'un modifieur du nom sous la forme d'une dépendance sortante, les profils de l'EP ***jeter l'éponge*** et des non-EP associées pourraient être décrits de façon sommaire comme dans la Table 10.1 établie d'après les exemples 10.1 à 10.9. On y remarque qu'un trait peut être défini en termes de nécessité (lorsqu'un trait est exclusivement présent (+) ou absent (-)) ou de possibilité (lorsqu'un trait peut être présent ou absent, ce qui est signalé par +-). Il ressort de cette table qu'une lecture idiomatique n'est possible que lorsque le nom est (i) syntaxiquement relié au verbe et (ii) dépourvu de modifieur.

(10.1) *Il jette la serpillière à côté de l'éponge.* (lien-,modif-)

(10.2) *Il jette la serpillière à côté de l'éponge verte.* (lien-,modif+)

(10.3) *Quand on l'étreint, l'éponge jette des gouttes sur le sol.* (lien+,modif-)

(10.4) *Quand on l'étreint, l'éponge verte jette des gouttes sur le sol.* (lien+,modif+)

(10.5) *Après 1000 euros de réparations, tu* ***jetteras l'éponge****.* (lien+,modif-)

(10.6) *Il jette l'éponge dans l'évier.* (lien+,modif-)

(10.7) *Il jette l'éponge verte dans l'évier.* (lien+,modif+)

(10.8) *Il jette une dizaine d'éponges dans l'évier.* (lien-,modif-)

(10.9) *Il jette une dizaine d'éponges vertes dans l'évier.* (lien-,modif+)

On ne peut malheureusement pas établir de profil universel pour l'ensemble des EP. En effet, contrairement à l'EP ***jeter éponge***, l'EP ***prendre habit*** ⇒ 'entrer dans les ordres' tolère à la fois la modification du nom par un adjectif (Ex. 10.10) et son absence (Ex. 10.11).

(10.10) *Aniello Francesco Saverio Maresca* ***prend*** *l'****habit*** *franciscain*$_{\text{ADJ}}$ *en 1844 sous le nom de 'Simpliciano della Natività'* (Web)

(10.11) *Tout change avec Dimitri Donskoï qui ne* ***prend*** *pas l'****habit*** *en 1389.* (Web)

Dresser un profil plus précis requiert la prise en compte d'un nombre de traits supérieur aux deux évoqués ici, en s'appuyant par exemple sur les indicateurs potentiellement pertinents mis en évidence dans le chapitre 9, complétés par d'autres traits tels que ceux de Fazly *et al.* (2009), ce qui sera abordé dans la section 10.3.2.

10.2 Limites de la définition de profil

Notre objectif est de tirer profit des informations disponibles dans des corpus annotés en EP pour établir le profil de ces EP grâce à des méthodes d'apprentissage automatique. Or, cette approche est limitée par la couverture en EP des corpus, la majorité des EP n'y apparaissant pas ou peu (comportement zipfien). De plus, pour établir le profil de variabilité d'une EP, on se heurte à trois difficultés majeures : la qualité du corpus source (section 10.2.1), la qualité d'annotation en EP de ce corpus (section 10.2.2) – qui conditionne en partie [1] la validité de la *LemmNorm* (section 10.2.3) – et, enfin, notre capacité à en extraire un profil valide (section 10.2.4).

10.2.1 Qualité du corpus source

La qualité du corpus source s'apprécie selon deux critères : la qualité d'annotation morphosyntaxique et en dépendances syntaxiques d'une part et sa représentativité d'autre part. En effet, le corpus peut comporter des erreurs, comme des lemmatisations incorrectes relevées pour certaines $\text{IRV}_{1.1}$ de `FR-train1.1` : ***assurez-vous*** → ***assurer vous*** au lieu de ***assurer se***, et ***nous amuser*** → ***le amuser*** au lieu de ***se amuser***. De plus, un corpus étant considéré comme un échantillon représentatif de textes, tout défaut de représentativité s'avère préjudiciable, comme la sur-représentation textes médicaux dans le corpus `FR-test1.0` par rapport à celui d'entraînement `FR-train1.0` (Candito *et al.*, 2017), ce qui implique la présence inhabituellement fréquente d'EP de nature médicale (p. ex. ***subir*** *une* ***ICP***, ***présenter*** *une* ***endocardite***). Si, comme l'affirme Rastier (2005), "[a]ucun corpus ne représente la langue" et ne peut donc se targuer d'être *représentatif*, la pertinence des données textuelles repose à notre avis sur l'absence de biais manifeste (registre de langue par exemple) et le choix d'un échantillon de taille suffisante.

10.2.2 Qualité d'annotation d'EP

Les corpus `FR-corpus-1.0` et `FR-corpus-1.1` ont été manuellement annotés en EP (c.-à-d. leurs composants ont été mis en évidence) avec leur catégories (LVC, etc.) par plusieurs annotateurs (Sec. 8.1.1). L'accord inter-annotateurs, mesuré au moyen du coefficient *Kappa* de Cohen κ, permet de juger la qualité d'annotation en EP et, de façon indirecte, la qualité du guide d'annotation. Un faible accord remettrait en question la fiabilité des EP annotées, puisque cela signifierait des choix d'annotation différents d'un annotateur à l'autre (choix des composants de l'EP et/ou de sa catégorie). Cette évaluation a été menée sur un échantillon de 803 phrases dans le corpus `FR-train1.1`. Avec des valeurs $\kappa_{\text{composants}} = 0{,}729$ et $\kappa_{\text{catégorie}} = 0{,}960$ (Ramisch *et al.*, 2018), l'accord inter-annotateurs est jugé satisfaisant et confirme la pertinence de l'annotation manuelle.

10.2.3 Validité de la *LemmNorm* pour la fusion de types d'EP

Notre postulat de départ est que toute EP se voit caractérisée par une *LemmNorm* qui lui est propre. Or, une même *LemmNorm* peut correspondre à des EP distinctes :

1. La validité de la *LemmNorm* dépend essentiellement de la lemmatisation correcte des composants.

⟨le ;prendre ;tête⟩ pour ***prendre la tête*** peut signifier 'se retrouver en première position d'une compétition non encore terminée' ou 'faire perdre patience' [2]. Il s'agit donc de deux types d'EP dont témoigne d'ailleurs la nature différente de leurs arguments : le premier sens implique un complément non-humain (*il* ***prend la tête*** *de la course*$_{\text{OBL}}$) contrairement au second (*ça me*$_{\text{OBJ}}$ ***prend la tête***). A l'inverse, certaines EP se voient attribuer deux *LemmNorm* en raison de l'élision de composants, comme pour ***s'agissant*** et ***il s'agit*** respectivement associés à ⟨agir ;se⟩ et ⟨agir ;il ;se⟩. De même, pour ***(il) y a*** apparaissant sous la forme standard ⟨avoir ;il ;y⟩ et sous la forme familière ⟨avoir ;y⟩. De tels cas demeurent cependant marginaux et ne remettent pas en cause la fusion de variantes par la *LemmNorm*.

10.2.4 Représentativité limitée du profil de variabilité

La limite majeure de ces corpus établis manuellement est leur taille réduite qui risque de fournir un profil de variabilité incomplet. En effet, le fait de ne pas observer un mode de variabilité en corpus ne signifie pas nécessairement qu'il est interdit. Pour résoudre cette difficulté, et augmenter les données exploitables, nous avons également utilisé les corpus `CoNLL17` et `WebSample`. Ces corpus complémentaires permettront de vérifier deux hypothèses en lien avec une dégradation ou un gain de performances pour la tâche d'identification de variantes (voir partie IV) :

H3a Utilisation d'autres types de corpus : en comparant les performances d'identification de variantes d'EP dans un corpus de `test` de nature similaire à celui de l'entraînement (à l'origine du modèle de classification) et dans un corpus différent, on peut s'attendre à ce que les performances se dégradent pour une utilisation sur un nouveau corpus.

H4 Enrichir un corpus manuellement annoté en EP de taille restreinte par un autre corpus annoté automatiquement en EP de la façon la plus fiable possible devrait permettre d'enrichir le profil de variabilité. Nous nous attendons donc à un gain de performances par rapport à la seule utilisation du corpus d'entraînement initial de taille réduite.

10.3 Représentation graphique du profil de variabilité

La modélisation d'un profil de variabilité multidimensionnel des EP nécessite de définir quelles dimensions seront prises en compte (section 10.3.1). Une fois ce choix opéré, il convient de déterminer comment cette variabilité sera appréciée. Pour cela, deux EP servent d'illustration pour le calcul de variabilité, ce qui donne lieu à une représentation graphique de leurs profils de variabilité respectifs (section 10.3.2).

10.3.1 Choix des traits

Pour définir un profil de variabilité, nous nous appuyons sur les cinq caractéristiques suivantes, les trois dernières s'inspirant des travaux de Fazly *et al.* (2009).

2. `https://fr.wiktionary.org`

1. Distance syntaxique : ce trait s'appuie sur le fait que la majorité des EP du corpus français comporte deux composants disposant d'une connexion syntaxique, donc d'une distance syntaxique nulle. Il s'agit essentiellement d'une relation d'objet direct entre le verbe et le nom pour les EP de catégories $LVC_{1.1}$/$VID_{1.1}$), et d'objet entre le pronom et le verbe pour les $IRV_{1.1}$,
2. Distance linéaire : ce trait représente la longueur de la discontinuité entre les composants, autrement dit le nombre d'éléments de la fenêtre d'annotation (excluant les composants [3]), en ignorant l'arbre syntaxique. Ce trait vise notamment à écarter les co-occurrences fortuites, dont on suppose les composants linéairement plus éloignés que dans le cas de lectures idiomatiques ou littérales,
3. Modification du nom : pour les EP comportant un nom, nous nous intéressons à l'existence de modifieurs car, comme dans l'exemple *jeter l'éponge verte*, l'ajout d'un modifieur peut interdire la lecture idiomatique. A l'inverse, certaines EP nécessitent un modifieur non lexicalisé : ***filer le*** *(grand/parfait/...)* ***amour***,
4. Flexion du nom : certaines EP bloquent en effet la flexion nominale en nombre : *jeter les éponges*,
5. Variabilité du déterminant : ce trait a pour objectif de mesurer la tolérance de variabilité du déterminant, fortement dépendante des types d'EP : *jeter cette/mon/... éponge* vs. ***prendre*** *cette/ma/...* ***décision***.

Ces traits peuvent désormais être mis à profit pour dresser le profil de variabilité graphique d'EP de patron identique (VERB-NOUN), mais de catégorie différente ($LVC_{1.1}$, $VID_{1.1}$) afin de déterminer si la catégorie a une influence sur le profil de variabilité.

10.3.2 Représentation graphique

Nous cherchons ici à représenter graphiquement le profil de variabilité de la co-occurrence de lemmes d'une EP au sein d'une phrase, qu'il s'agisse de lectures idiomatique / littérale ou de co-occurrences fortuites, en nous appuyant sur les traits définis dans la section précédente. On considère pour cela deux EP, ***prendre habit*** et ***prendre décision***, dont le patron syntaxique est similaire, car composé du verbe *prendre* et d'un nom (*habit, décision*). Cependant, la première est de catégorie $VID_{1.1}$ tandis que la seconde est une $LVC_{1.1}$. Le corpus `CoNLL17` comporte 1121 co-occurrences des lemmes *prendre* et *habit* dans une même phrase. En ce qui concerne l'EP ***prendre décision***, qui est bien plus fréquente, nous nous limitons à un échantillon aléatoirement choisi de 2179 occurrences. Après cette phase de sélection automatique, une analyse manuelle des occurrences met en évidence les répartitions suivantes :

— Pour *prendre + habit* : 396 lectures idiomatiques, 134 lectures littérales et 591 co-occurrences fortuites.
— Pour *prendre + décision* : 1955 lectures idiomatiques et 224 co-occurrences fortuites. Aucun exemple de lecture littérale ne figure dans l'échantillon traité (p. ex. *j'ai pris la décision qui était sur le bureau*).

Pour définir un profil de variabilité, chacun des cinq traits illustré par les exemples 10.12-10.22 est assorti de sa valeur calculée comme suit :

3. Pour rappel, nous distinguons les composants (lexicalisés) de l'EP des éléments (non lexicalisés).

1. NOConnexion VERBE-NOM : pourcentage d'exemples dont le verbe et le nom ne sont pas syntaxiquement reliés,
2. LongueurDiscont : longueur moyenne des discontinuités,
3. ModifNom : pourcentage d'exemples dont le nom est modifié par un adjectif,
4. PlurNom : pourcentage d'exemples avec une flexion plurielle du nom,
5. VarDet : pourcentage de variabilité du déterminant. L'étalonnage choisi repose sur 6 variations : absence de déterminant, déterminant défini, indéfini, démonstratif, possessif, numéral. Comme une EP satisfait au minimum l'un de ces cas, nous cherchons à établir le nombre d'autres cas (entre 1 et 5) attestés au moins une fois en corpus. De cette façon, si une EP apparaît uniquement avec un déterminant défini et indéfini, l'indicateur VarDet vaudra 1/5, soit 20%.

(10.12) *Il finit ses jours au monastère de Blessac, après avoir **pris** l'**habit**.* (référence)

(10.13) *En 1646, il **prend** l'**habit** franciscain.* (ModifNom)

(10.14) *Thurstan démissionne et **prend** les **habits** de moine de l'abbaye de Cluny.* (PlurNom)

(10.15) *Ils prennent l' habit noir.* (ModifNom + VarDet : Déterminant défini)

(10.16) *J'ai pris les habits et nous sommes partis.* (PlurNom)

(10.17) *Prends tes habits et va à la salle de bain.* (VarDet (Possessif) + PlurNom)

(10.18) *Prenez ces deux habits et utilisez les pour mon linceul* (VarDet (Numéral, démonstratif) + PlurNom)

(10.19) *Ils lui prirent Chapeau, casaque, habit, bourse, et cheval ;* (VarDet (aucun déterminant))

(10.20) *Elle prend un habit, s'amuse avec, le met quelque part.* (VarDet (Déterminant indéfini))

(10.21) *Prendre une douche et laver leurs habits.* (VarDet (Possessif) + PlurNom)

(10.22) *Elle peut prendre une pose de mannequin avec des habits tendances* (ModifNom)

Les profils des EP ***prendre habit*** et ***prendre décision***, ainsi que des co-occurrences fortuites et lectures littérales associées, sont représentés sur les Figures 10.2 et 10.3. L'intersection entre EP et non-EP y est bien plus restreinte pour les $VID_{1.1}$ que pour les $LVC_{1.1}$, ce qui témoigne de leur profil de variabilité plus réduit. Pour l'EP de catégorie $VID_{1.1}$, les traits les plus discriminants sont la distance syntaxique entre le verbe et le nom, la longueur moyenne des discontinuités (en moyenne 1,1 pour les EP et 10,5 pour les non-EP), la possibilité de faire varier le déterminant et la capacité de flexion nominale. Pour celle de catégorie $LVC_{1.1}$, seuls deux traits se distinguent, ce qui traduit une variabilité supérieure aux $VID_{1.1}$: une distance syntaxique nulle entre le verbe et le nom ainsi que la longueur moyenne des discontinuités (en moyenne 2,3 pour les EP et 12,1 pour les non-EP). Ces différences de profil de variabilité entre EP de catégories $LVC_{1.1}$ et $VID_{1.1}$ laissent supposer qu'une classification spécifique par catégorie d'EP devrait être envisagée.

10.4 Correspondance transformation-trait(s)

Les traits définis par Tutin (2016) pour le français peuvent être décrits sous la forme de combinaisons de nos traits : la passivation requiert en premier lieu que le nom précède

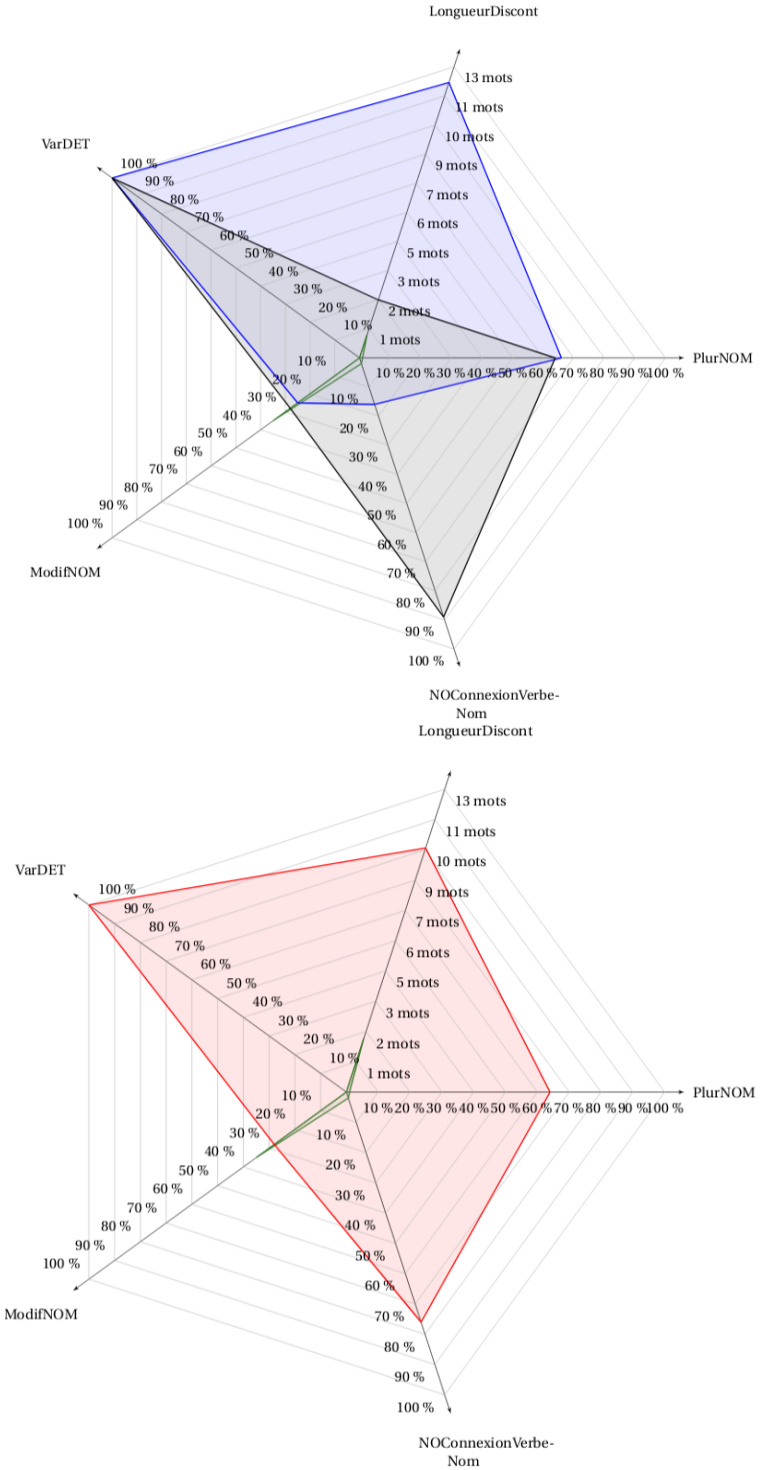


FIGURE 10.2 – Profil de variabilité de l'EP ***prendre l'habit*** (en vert) par rapport aux lectures littérales (en bleu) et co-occurrences fortuites (en noir) dissociées ou considérées conjointement comme des non-EP (en rouge). Le profil de couleur verte est d'autant plus réduit que l'EP est figée.

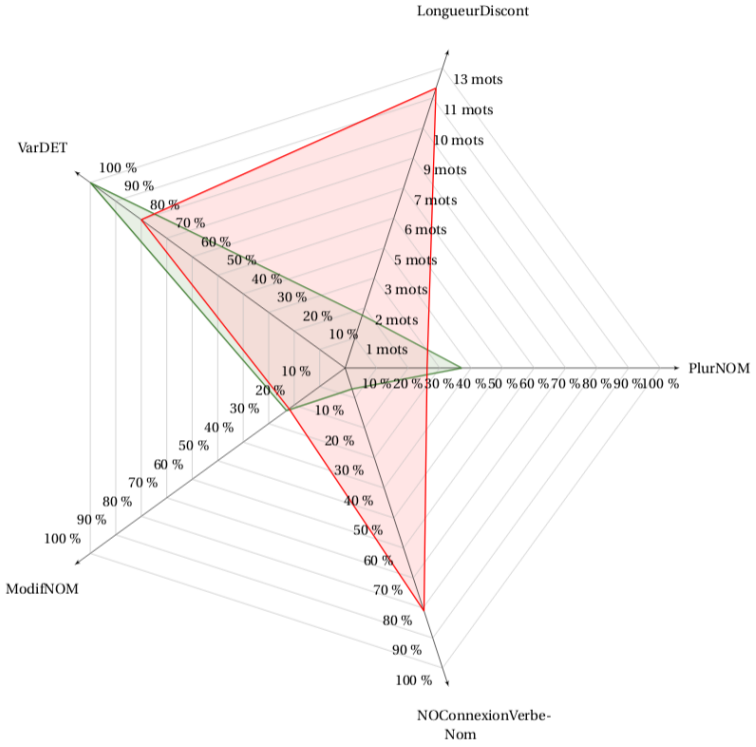


FIGURE 10.3 – Profil de variabilité de l'EP ***prendre décision*** (en vert) par rapport aux co-occurrences fortuites (en rouge).

le verbe[4] : {ordre = NOUN-VERB}. On distingue ensuite trois cas selon le type de passive, prenant en compte à la fois les étiquettes de relations syntaxiques et la présence de certaines insertions.

— Pour les passives avec auxiliaire : $\{\text{VERB}_{\text{VerbForm=Part ;Voice=Pass}} + \textit{être}_{\text{AUX(:PASS))}} \in$ insertions$\}$,
— Pour les passives sans auxiliaire : $\{\text{VERB}_{\text{VerbForm=Part ;Voice=Pass}} + \textit{être}_{\text{AUX}} \notin$ insertions$\}$,
— Pour les passives en SE : $\{\text{VERB}_{\neq\text{(VerbForm=Part ;Voice=Pass)}} + \text{PRON}_{\text{OBJ}} = \text{'}se\text{'} \in$ insertions$\}$.

Ces traits ne sont pas exhaustifs, on peut également ajouter les prépositions *de* et *par* introduisant un complément d'agent dépendant syntaxiquement du verbe avec une relation OBL. Elles peuvent en effet aider à différencier des passives de configurations similaires qui n'en sont cependant pas : *l'**étude** est **menée** par* X_{OBL} vs. **le symptôme est survenu par* X_{OBL}.

L'inconvénient d'une telle description est que, malgré sa généricité, elle doit être adaptée selon les différentes langues : en espagnol par exemple, des constructions passives peuvent faire appel à différents verbes auxiliaires, comme *quedar* (*rester*) dans *el trabajo quedó hecho* (*le travail resta fait*). En revanche, si dans cette langue, les passives en SE existent bien, ce n'est pas le cas en anglais, d'où l'intérêt d'extraire de façon automatique et différenciée par langue les informations relatives aux modalités du passif et des transformations autorisées.

4. Les exceptions sont rares, par exemple : *une fois* $\textbf{\textit{prise}}_{\text{VERB}}$, *la* $\textbf{\textit{décision}}_{\text{NOUN}}$ *est irrévocable.*

Chapitre 11

Quantification de la variabilité des EP

La quantification de la variabilité des EP peut s'effectuer de deux façons selon que l'on privilégie une approche discrète ou continue du phénomène : soit en regroupant les EP par niveaux de variabilité (section 11.1), soit en leur attribuant un score sous la forme d'une valeur numérique de variabilité (section 11.2). Un bilan est proposé en section 11.3.

11.1 Niveaux de variabilité d'EP VERB-(DET)-NOUN

On propose d'attribuer automatiquement des niveaux de variabilité aux EP de patron VERB – (DET) – NOUN d'après la diversité des dépendances syntaxiques associées au nom. Les résultats sont ensuite comparés avec les niveaux établis manuellement par Tutin (2016).

11.1.1 Variabilité des EP d'après leurs dépendances

L'étude des discontinuités sous-entend une non-prise en compte des éléments situés en dehors de la fenêtre d'annotation, comme l'adjectif souligné dans l'exemple 11.1 :

(11.1) *les Byzantins* [***remportent*** *une* ***victoire***] *décisive*$_{\text{ADJ}}$. (`FR-train1.0`)

Pour bénéficier d'une vue plus large des contraintes exercées sur les 29 EP de patron VERB – NOUN les plus fréquentes n'ayant pas de déterminant figé, nous avons donc utilisé les relations de dépendance fournies dans le corpus. Dans cet exemple, l'adjectif post-posé au nom bénéficie en effet d'un lien syntaxique vis-à-vis de ce nom de même nature (AMOD pour *modifieur adjectival*) que s'il lui était antéposé et, dans ce cas, également compté comme une discontinuité.

L'exploitation des relations syntaxiques permet également une analyse plus aisée des longues discontinuités (soulignées dans l'Ex. 11.2).

(11.2) *Il* ***prend*** *également, et cette fois-ci seul et sans vote de l'assemblée, des* ***décisions*** *réglementaires.* (`FR-train1.1`)

Pour chaque EP testée, chaque sorte de dépendance sortante du nom est comptabilisée. En effet, nous nous intéressons à la capacité du nom à être modifié plutôt qu'au nombre de

modifieurs[1]. On obtient 374 patrons de dépendances pour les noms inclus dans des $ID_{1.0}$ et 434 dans des $LVC_{1.0}$, valeurs suffisamment proches pour permettre une comparaison. L'absence de dépendances est trois fois et demie plus fréquente dans les $ID_{1.0}$ que dans les $LVC_{1.0}$ (48% vs 13,6%) et, dans une proportion similaire, le déterminant non lexicalisé est plus fréquent dans les $LVC_{1.0}$ (30,4% vs 7,5%). Alors que le nom garde tout son sens dans les $LVC_{1.0}$, cela se produit rarement dans les $ID_{1.0}$ (***faire office***) sans pour autant être impossible (***faire prisonnier***). Cela peut expliquer cette absence plus fréquente de dépendances du nom dans les $ID_{1.0}$ (***faire*** **un grand* ***office***) que dans les $LVC_{1.0}$ (***faire*** *une grande* ***révélation***).

En s'appuyant sur l'exploitation des dépendances, nous pouvons établir un classement des EP de patron VERB − (DET) − NOUN selon leur degré de variabilité et le comparer à l'état de l'art afin d'estimer la pertinence de cet indicateur.

11.1.2 Comparaison avec Tutin (2016)

Tutin (2016) a étudié la variabilité des 30 EP (idiomes et constructions à verbe support) de patron VERB − (DET) − NOUN les plus fréquentes en français. Les propriétés les plus discriminantes ainsi identifiées sont la pluralisation du nom, les constructions relatives et passives (*le* ***rôle*** *est* ***joué*** *;* ***l'attention prêtée***). 2 EP ont un comportement inattendu dans le classement de Tutin (2016) : la $LVC_{1.0}$ ***avoir recours*** s'avère moins variable que l'$ID_{1.0}$ ***avoir*** *du* ***mal***[2], alors que l'on s'attend à ce que les $ID_{1.0}$ se situent toujours sur les niveaux de variabilité les plus faibles et les $LVC_{1.0}$ sur les plus élevés.

Nous proposons d'établir un classement des EP de patron VERB-(DET)-NOUN reposant sur un unique critère : l'analyse automatique de la diversité des dépendances syntaxiques associées au nom de l'EP, selon le principe présenté en section 11.1.1. On suppose en effet que plus une EP est variable, plus la nature de ses dépendances est susceptible d'être, elle-aussi, variée, et inversement. La partie haute de la Table 11.1 montre la comparaison entre le niveau de variabilité défini par Tutin (2016) – qui constitue notre référence – et la variabilité déduite des dépendances syntaxiques attachées au nom dans `FR-train1.0`. On se focalise pour cela sur six EP de niveaux 0, 1 et 4 (deux par niveau), sélectionnées en raison d'une fréquence suffisante dans notre corpus. Pour ces six EP, les résultats sont cohérents pour les degrés extrêmes de (non-)variabilité (0 et 4) mais moins marqués pour le niveau 1. En effet, dans notre corpus, l'expression ***faire partie*** (Niveau 1) obtient un score assez similaire à celui de ***faire appel*** (Niveau 0) (0,03 vs. 0,09 dépendances différentes). Les cellules grisées, représentant les invariabilités, sont essentiellement affectées aux niveaux 0 et 1, mais jamais au niveau 4. De même, l'absence de discontinuités est majoritaire sur ces faibles niveaux (plus de 90%), mais minoritaire pour le niveau 4 (moins de 15%).

Le nombre de dépendances différentes associées au nom de l'EP semble un indicateur intéressant : il corrobore effectivement la tendance de variabilité supérieure des EP du niveau 4 (plus de 0,39 dépendances différentes en moyenne) par rapport à celles de niveau

1. Par exemple, la phrase *Les femmes*$_{\text{NSUBJ}}$ *jouent au football*$_{\text{NMOD}}$ *depuis la fin*$_{\text{NMOD}}$ *du XIX*^{e} *siècle en Angleterre*$_{\text{NMOD}}$ aura pour patron de dépendances associées au verbe *jouer* : {NSUBJ,NMOD} et non {NSUBJ,NMOD,NMOD,NMOD}.

2. Dans `FR-train1.0` : ***avoir*** *(du)* ***mal***, ***avoir*** *(également) beaucoup de* ***mal***, *n'****avoir*** *aucun* ***mal***, ***avoir*** *plus* ***mal***.

0 (moins de 0,18).

La partie basse de la Table 11.1 met également en évidence d'autres tendances : si l'on s'intéresse aux traits pris en compte dans le classement de Tutin (2016), on s'aperçoit que la variabilité en nombre du nom et l'emploi de relatives ne sont observés dans notre corpus qu'au niveau 4. La passivation n'est quant à elle jamais observée au niveau 0. On peut par ailleurs ajouter deux autres critères, non pris en compte par Tutin (2016). Par exemple, une absence de discontinuité prédomine pour les niveaux 0 et 1 (au moins 92% des cas), mais pas pour le niveau 4 (au maximum 14% des cas). De plus, l'ordre VERB-NOUN est systématiquement observé pour trois des EP de niveaux 0 et 1, ce qui n'est pas le cas au niveau 4. Cette constatation est cohérente avec la fréquence de tournures passives ou relatives qui engendrent cet ordre de composants.

	EP (type)	***jouer rôle***	***poser question***	***faire partie***	***tenir compte***	***faire appel***	***donner lieu***
	EP (tokens)	33 occ.	14 occ.	90 occ.	17 occ.	32 occ.	12 occ.
	Catégorie	$LVC_{1.0}$	$LVC_{1.0}$	$ID_{1.0}$	$ID_{1.0}$	$LVC_{1.0}$	$ID_{1.0}$
Niveau de variabilité de Tutin (2016)		Niveau 4		Niveau 1		Niveau 0	
Propriétés de Tutin (2016)	Nom variable en nombre	oui	oui	non	non	non	non
	Passivation	oui	oui	non	oui	non	non
	Relativation	oui	oui	non	non	non	non
Notre mesure	Nombre moyen de dépendances sortantes différentes du nom (y compris absence de dépendance)	0,39	0,57	0,03	0,29	0,09	0,17
Autres propriétés	Ordre VERB-NOUN	79%	50%	100%	71%	100%	100%
	Ordre NOUN-VERB	21%	50%	0%	29%	0%	0%
	Discontinuités :						
	longueur = 0	9%	14%	93%	94%	94%	92%
	longueur = 1	76%	36%	7%	6%	6%	8%
	longueur = 2	9%	36%	0%	0%	0%	0%
	longueur > 2	6%	14%	0%	0%	0%	0%

TABLE 11.1 – Comparaison des degrés de variabilité observés en corpus avec ceux établis par Tutin (2016). Les cellules grisées indiquent des invariabilités. Le niveau 0 correspond à la variabilité minimale et le niveau 4 à la variabilité maximale.

11.2 Score de variabilité d'EP VERB-(DET)-NOUN

A la différence des niveaux de variabilité permettant de classer les EP selon des paliers de variabilité chez Tutin (2016), nous proposons une approche novatrice conduisant à un classement plus fin en leur attribuant un score de variabilité. On se restreint pour cela à un ensemble d'EP homogène du point de vue de leur patron syntaxique (VERB−(DET)−NOUN) mais hétérogène du point de vue des catégories d'EP concernées ($ID_{1.0}$/$LVC_{1.0}$) afin de valider la méthode. Le choix de ce patron permet de limiter l'influence de la restriction opérée car il s'agit du patron le plus fréquent en corpus, d'où un impact moins marqué sur la couverture des EP considérées. Grâce à ce score de variabilité, on pourrait distinguer

les EP ayant un score élevé des autres EP. Ainsi, savoir qu'un type d'EP donné dispose d'un score très élevé laisserait sous-entendre qu'il tolère un large éventail de réalisations, y compris non observées auparavant. A l'inverse, toute réalisation non observée au préalable pour un type d'EP de score faible se verrait plus facilement considérée comme une 'non-EP'. On s'attend également à ce que des EP de catégorie $ID_{1.0}$ soient moins variables que des $LVC_{1.0}$. On pourrait donc exploiter le score de variabilité pour permettre la catégorisation d'EP, ce qui pourrait ensuite être utile pour des tâches telles que la traduction automatique car les idiomes bénéficient souvent d'une double lecture littérale/idiomatique et nécessitent donc un traitement particulier. Pour calculer le score de variabilité d'une EP, nous nous appuyons sur la similarité entre les occurrences de cette même EP. De façon similaire avec la section précédente, la validité de cette méthode est évaluée par comparaison avec les niveaux de variabilité mis en évidence par Tutin (2016). L'intérêt de ce score de variabilité est également discuté pour la distinction entre les catégories $LVC_{1.0}$ et $ID_{1.0}$, ainsi que les limites de cette approche.

11.2.1 Mesure de similarité et de variabilité

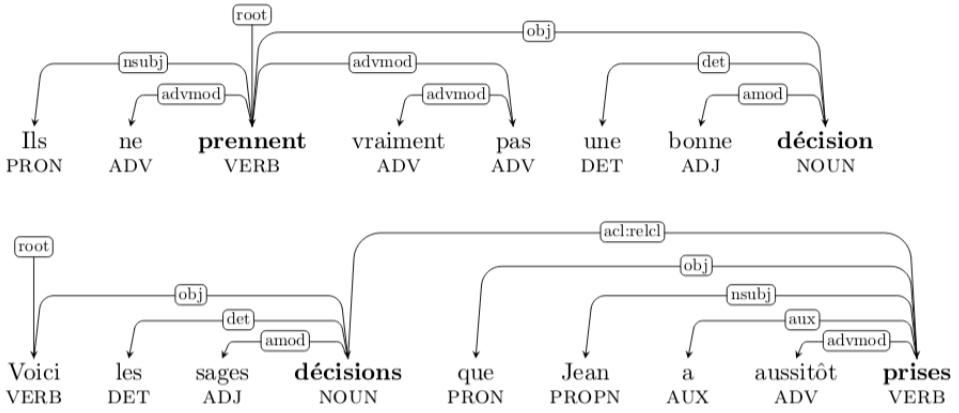


FIGURE 11.1 – Deux exemples de l'EP ***prendre*** *une* ***décision*** étiquetés en POS et avec parsing en dépendances.

Considérons par exemple l'EP ***prendre décision***, nous cherchons à savoir à quel point deux réalisations sont similaires par comparaison de leurs tokens deux à deux, par exemple *prennent* vs. *prises* et *décision* vs. *décisions* (Fig. 11.1). La mesure de variabilité de cette EP s'appuie donc sur une mesure préalable de similarité. La Figure 11.1 illustre deux arbres de dépendances pour des phrases qui sont des variantes l'une de l'autre, désormais nommées V_1 et V_2, de cette EP. V_1 et V_2 possèdent des propriétés syntaxiques et linéaires (c.-à-d. du point de vue de leurs discontinuités) qui se ressemblent sur certains points tout en divergeant sur d'autres. Par exemple, le nom *décision* gouverne un déterminant (DET) et un modifieur adjectival (AMOD) à la fois dans V_1 et dans V_2, mais ce n'est que dans V_2 qu'il gouverne également une relative (ACL :RELCL). Le verbe *prendre* gouverne un sujet nominal (NSUBJ), un complément d'objet (OBJ) et un modifieur adverbial (ADVMOD) à la fois dans V_1 et dans V_2, mais un auxiliaire (AUX) uniquement dans V_2. Quant aux

discontinuités, leurs POS sont ADV (deux fois), DET et ADJ dans V_1, et PRON, PROPN, AUX et ADV dans V_2. Autrement dit, une seule POS (ADV) est commune aux deux variantes.

Afin de mesurer simultanément les caractéristiques communes et distinctes entre deux variantes, nous définissons leur similarité d'après la similarité de leurs composants et des éléments externes insérés. Toutes les variantes d'une EP E sont regroupées au moyen de leur *LemmNorm*. Par exemple, dans la Figure 11.1, les composants $C_1 = \textit{décision}$ et $C_2 = \textit{prendre}$ conduisent à la *LemmNorm* $E = \langle \textit{décision}; \textit{prendre} \rangle$. De cette façon, C_i^j représente la forme qu'un composant i prend dans la variante j :

$C_1^1 = \textit{décision}$ et $C_2^1 = \textit{prennent}$ $\qquad$ $C_1^2 = \textit{décisions}$ et $C_2^2 = \textit{prises}$

La similarité des objets (composants ou EP) est mesurée par le coefficient de Sørensen–Dice, défini comme :

$$S_{\text{Dice}}(O_1, O_2) = \frac{2 \times |P(O_1) \cap P(O_2)|}{|P(O_1)| + |P(O_2)|}$$

où $P(O_i)$ est un ensemble de propriétés pertinentes de l'objet O_i. On définit désormais deux mesures de similarité entre variantes : la similarité syntaxique – se focalisant sur les dépendances sortantes – et la similarité linéaire – reposant sur les discontinuités.

11.2.1.1 Similarité syntaxique

La similarité syntaxique S^S s'appuie sur les dépendances entre les composants d'une EP et les autres éléments de la phrase. Cela permet de prendre en compte des arguments parfois très éloignés et des modifieurs en dehors de la fenêtre d'annotation. La similarité de chaque paire de composants est d'abord calculée (c.-à-d. le verbe de V_1 vs. celui de V_2, le nom de V_1 vs. celui de V_2, etc.) puis moyennée pour l'EP complète d'après le poids que l'on souhaite attribuer à chaque composant par exemple une pondération nulle pour le déterminant, peu informatif car fortement dépendant du nom auquel il est associé. Pour chaque composant, l'ensemble des dépendances sortantes est considéré et les relations de chaque sorte de dépendances ne sont comptabilisées qu'une seule fois. Dans les deux phrases illustrées par la Figure 11.1, la similarité syntaxique du nom C_1 et du verbe C_2 est :

$$\begin{aligned} S^S(C_1^1, C_1^2) &= \frac{2 \times |\{\text{AMOD,DET}\}|}{|\{\text{ACL :RELCL,AMOD,DET}\}| + |\{\text{AMOD,DET}\}|} \\ &= \frac{4}{5} \\ S^S(C_2^1, C_2^2) &= \frac{2 \times |\{\text{ADVMOD,NSUBJ,OBJ}\}|}{|\{\text{ADVMOD,NSUBJ,OBJ}\}| + |\{\text{ADVMOD,AUX,NSUBJ,OBJ}\}|} \\ &= \frac{6}{7} \end{aligned}$$

La similarité syntaxique entre les variantes V_1 et V_2 est la moyenne pondérée du score individuel obtenu pour chaque composant :

$$S^S(V_1, V_2) = \sum_{i=1}^{n} w_i \times S^S(C_i^1, C_i^2)$$

où la somme des poids $w_1, \ldots, w_n$ est de 1. Par exemple, avec des poids uniformes $w_1 = w_2 = \frac{1}{2}$, on obtient :

$$S^S(V_1, V_2) = \frac{1}{2} \times \frac{4}{5} + \frac{1}{2} \times \frac{6}{7} = \frac{29}{35}$$

11.2.1.2 Similarité linéaire

La similarité linéaire S^L est définie entre une variante d'EP et une forme de référence en fonction des POS de leurs discontinuités respectives. La longueur de discontinuités pour une même POS n'est pas considérée. De cette façon, nous nous focalisons sur le fait d'admettre une insertion d'une POS donnée, plutôt que sur leur quantité. Par exemple, les deux insertions adverbiales ADV dans V_1 (*vraiment* et *pas*) n'apportent pas davantage d'information qu'une seule, c'est pourquoi elles ne sont comptées qu'une seule fois :

$$\begin{aligned} S^L(V_1, V_2) &= \frac{2 \times |\{\text{ADV}\}|}{|\{\text{ADJ,ADV,DET}\}| + |\{\text{ADV,AUX,PRON,PROPN}\}|} \\ &= \frac{2}{7} \end{aligned}$$

11.2.1.3 De la similarité à la variabilité des EP

Étant donné deux mesures de similarité S^S et S^L entre V_1 et V_2 pour une EP E, le score de *rigidité* de E est calculé d'après la moyenne des similarités entre toutes les paires de variantes de E. Par exemple, si ***prendre décision*** apparaît 6 fois en corpus, on moyenne les scores S^S et S^L sur $\binom{6}{2} = 15$ paires :

$$R^X(E) = \frac{1}{\binom{m}{2}} \times \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} S^X(V_i(E), V_j(E))$$

où $X \in \{\text{S}, \text{L}\}$, m est le nombre de variantes de E dans le corpus, et $V_i(E)$ est la $i^{\text{ème}}$ variante. La valeur de rigidité de E s'échelonne entre 0 et 1. La *variabilité* de E peut dès lors être définie comme complémentaire de cette rigidité : $V^X(E) = 1 - R^X(E)$.

La pertinence et l'utilité de ces mesures sont évaluées dans la section 11.2.2. Les valeurs des paramètres, choisies de façon empirique [3], sont les suivants :

- Poids w des composants : $w_{\text{VERB}} = 0$, $w_{\text{NOUN}} = 1$, $w_{\text{DET}} = 0$,
- Traits pris en compte pour S^L (c.-à-d. les POS des insertions) : ADJ, ADV, INTJ, NOUN, CCONJ, NUM, PROPN, VERB, AUX, SCONJ, ADP, PRON, X, PART, SYM, DET. Seule PUNCT est donc exclue ici.
- Traits pris en compte pour S^S au format UD : AUX :PASS, NMOD :POSS, NUMMOD, DET, NSUBJ :PASS, ACL :RELCL, AMOD, ACL. Parmi les relations exclues (car supposées être moins porteuses d'informations) figurent par exemple : PUNCT, ADVCL, etc.

11.2.2 Validation de la méthode

Afin d'estimer la pertinence de nos mesures, nous nous référons de nouveau à l'étude en corpus de Tutin (2016) et les 6 niveaux de variabilité obtenus pour les 30 EP les plus

3. A long terme ces paramètres devraient être estimés expérimentalement, peut-être d'après des applications spécifiques.

Niveau de variabilité de Tutin	0	1	2	3	4	5	Total
Nombre d'EP dans `FR-train1.0`	6	3	2	3	1	3	18
Nombre de tokens dans `FR-train1.0`	69	114	8	18	7	54	270
Niveaux regroupés	N_{0-1}		N_{2-4}			N_5	N

TABLE 11.2 – Distribution des EP extraites du corpus `FR-train1.0` parmi les classes de Tutin (2016)

fréquentes en français de patron VERB – (DET) – NOUN. Les modes de variabilité définis par Tutin (2016) sont définis d'après des phénomènes linguistiques complexes comme le fait d'admettre la passivation et des constructions relatives, et qui doivent être validées manuellement. A l'inverse, notre approche est motivée par la recherche de procédures automatisables et multilingues, ce qui implique une appréciation différente de la variabilité des EP. Il est intéressant de savoir à quel point ces deux approches s'accordent au niveau de leurs conclusions.

Dans ce but, nous extrayons de `FR-train1.0` toutes les occurrences des 30 EP répertoriées par Tutin (2016) et nous ne retenons que celles apparaissant au minimum deux fois (car la mesure de similarité requiert au bas mot deux tokens). La Table 11.2 montre la distribution de l'ensemble résultant N des 18 EP d'après les niveaux de Tutin. Tandis que leur fréquence en corpus est relativement élevée aux niveaux 0, 1 et 5, elle est faible aux niveaux 2, 3 et 4. Par conséquent, certains niveaux voisins sont regroupés dans trois sous-ensembles : N_{0-1}, N_{2-4} and N_5. Pour chaque EP de N, nous calculons V^L et V^S avec le poids $w_i = 1$ pour le nom et 0 pour le verbe et le déterminant (s'il y en a un). Comme illustré dans les boxplots de la Figure 11.2 (a–b), V^L tend à augmenter avec le niveau de Tutin. Autrement dit, plus une EP est variable (d'après le jugement d'un linguiste expert sur la base d'une étude en corpus manuelle), plus sa valeur de variabilité linéaire est importante. Les niveaux extrêmes 0–1 et 5 sont particulièrement bien discriminés par V^L [4].

Aucune tendance intéressante n'a toutefois été observée pour la variabilité syntaxique du nom [5] bien que les dépendances sortantes différentes puissent *a priori* jouer différents rôles dans la modélisation de la variabilité syntaxique. Par exemple, dans l'EP ***aller dans** le bon **sens*** ⇒ 'évoluer positivement', la dépendance entre le nom et le modifieur *bon* nous en apprend probablement davantage sur la rigidité de cette EP que sa préposition *dans* ou le déterminant *le*. Dans des travaux futurs, nous aimerions estimer expérimentalement le poids des différentes relations de dépendance dans S^S.

11.2.3 Utilisation de la méthode : discrimination $\text{LVC}_{1.0}$ vs. $\text{ID}_{1.0}$

Les constructions à verbe support (ici représentées par les $\text{LVC}_{1.0}$) sont réputées avoir un fonctionnement morphosyntaxique relativement régulier en comparaison des idiomes (ici

4. Le test de Wilcoxon-Mann-Whitney (WMW) (Wilcoxon, 1946; Mann et Whitney, 1947) confirme que la variabilité moyenne des EP du niveau N_5 diffère de celles des niveaux N_{0-1} avec une significativité $\alpha = 0.05$.

5. Celle du verbe a été ici écartée en raison de son manque de pouvoir discriminatoire : la majorité des EP tolère en effet des modifieurs adverbiaux par exemple.

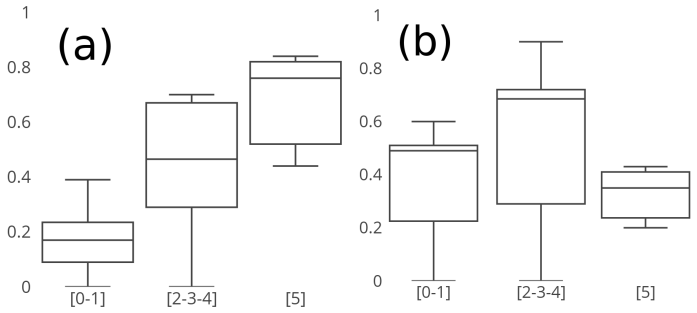


FIGURE 11.2 – Boxplots de Tukey de V^L (a) et V^S (b) (en ordonnée) en fonction des niveaux de Tutin (en abscisse).

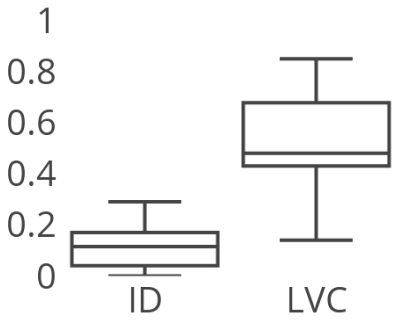


FIGURE 11.3 – Boxplots de Tukey de V^L (en ordonnée) en fonction des catégories d'EP (en abscisse) : $\text{ID}_{1.0}$ et $\text{LVC}_{1.0}$

$\text{ID}_{1.0}$) qui tendent à être davantage rigides. Nous nous attendons donc à ce que nos mesures de variabilité permettent de discriminer ces catégories. Pour cela, nous sélectionnons les EP les plus fréquentes (au moins 10 tokens [6]) dans `FR-train1.0`, ce qui est le cas pour les 12 $\text{ID}_{1.0}$ et 17 $\text{LVC}_{1.0}$ de la Table 9.1. V^S et V^L ont été calculés pour chacune de ces EP et, comme le montre la Figure 11.3, V^L discrimine fortement les $\text{ID}_{1.0}$ des $\text{LVC}_{1.0}$ car la variabilité des $\text{ID}_{1.0}$ ne dépasse jamais 0,30, alors qu'elle atteint 0,94 pour les $\text{LVC}_{1.0}$ [7].

La capacité de cette mesure de variabilité linéaire à discriminer les idiomes des constructions à verbe support pourrait être mise à profit pour développer des stratégies différentes dans la gestion des EP d'après leur catégorie.

11.2.4 Limites et biais

Cette étude de la variabilité se trouve confrontée à différentes limites et/ou biais. Notons par exemple que le coefficient de Sørensen–Dice ne respecte pas l'inégalité triangulaire et ne peut donc être qualifié de métrique de distance *stricto sensu*. Il dispose cependant d'un

6. Ce seuil représente un compromis entre le fait de conserver suffisamment de paires de variantes pouvant être comparées pour modéliser le profil de variabilité d'une EP d'une part, et suffisamment d'EP pour évaluer V^S et V^L d'autre part. Le fait d'augmenter ce seuil à un minimum de 20 tokens conduirait à 190 comparaisons par EP (vs. 45 ici) mais avec seulement 8 EP.

7. Ces résultats sont significativement significatifs à $\alpha = 0.01$ d'après le test de WMW.

atout majeur : en mettant davantage l'accent sur l'intersection des propriétés des deux objets (par rapport au coefficient de Jaccard par exemple), il souligne la similitude entre tokens.

Par ailleurs, notre focalisation sur les expressions de patron VERB − (DET) − NOUN se limite *de facto* aux $LVC_{1.0}$ et $ID_{1.0}$, autrement dit à une portion limitée du corpus (47% des tokens). Néanmoins, les scores de similarité définis ici sont généralisables à des EP d'autres patrons puisqu'il s'agit de comparer les réalisations de chaque composant d'un même type d'EP et de prendre en compte leurs discontinuités. Notre corpus étant par ailleurs de taille restreinte, on établit un profil en fonction d'un nombre de réalisations qui n'est pas obligatoirement pertinent. Que penser en effet d'une EP qui n'apparaîtrait que deux fois en corpus ? Malgré tout, nous supposons que plusieurs EP peuvent partager des similarités comme le fait de tolérer/d'interdire la flexion nominale en nombre, ce qui devrait permettre de compenser les nombreux cas où une EP apparaît rarement.

11.3 Bilan

Cette étude descriptive de la variabilité en corpus a permis de confronter nos hypothèses de travail (formulées en page 121) avec la réalité du comportement des EP (section 11.3.1). Il est également intéressant d'évaluer l'impact de cette variabilité sur les performances obtenues par les systèmes d'identification automatique (section 11.3.2) avant d'aborder le fonctionnement de notre système.

11.3.1 Retour sur les hypothèses

Concernant l'hypothèse H_1, les représentations graphiques de profils de variabilité sous forme de radar (section 10.3.2) illustrent que des EP différentes ont des profils de variabilité qui leur sont propres et que certains marqueurs morpho-syntaxiques sont particulièrement discriminants pour distinguer les EP de non-EP. Bien que ces exemples illustratifs ne puissent avec certitude (in)valider l'hypothèse H_1, ils constituent des indices en faveur de la notion de profil de variabilité. Cette modélisation multidimensionnelle, sous-tendue par H_2, peut donc s'avérer pertinente.

De même, l'hypothèse H_{2a} sur la possibilité de quantifier la variabilité se voit confortée par une tendance à un score de variabilité supérieur des $LVC_{1.0}$ par rapport à celui des $ID_{1.0}$ si l'on se focalise sur les types d'insertions. En revanche, ces catégories ne sont pas complètement disjointes, certains $ID_{1.0}$ ayant un niveau équivalent à celui des $LVC_{1.0}$. On note d'ailleurs une plus grande dispersion de valeurs de variabilité pour les $LVC_{1.0}$ que pour les $ID_{1.0}$, ce qui pourrait laisser supposer que les profils de variabilité des constructions à verbe support sont plus variés que ceux des idiomes, autrement dit que cette dernière catégorie serait plus homogène du point de vue de leur variabilité linéaire.

11.3.2 Variabilité des EP et performance des systèmes

Au terme de cette étude de la variabilité, il serait intéressant de savoir si (i) des EP moins variables sont mieux identifiées par des systèmes d'identification automatique, par

exemple ceux évalués lors de la ST [8], (ii) certaines formes de variabilité sont mieux identifiées. La Figure 11.4 [9] résume les performances de 15 systèmes évalués (pour le français) sur leur capacité à identifier des EP dans un nouveau corpus lors de la ST. On remarque que ces systèmes tendent à rencontrer davantage de difficultés pour l'identification de tokens d'EP apparaissant sous des formes différentes de celles vues durant l'entraînement : sur 15 systèmes, la F-mesure diminue de 0,10 à 0,33 points pour le français entre l'identification des *identical-to-train* et celle des *variant-of-train*.

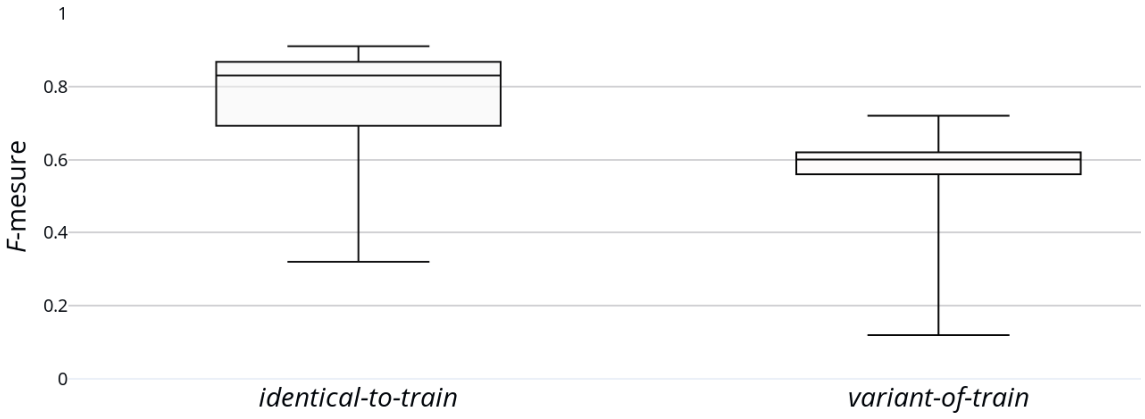


FIGURE 11.4 – F-mesure des 15 systèmes de la compétition PARSEME 1.1 pour l'identification d'EP déjà vues selon que l'on considère des variantes identiques à celles vues (*identical-to-train*) ou différentes (*variant-of-train*).

Il est toutefois difficile de tirer des conclusions précises sur l'aptitude des systèmes à gérer la variabilité, faute de disposer d'un nombre suffisant de variantes et de modes de variabilité par expression dans le corpus de `test`. C'est ce dont on s'aperçoit d'après les performances obtenues par le système TRAVERSAL (qui fut le plus performant durant la ST) pour des EP fréquentes de patron VERB=*prendre*-NOUN, 2 étant des expressions de catégorie $\text{VID}_{1.1}$ (21 occurrences de ***prendre part*** et 5 de ***prendre conscience*** dans `FR-train1.1`) et deux autres de catégorie $\text{LVC}_{1.1}$ (6 occurrences de ***prendre mesure*** et 8 de ***prendre décision***). Ces EP sont présentes dans le corpus `FR-test1.1` à hauteur respective de 2, 1, 2 et 3 occurrences. Concernant les $\text{VID}_{1.1}$, tous sont identifiés hormis une expression erronée au niveau du participe passé (*écoliers ayant pris part à l'expérience*). 3 des 5 occurrences de $\text{LVC}_{1.1}$ sont identifiées mais, alors que le système reconnaît *des* ***mesures*** *décisives*$_{\text{ADJ}}$ *seront*$_{\text{AUX}}$ ***prises*** (vs. *les* ***mesures prises*** dans `FR-train1.1`), il échoue à deux reprises sur une configuration similaire malgré les similitudes entre les POS des mots du contexte : *une telle*$_{\text{ADJ}}$ ***décision*** *est*$_{\text{AUX}}$ ***prise*** (vs. *la* ***décision prise*** dans `FR-train1.1`).

De façon générale, on remarque que ce système a une F-mesure sur les EP déjà vues de 64,71 pour les $\text{IRV}_{1.1}$, de 55,77 pour les $\text{VID}_{1.1}$ et de 42,92 pour les $\text{LVC.full}_{1.1}$, ce qui tend

8. Pour rappel, ce sigle fait référence à la compétition PARSEME de 2018.

9. Graphique établi à partir des résultats des systèmes ayant participé à la ST. Ces données, ainsi que les informations relatives aux différents système soumis, sont disponibles à l'adresse `http://multiword.sourceforge.net`.

à confirmer que les catégories d'EP les plus sujettes à variabilité sont plus difficilement identifiables.

Quatrième partie

Vers une identification multilingue de variantes d'EP *pour faire avancer le schmilblick*

Cette partie est consacrée aux deux systèmes (VarIDE et VarIDE+) que nous avons développés pour la tâche spécifique d'identification de variantes d'EP, à laquelle nous ferons désormais référence sous l'appellation de *Tâche*. Le principe de fonctionnement de VarIDE et VarIDE+, développés pour résoudre la Tâche, s'appuie sur l'hypothèse H_3, qui est l'hypothèse principale de cette partie :

H3 Connaître le profil de variabilité de chaque EP connue peut aider à identifier ses variantes dans un nouveau texte. Par exemple, nous considérerons la séquence ***posera un lapin*** comme étant une variante de l'EP ***posais un lapin*** mais pas *le lapin posé, poser des lapins, etc.* Il serait donc possible de développer un système d'identification d'EP centré sur les variantes grâce à un classifieur entraîné sur ces profils.

D'un point de vue théorique, nous avons en effet observé que les profils de variabilité d'EP et de non-EP – reposant sur des traits linguistiquement pertinents liés à la morphologie, aux relations de dépendances syntaxiques et aux discontinuités – étaient distincts (section 10.3.2). Chercher à décrire les profils d'EP de façon monolithique malgré les divergences observées d'une EP à l'autre ne nous a pas semblé l'approche la plus adaptée, compte tenu de notre objectif de généralisation multilingue et des difficultés d'appréciation des traits pertinents pour chaque langue. C'est pourquoi nous proposons d'effectuer un apprentissage automatique des spécificités des EP vs. non-EP dans le cadre d'une approche supervisée. Cette approche, justifiée par la disponibilité de corpus annotés en EP, suggère en définitive que le modèle de classification pourra s'appuyer de façon implicite sur les profils de variabilité.

Concrètement, nous extrayons tout d'abord un large ensemble de candidats dotés des mêmes co-occurrences de lemmes que des EP attestées (donc connues).

Le chapitre 12 présente le système multilingue VarIDE que nous avons soumis à la compétition PARSEME 1.1 (ST) (Ramisch *et al.*, 2018). Le développement de ce premier système constitue une étude de faisabilité puisque nous ignorions alors la pertinence de notre méthodologie. Le contexte de la compétition permettait en outre de comparer ses performances pour la Tâche par rapport à d'autres approches. L'une des limites majeures de ce premier système est, pour le français, le grand nombre de traits utilisés pour décrire les candidats (18 000), c'est-à-dire environ le double de la taille du corpus d'apprentissage, ce qui expose au problème de surapprentissage.

C'est pourquoi nous proposons une optimisation de VarIDE (chapitre 13) bénéficiant d'une sélection des traits les plus pertinents et d'autre part évalué différentes techniques de classification, au lieu de l'utilisation restreinte à un classifieur Naïve Bayes pour VarIDE. Ce second système, VarIDE+, n'a pour l'instant été évalué que sur le français afin de bénéficier de notre connaissance de cette langue pour (i) comprendre la pertinence des traits sélectionnés, (ii) mener à bien l'analyse d'erreurs.

Chapitre 12

VarIDE : *un ballon d'essai*

Dans ce chapitre, nous présentons notre système VarIDE dédié à la Tâche, à la fois du point de vue de son fonctionnement (extraction de candidats suivie de leur classification) et des résultats obtenus. La tâche de classification supervisée s'appuie sur des exemples étiquetés, dans notre cas (non-)'EP' grâce aux corpus annotés en EP à notre disposition, pour élaborer un modèle de classification capable de prédire l'étiquette associée à de nouvelles données. Ce système multilingue a été soumis à la compétition PARSEME 1.1 (ST) (Ramisch *et al.*, 2018) et évalué sur 19 des 20 langues proposées [1].

Après avoir détaillé les hypothèses en rapport avec sa conception et son optimisation (section 12.1), la section 12.2 présente la phase d'extraction, requise pour obtenir les exemples positifs et négatifs de tout corpus d'entraînement annoté en EP (noté `train`) permettant au classifieur de paramétrer son modèle de classification en fonction des traits (section 12.3) associés aux différents exemples. Survient ensuite la phase de classification des candidats d'un corpus (quelconque) de test, noté `test` (section 12.4). Chaque EP candidate du `test` à laquelle est attribuée l'étiquette 'EP' est finalement catégorisée d'après les catégories (p. ex. $\text{LVC}_{1.1}$) définies dans la ST. Pour conclure, un bilan est proposé dans la section 12.5.

12.1 Hypothèses

De l'hypothèse H_3 dérivent deux sous-hypothèses portant sur la capacité de généralisation de notre modèle :

H3 Connaître le profil de variabilité de chaque EP peut aider à identifier les variantes d'une EP apparaissant dans un nouveau texte. Par exemple, nous considérerons la séquence ***jeter l'éponge*** comme étant une EP mais pas *une éponge jetée, jeter des éponges, etc.* Il serait donc possible d'utiliser cette modélisation pour développer un système d'identification d'EP centré sur les variantes grâce à un classifieur entraîné sur ces profils.

H3a Utilisation d'autres types de corpus : en comparant les performances d'identification de variantes d'EP dans un corpus de `test` de nature similaire à celui de l'entraînement (à l'origine du modèle de classification) et dans un corpus différent, nous

1. L'arabe, optionnel en raison d'une mise à disposition tardive, n'a pas été pris en compte.

nous attendons à ce que les performances se dégradent dans le second cas.

D'autres hypothèses portent sur un gain de performances (H_4) et sur la pertinence des traits en général ou par famille de traits (H_5). La recherche d'optimisation du choix des traits sera abordée dans le chapitre 13.

H4 Enrichir le corpus d'entraînement initial par un autre corpus bien plus volumineux et annoté automatiquement en EP de la façon la plus fiable possible devrait permettre d'enrichir le profil de variabilité. Nous espérons donc de meilleures performances d'identification de variantes qu'avec le corpus initial de taille réduite.

H5 Certains traits sont plus pertinents que d'autres : certains traits peuvent notamment s'avérer plus intéressants selon qu'il s'agit d'EP fréquentes ou rares.

12.2 Extraction de candidats (ExtractCands)

Grâce à des EP observées dans les données d'entraînement, VarIDE est conçu pour en identifier des variantes. Il repose sur l'hypothèse que plus une expression candidate c est similaire à au moins l'une des EP annotées e, plus il est probable que c soit une EP. On estime cette similarité grâce aux caractéristiques (ou *traits*) observées chez c et e. Ces traits sont utilisés par un classifieur pour déterminer si c est une EP avérée (c'est-à-dire une variante d'une EP observée) ou non (lecture littérale ou co-occurrence fortuite).

L'entraînement du classifieur binaire requiert des candidats positifs ('EP') et négatifs ('non-EP') extraits à partir du `train`. Comme seuls les premiers sont fournis dans le `train`, des candidats négatifs sont obtenus par la même méthode d'extraction de candidats que celle qui sera appliquée sur le `test`. Cette extraction, dite `ExtractCands`, consiste à rechercher des co-occurrences de lemmes identiques à ceux d'EP annotées, tout en ajoutant une contrainte portant sur les permutations de POS tolérées. Cette contrainte a pour objectif de diminuer le bruit vis-à-vis de co-occurrences libres de lemmes, sachant de plus que les traits dépendent fortement des patrons syntaxiques de c et e.

12.2.1 Normalisation et génération de patrons

La simple recherche de co-occurrences de lemmes conduit à un bruit important, notamment si un composant est particulièrement fréquent (déterminant par exemple). Cela risque donc de générer un grand nombre de co-occurrences fortuites. Or, pour une EP telle que ***jeter l'éponge***, on sait qu'en français le déterminant ne pourra jamais être post-posé au nom. Nous proposons donc d'intégrer à la phase d'extraction de candidats une contrainte sur l'ordre des composants, apprise grâce aux EP attestées du `train`. Il faut pour cela tenir compte du fait que les composants d'une EP peuvent parfois tolérer différentes permutations de POS (VERB − DET − NOUN ou DET − NOUN − VERB pour ***tournera la page*** vs. ***la page tournée***). Nous recherchons donc, pour chaque ensemble de POS, les ordres les plus fréquemment observés en corpus, pour chacune des langues. Une double phase de normalisation est alors requise, la première étant la *LemmNorm* (définie en section 7.3) pour regrouper les tokens d'une même EP. La seconde, dite *POSnorm* et détaillée dans la section 12.2.1.1, spécifie les POS de chaque EP. En effet, d'après notre définition des composants, une modification syntaxique (p. ex. la passivation) peut modifier leur ordre, mais

n'affecte pas leur existence et n'entraîne pas non plus l'ajout de nouveaux composants. Les POS des composants subsistent donc, quelque soit leur patron syntaxique.

12.2.1.1 Normalisation des séquences de POS (*POSnorm*)


A l'instar de la *LemmNorm* qui représente la normalisation de lemmes, nous introduisons la *séquence de POS normalisée* (désormais *POSnorm*) afin de représenter toute EP comme une séquence d'étiquettes de POS dans l'ordre lexicographique à partir des POS de ses composants sous la forme $\langle \text{pos}_1 ; \text{pos}_2 ; \ldots ; \text{pos}_n \rangle$.

Comme signalé pour l'EP ***tourner la page***, ces étiquettes peuvent apparaître dans des ordres variables selon que l'on utilise par exemple la voix active ou passive. La *POSnorm* permet de s'affranchir de cet ordre variable. Les deux exemples précédents auraient en effet pour *POSnorm* ⟨DET ;NOUN,VERB⟩. La *POSnorm* conserve les cas de duplications de POS, par exemple ⟨ADP ;DET ;DET ;NOUN ;NOUN ;VERB⟩ pour ***prendre le***$_{\text{DET}}$ ***taureau***$_{\text{NOUN}}$ ***par les***$_{\text{DET}}$ ***cornes***$_{\text{NOUN}}$ qui contient deux déterminants et deux noms.

12.2.1.2 Génération de patrons

Chaque token d'EP attestée du `train` est identifié par sa *LemmNorm* et sa *POSnorm*. Nous lui adjoignons également la séquence observée des POS de ses composants (dite *POSseq*), comme illustré par la Table 12.1. Après avoir parcouru l'intégralité du corpus, chaque *POSnorm* est associée à la fréquence des *POSseq* correspondantes. La *POSnorm* ⟨NOUN ;VERB⟩ est ainsi davantage associée à la *POSseq* VERB − NOUN qu'à l'ordre inverse. Comme certaines EP peuvent ne pas montrer toute l'étendue des *POSseq* possibles dans le `train`, nous effectuons une extrapolation pour les EP rares ou affichant un nombre restreint de *POSseq* dans le `train` : en consultant la Table 12.1 de droite à gauche, nous obtenons une liste de permutations autorisées pour chaque EP partageant la même *POSnorm*. Par exemple, les EP associées à la *POSnorm* ⟨NOUN ;VERB⟩ (p. ex. ***prendre décision***) ont comme *POSseq* VERB − NOUN ou NOUN − VERB, contrairement à celles associées à la *POSnorm* ⟨PRON ;VERB⟩ (p. ex. ***en savoir***), qui n'apparaissent qu'avec la *POSseq* PRON-VERB.

De cette façon, la *LemmNorm* ⟨adresser ;refus⟩ sera associée aux deux *POSseq* NOUN − VERB et VERB − NOUN, même si cette dernière n'a jamais été observée pour l'EP en question. Cette extrapolation permet de prendre en compte la variabilité d'ordre des composants, mais cela ne signifie pas que l'intégralité des permutations est effectivement tolérée par toutes les EP partageant une même *POSnorm* (p. ex. ***le***$_{\text{DET}}$ ***bât***$_{\text{NOUN}}$ ***blesse***$_{\text{VERB}}$ vs. *blesse*$_{\text{VERB}}$ *le*$_{\text{DET}}$ *bât*$_{\text{NOUN}}$). Lorsqu'une même *LemmNorm* est associée à plus d'une *POSnorm* (p. ex. en raison d'erreurs d'annotation des POS), seule la plus fréquente est conservée. Par exemple, l'EP ***mener à bien*** ⇒ 'réussir' de *LemmNorm* ⟨bien ;mener ;à⟩ apparaît 5 fois avec la *POSnorm* ⟨ADP ;ADV ;VERB⟩ et une seule fois sous la forme ⟨ADP ;NOUN ;VERB⟩. Dans ce cas, l'annotation erronée de *bien* comme nom (comme dans *le bien a été vendu*) au lieu d'adverbe est minoritaire, la *POSnorm* associée à cette *LemmNorm* sera donc ⟨ADP ;ADV ;VERB⟩. Si plusieurs *POSnorm* ont la même fréquence, nous choisissons de façon arbitraire de conserver celle qui apparaît en premier dans l'ordre lexicographique.

12.2.2 Extraction de candidats positifs ('EP') et négatifs ('non-EP')

Pour être extrait par ExtractCands, un candidat doit non seulement disposer de lemmes identiques à l'EP de référence, il doit également respecter l'une des séquences de POS autorisée par sa *POSnorm*, condition qui n'est pas satisfaite dans *Il a apporté*$_{\text{VERB}}$ *"des*$_{\text{DET}}$ *éléments positifs"* à *Orange*$_{\text{PROPN}}$ [2] qui n'est pas l'EP ***apporter des oranges*** (PROPN à la place de NOUN), ni dans *sachez*$_{\text{VERB}}$ *en*$_{\text{PRON}}$ *profiter* (ordre VERB-PRON non autorisé). On conserve les n *POSnorm* les plus fréquentes pour chaque langue et leurs séquences de POS associées. Pour nos expériences, présentées dans le chapitre 12, la valeur de n a été fixée à 10. Cette valeur de 10 *POSnorm* permet de couvrir la majorité des occurrences lorsqu'on l'applique du corpus d'entraînement : sur les 19 langues, la couverture moyenne est de 92,5% [3] des occurrences au lieu de 84,7% avec les 5 *POSnorm* les plus fréquentes. Pour chaque *LemmNorm* du train dont la *POSnorm* appartient à ce top 10, nous générons toutes les permutations autorisées de lemmes et nous les recherchons dans le corpus, sans imposer de contrainte sur leur contiguïté, telle qu'une distance maximale entre les composants pris deux à deux.

Cependant, pour limiter la quantité de candidats non pertinents dans certaines langues (p. ex. à cause d'erreurs de segmentation), nous restreignons la longueur globale des discontinuités à une valeur maximale de 20. Cette valeur, choisie arbitrairement, correspond au maximum observé pour les EP annotées du corpus français. Cette contrainte est désormais nommée *Filtre20* et sera appliquée (ou non) selon qu'elle améliore (ou non) la performance de la Tâche sur le dev. Une étape de post-traitement, dédiée aux données du français, a également pour but de gérer certaines erreurs de lemmatisation observées pour les IRV$_{1.1}$. Par exemple, le token ***assurez-vous*** de *POSnorm* ⟨PRON ;VERB⟩ est lemmatisé ⟨assurer ;vous⟩ une fois dans train au lieu de ⟨assurer ;se⟩. Comme on restreint l'extraction aux EP vues 2 fois d'après leur *LemmNorm*, ⟨assurer ;vous⟩ ne serait pas extrait par ExtractCands, d'où cette étape de post-traitement qui (i) extrait les occurrences satisfaisant la *POSnorm* ⟨PRON ;VERB⟩ (figurant dans le top-10) et ayant une *LemmNorm* de fréquence 1 dans train, (ii) modifie ensuite les pronoms mal lemmatisés et ajoute ces candidats aux autres candidats extraits (qui bénéficient également de cette correction si nécessaire) à condition d'atteindre un total de 2 tokens pour l'EP considérée.

En effet, comme nous souhaitons vérifier la validité de notre hypothèse sur les profils de variabilité, nous ne retenons que les EP dont le nombre d'occurrences est suffisamment élevé pour être représentatif de leur variabilité. Cependant, ce seuil de fréquence ne peut pas être trop élevé, sinon la taille des données annotées diminuerait de façon importante. Pour un raisonnable compromis entre ces deux facteurs, seuls les candidats dont l'EP attestée apparaît au moins deux fois dans le train, ce qui représente 78% des EP annotées du corpus (3582 tokens), seront extraits.

Tout candidat extrait du train est alors étiqueté 'EP' s'il a été manuellement annoté comme EP, et 'non-EP' sinon, ce qui nous fournit respectivement des exemples positifs et négatifs qui seront utiles pour la phase d'entraînement du classifieur. Ce classifieur nécessite en outre que chaque exemple soit représenté par un ensemble de traits.

2. *Le Point*, 06/06/2013.
3. La médiane se situe à 95,3%.

LemmNorm	Exemple	*POSseq*	*POSnorm* par *LemmNorm*	*POSnorm* (fréquence)
⟨décision ;prendre⟩	***décisions prises***	NOUN − VERB	⟨NOUN ;VERB⟩ (2)	⟨NOUN ;VERB⟩ (7)
	prendre *une* ***décision***	VERB − NOUN		
⟨faire ;part⟩	***faisant part***	VERB − NOUN	⟨NOUN ;VERB⟩ (4)	
	fait part *(3)*			
⟨adresser ;refus⟩	***adresse*** *un* ***refus***	NOUN − VERB	⟨NOUN ;VERB⟩ (1)	
⟨en ;savoir⟩	*pour* ***en savoir*** *plus*			⟨PRON ;VERB⟩ (5)
	en savaient *trop*	PRON − VERB	⟨PRON ;VERB⟩ (2)	
⟨en ;finir⟩	*désireux d'****en finir*** (3)	PRON − VERB	⟨PRON ;VERB⟩ (3)	

TABLE 12.1 – Exemples d'EP, de leur *LemmNorm*, des séquences de POS observées et de leur *POSnorm*. Les patrons d'extraction autorisés (d'après ces exemples) sont : ⟨NOUN ;VERB⟩⇒ {NOUN − VERB, VERB − NOUN}, ⟨PRON ;VERB⟩⇒ {PRON − VERB}

En résumé, les contraintes exercées par `ExtractCands` sont rappelées ci-dessous :
— même *LemmNorm* que l'expression E de référence,
— même *POSnorm* que E, cette *POSnorm* faisant partie des 10 *POSnorm* les plus fréquentes pour la langue considérée,
— même POS de chaque composant que E,
— permutation de POS autorisée par la *POSnorm*,
— (optionnellement) filtre sur la longueur des discontinuités grâce à *Filtre20*, d'après les résultats obtenus sur le corpus `dev`,
— restriction aux EP vues 2 fois dans le `train`. Comme évoqué en section 11.2.2, il faut au moins deux tokens par EP pour apprécier la variabilité de celle-ci et, par conséquent, rendre les traits relatifs pertinents.

12.3 Choix des traits pour la classification

12.3.1 Traits adaptables par langue

Chaque exemple est décrit par un ensemble de paires trait-valeur. Un *trait* est défini comme une propriété (p. ex. la forme verbale au format UD VERBFORM) qui est associée à une valeur d'après l'ensemble des valeurs possibles pour ce trait (p. ex. *Inf*, *Ger*, *Conv*). On ne peut toutefois pas définir un ensemble figé de traits et de valeurs en raison des spécificités de chaque langue (p. ex. VERBFORM=*Conv(erb)* existe en croate mais pas en anglais). De telles spécificités s'observent à différents niveaux : dans les étiquettes de POS, dans les relations de dépendance et dans les caractéristiques morphologiques. On parcourt donc le corpus d'entraînement d'une langue donnée afin de répertorier l'ensemble des informations *a priori* pertinentes, soit au niveau des composants de l'expression (p. ex. le temps du verbe), soit au niveau de l'expression elle-même (p. ex. sa *LemmNorm*).

Les traits que nous utilisons peuvent par ailleurs être absolus ou relatifs.

12.3.2 Traits absolus (ABS) vs. relatifs (REL)

Pour rappel, nous émettons l'hypothèse H_3 que les profils de variabilité des EP peuvent aider à identifier correctement des occurrences d'EP déjà vues. Représenter ces profils de façon directe n'est pas toujours simple (p. ex. pour la passivation), surtout dans une optique d'indépendance des traits vis-à-vis de la langue. Notre point de départ est donc un ensemble de traits absolus (ABS) – dont nous fournissons des exemples dans la Table 12.2 – que l'on peut scinder en 5 familles :

- les traits liés aux EP annotées en corpus, ayant la même *LemmNorm* que l'EP candidate : leur catégorie (ABS_CATEP) et leur *LemmNorm* (ABS_LEMMNORM),
- les lemmes des composants (ABS_LEMME),
- les traits morphologiques représentant les propriétés de chaque composant de l'EP [4],
- les traits syntaxiques signalent les relations de dépendances sortantes d'un composant (p. ex. ABS_DEPSYN_NOUN $\in$ {NSUBJ, NOBJ}). Le trait ABS_DISTSYN_2COMP indique la valeur de la distance syntaxique dans le cas d'EP à 2 composants et ABS_DISTSYN_V-N dans le cas d'EP comportant un unique verbe et un unique nom. Ces deux traits ont été définis d'après les spécificités du français : 77% des EP du corpus `FR-train1.0` ont deux composants, et les patrons VERB-(PREP)-(DET)-NOUN couvrent 97% des $LVC_{1.0}$ et 54% des $ID_{1.0}$. ABS_TYPEDISTSYN précise par ailleurs si les composants sont directement connectés – ce qui est une information redondante avec une distance syntaxique nulle – ou s'ils bénéficient d'une connexion indirecte en série ou en parallèle (définies en section 9.3).
- les traits portant sur les discontinuités, c'est-à-dire les éléments insérés entre les composants, relèvent également de la syntaxe. Ils tiennent compte de la longueur des discontinuités définie par plages (ABS_LONGUEURDISCONT_0À5, etc.) et de leur séquence de POS sans intégrer d'informations sur leur contiguïté. La séquence de POS de la discontinuité sera ainsi ABS_DISCONTSEQ = ADV − ADV à la fois dans *il* ***fait*** *très*$_{\text{ADV}}$ *souvent*$_{\text{ADV}}$ ***face*** *à des obstacles* et dans *j'aimerais* ***en avoir*** *vraiment*$_{\text{ADV}}$ ***le cœur*** *bien*$_{\text{ADV}}$ ***net*** bien que dans ce dernier cas les deux adverbes soient séparés par un composant. Un dernier trait précise si une POS particulière figure parmi les discontinuités, par exemple ABS_DISCONT_DET s'intéresse à la présence d'au moins un déterminant dans la discontinuité. Ce trait binaire vaut 1 si cette condition est remplie et 0 dans le cas contraire.

Étant donné que les EP comportent plusieurs composants, il est nécessaire de distinguer les caractéristiques de chacun. Au lieu d'établir manuellement la liste des traits utiles, les traits morphologiques et syntaxiques servant à décrire chaque composant sont générés de façon automatique d'après les POS, les relations de dépendances et les propriétés morphologiques disponibles pour chaque langue. Pour le français, des traits morphologiques sont par exemple créés en combinant les 18 POS (NOUN, VERB, etc.) avec les 15 propriétés présentes (number, gender, etc.), ce qui conduit à 270 combinaisons pour ce seul mode de formation de traits, par exemple : NOUN_NUMBER, NOUN_GENDER, VERB_NUMBER, VERB_GENDER etc.

Sachant par ailleurs que certains traits décrivent les tokens d'une expression d'après

4. Flexion (p. ex. ABS_MORPH_NOUN_GENRE $\in$ {*Masc(ulin)*, *Fem(inin)*)}) et catégorisation p. ex. ABS_MORPH_NUMTYPE $\in$ {*Ord(inal)*, *Card(inal)*}.

leur POS, il est parfois nécessaire de spécifier de quel composant il s'agit lorsqu'une EP contient par exemple deux noms. Seul le cas d'ambiguïté liée à la duplication de POS est ici pris en compte, en l'occurrence pas ceux de triplication [5], etc. A partir des exemples 12.2-12.4, nous distinguons trois cas (Cas 1, Cas 2 et Cas 3) résumés dans la Table 12.2. Notons que les traits correspondant à ces trois cas sont systématiquement générés, initialisés à la valeur -1 mais modifiés uniquement s'ils sont pertinents pour un exemple donné. Cela implique que des informations a priori similaires seront, selon les cas, représentées à des endroits différents selon qu'une EP comporte par exemple un seul ou deux noms. Les trois cas mentionnés sont décrits ci-après :

— Cas 1 : aucune POS n'est dupliquée, si bien que chaque composant est identifiable par sa POS, comme dans l'exemple 12.2 qui contient un unique verbe (*prendre*) et un unique nom (*décision*),
— Cas 2 : les POS sont dupliquées mais peuvent être distinguées par les relations de dépendances entrantes associées au composant. Dans l'exemple 12.3, le premier nom est étiqueté OBJ tandis que le second est NMOD,
— Cas 3 : les POS sont dupliquées mais ne peuvent être distinguées par les relations de dépendances associées au composant, ce qui est illustré par les deux adverbes de l'exemple 12.4 (tous deux ADVMOD).

(12.1) [...] *suivre la* ***décision***$_{\text{NOUN.OBJ}}$ *que*$_{\text{PRON}}$ *Tibère*$_{\text{PROPN}}$ *va*$_{\text{VERB}}$ ***prendre***$_{\text{VERB}}$*.* [6] (`FR-train1.1`)

(12.2) [...] *pour que la commission* ***prenne*** *des*$_{\text{DET}}$ ***décisions***$_{\text{NOUN.NOBJ}}$ (`FR-train1.1`)

(12.3) [...] *la Commune* ***tire son épingle***$_{\text{NOUN.OBJ}}$ ***du jeu***$_{\text{NOUN.NMOD}}$ [7] (`FR-train1.1`)

(12.4) [...] *les* ***malheurs***$_{\text{NOUN.NSUBJ}}$ ***ne***$_{\text{ADV.ADVMOD}}$ ***surviennent jamais***$_{\text{ADV.ADVMOD}}$ ***seuls***. (`FR-train1.1`)

(12.5) *Il* ***prend*** *d'*$_{\text{DET}}$ *importantes*$_{\text{ADJ}}$ ***mesures*** (référence)

(12.6) *Ils* ***prirent*** *cette*$_{\text{DET}}$ *nouvelle*$_{\text{ADJ}}$ ***mesure*** (candidat)

(12.7) *Il* ***prend*** *des*$_{\text{DET}}$ ***mesures*** *aussi (voire de plus en plus) importantes*$_{\text{ADJ}}$ (candidat)

Les traits portant sur les discontinuités et les dépendances syntaxiques peuvent véhiculer des informations similaires, comme dans l'exemple 12.5 où l'adjectif est compté à la fois comme une insertion et comme un modifieur du nom. Mais ils peuvent aussi être complémentaires, comme dans (Ex. 12.7), car les dépendances syntaxiques peuvent prendre en compte des modifieurs situés en dehors de la fenêtre d'annotation.

D'un autre côté, les traits relatifs (REL) sont obtenus en comparant les propriétés de chaque candidat d'EP c avec celles des tokens d'EP du même type annotés dans `train` ($\{e_1, e_2, \ldots, e_n\}$) nous servant de référence. Cette comparaison ne s'applique que si $c \neq e_i$ [8]

5. Dans de tels cas, on affecte la valeur -1 aux traits portant sur les POS multiples, par exemple : NOUN_NUMBER $= -1$ pour l'exemple *il* ***a un petit pois***$_{\text{NOUN}}$ ***à la place***$_{\text{NOUN}}$ ***du cerveau***$_{\text{NOUN}}$ $\Rightarrow$ 'il est peu intelligent' qui contient trois noms.

6. Cet exemple ne respecte pas les préconisations de UD (Nivre *et al.*, 2016) car ici le verbe *prendre* dépend de *va* qui lui-même dépend de *décision*, mais nous reproduisons ici les exemples tels qu'ils apparaissent en corpus.

7. Dans cet exemple, le nom *jeu* aurait dû bénéficier de l'étiquette OBL et non NOBJ.

8. Cette condition n'est requise que pour la phase d'entraînement puisque c est alors extrait du même corpus que e_i. En particulier, si $c = e_j$ avec $i \neq j$, c'est-à-dire si le candidat est une EP annotée, la comparaison s'applique, mais le candidat n'est jamais comparé à lui-même.

tout en partageant la même *LemmNorm*, c'est-à-dire *LemmNorm*(c) = *LemmNorm*(e_i). Ces traits relatifs ont pour objectif de capturer la similarité d'un candidat avec les EP annotées comparables. Ces traits relatifs prennent des valeurs binaires booléennes : *faux*, si aucune équivalence avec une quelconque occurrence attestée de la même EP n'a été trouvée, *vrai* si au moins une équivalence a été trouvée. Ils demeurent parfois à leur valeur initiale *-1* si la comparaison est impossible, ce qui se produit pour des hapax. Ainsi, le trait REL_DiscontSeq est *vrai* si les discontinuités de *c* sont identiques à celle d'au moins l'une des e_i, comme dans (Ex. 12.6) vs. (Ex. 12.5), et *faux* sinon, comme dans (Ex. 12.7) vs. (Ex. 12.5). De même, le trait REL_DepSyntax_VERB est *vrai* si les relations de dépendances sortantes du verbe de *c* sont identiques à celles observées pour au moins l'une des e_i, et *faux* sinon.

De façon générale, chaque trait ABS est assorti du trait REL correspondant, hormis les traits portant sur la séquence de lemmes et la catégorie d'EP : en comparant deux occurrences d'un unique type d'EP, leur valeur serait, par définition, toujours *vrai*, ce qui les rend inutiles. La proximité entre un candidat et le profil de variabilité d'une EP est définie au niveau des types d'EP (traits REL) plutôt que de leurs tokens (traits ABS). Toutefois, la distribution zipfienne des EP sous-entend un manque de fiabilité pour les types d'EP apparaissant rarement en corpus, d'où l'intérêt de considérer à la fois les traits ABS et les traits REL.

A l'issue du processus de génération de traits, 18 000 traits sont créés c'est-à-dire un nombre bien supérieur à celui des EP annotées du corpus d'entraînement, ce qui expose au risque de surapprentissage. Dans le cadre de notre démarche exploratoire, tous ont été conservés dans un premier temps bien que certains soient de toute évidence (i) invalides en raison de la combinatoire utilisée pour la génération des traits (p. .ex. le genre d'un adverbe), (ii) corrélés (p. ex. le genre d'un nom et celui de l'adjectif qui le modifie), (iii) peu ou non informatifs (p. ex. l'absence de propriétés morphologiques dans le corpus est signalée sous la forme d'un underscore, ce qui s'observe pour les prépositions, adverbes, etc. qui en sont dépourvus). Nous nous attendons à ce que les traits superflus soient aisément écartés de la phase de classification grâce à leur valeur initiale non modifiée.

Description des traits	Nom des traits t	Ex. 12.1 (ABS_t)	Ex. 12.2 (ABS_t)	Ex. 12.2 vs. 12.1 (REL_t)	Ex. 12.3 (ABS_t)	Ex. 12.4 (ABS_t)
LemmNorm	LEMMNORM	⟨décision ;prendre⟩		n/a n/a	⟨de ;épingle ;jeu le ;son ;tirer⟩	⟨jamais ;malheur ; ne ;seul ;survenir⟩
Catégorie d'EP	CATEP	*LVC.full*	*LVC.full*	n/a	*VID*	*VID*
Séquence de POS insérées	DISCONTSEQ	PRON-PROPN-VERB	DET	*faux*	∅	∅
Insertion de chaque POS existante (v = *1* si présent, *0* sinon)	DISCONT_DET DISCONT_VERB	*0* *1*	*1* *0*	*faux* *faux*	*0* *0*	*0* *0*
Si EP avec 1 seul verbe et 1 seul nom (sinon v = *-1*) : • Distance syntaxique • Nature du lien de dépendance syntaxique : *direct* (parent-enfant), *série* (ancêtre indirect), ou *parallèle* (même ancêtre)	 DISTSYN_V-N TYPEDISTSYN_V-N	 *1* *série*	 *0* *direct*	 *faux* *faux*	 *-1* *-1*	 *0* *direct*
Si EP à 2 composants (sinon v = *-1*) : • Distance syntaxique • Nature du lien de dépendance syntaxique	 DISTSYN_2ELTS TYPEDISTSYN_2ELTS	 *1* *série*	 *0* *direct*	 *faux* *faux*	 *-1* *-1*	 *-1* *-1*
Lemme de chaque composant (v = *-1* si non pertinent) Ajout de l'étiquette de dépendance syntaxique pour gérer les cas de POS dupliquées	LEMME_NOUN LEMME_ADV LEMME_NOUN-OBJ LEMME_NOUN-NMOD LEMME_ADV-ADVMOD	*décision* [cas 1] *-1* *-1* [cas 1] *-1* [cas 1] *-1*	*décision* [cas 1] *-1* *-1* [cas 1] *-1* [cas 1] *-1*	*vrai* *-1* *-1* *-1* *-1*	*-1* [cas 2] *-1* *épingle* [cas 2] *jeu* [cas 2] *-1*	*malheur* *-1* [cas 3] *-1* *-1* *-1* [cas 3]
Morphologie de chaque composant (v = *-1* si non pertinent)	MORPH_VERB-MOOD MORPH_VERB-PERSON	*-1* *-1*	*Ind* *3*	*faux* *faux*	*Ind* *3*	*Ind* *3*
Dépendances entrantes de chaque composant (v = *1* si satisfait, *0* sinon)	DEPSYN_NOUN-OBJ	*1*	*0*	*faux*	*1*	*0*

TABLE 12.2 – Aperçu de traits ABSolus et RELatifs pour les exemples 12.2 (RELatif à 12.1), 12.3 et 12.4. Les traits relatifs portant sur la *LemmNorm* et la catégorie sont indiqués comme n'étant pas applicables car, par définition, ils seraient systématiquement *vrai*.

Grâce aux exemples provenant du `train` décrits au moyen des traits absolus et relatifs et étiquetés selon les deux classes (non-)'EP', un modèle de classification tente de garantir les prédictions les plus fiables possibles. Une fois ce modèle établi, il est appliqué sur les nouveaux exemples du corpus de `test` afin de prédire la classe devant leur être attribuée.

12.3.3 Classification d'EP et attribution de catégories

Tout d'abord, pendant la phase de prédiction, nous extrayons des candidats du `test` suivant la procédure `ExtractCands` décrite dans la section 12.2.2, sauf que l'on ignore ici quels sont les exemples positifs et négatifs. Ensuite, nous utilisons le classifieur Naïve Bayes de NLTK [9] (décrit en section 6.1.2.2) pour classer les candidats comme positifs/négatifs d'après leurs valeurs associées aux différents traits. Les valeurs des traits absolus sont obtenues comme décrit pour l'extraction de candidats dans le corpus `train`. Quant aux traits relatifs, leurs valeurs sont obtenues par comparaison de la candidate c du corpus de `test` avec toutes les EP du `train` de même *LemmNorm* : pour un trait donné, si la même valeur absolue est trouvée dans le candidat à classifier et dans au moins une occurrence de E dans le `train`, alors le trait relatif booléen prendra la valeur *vrai*, et *faux* sinon.

Les traits absolus incluent également la *LemmNorm* de l'EP, p. ex. ⟨décision ;prendre⟩, ainsi que leur catégorie (p. ex. LVC) dans le corpus d'entraînement. Après classification, la catégorie d'EP (p. ex. $\text{LVC}_{1.1}$) du candidat est obtenue d'après la catégorie la plus fréquemment assignée à sa *LemmNorm* dans le `train`. Nous précisons à ce propos que la catégorisation des EP n'intervenait pas dans le calcul de la F-mesure comme pour la ST.

12.4 Évaluation de VarIDE

Les corpus utilisés pour VarIDE sont `xx-train1.1` pour l'entraînement du système et `xx-test1.1` pour l'évaluation, xx indiquant la langue considérée suivant la convention mentionnée en section 8.1.1. La Table 12.3 montre le nombre de candidats idiomatiques ('EP') et non-idiomatiques ('non-EP') extraits du `train` pour l'entraînement du classifieur, avec le ratio d'exemples 'EP' (% EP). On mentionne également le rappel $R_{\text{variant-of-train}}$ avant classification (c.-à-d. après l'extraction de candidats) et après classification.

12.4.0.1 `ExtractCands` : extraction de candidats

Nous obtenons une couverture satisfaisante avec le top-10 *POSnorm*, avec $R > 80$ pour 17 des 19 langues (62 et 75 pour l'italien et l'allemand). Le rappel à ce stade pour le français est de 93. Ces valeurs de rappel permettent d'évaluer l'influence de la sélection de 10 *POSnorm* : l'italien faisait en effet partie des deux langues pour lesquelles la couverture évaluée sur le `train` était la plus faible (84%). Toutefois, l'hébreu qui bénéficiait de la couverture la plus faible (75% dans le `train`) dispose dans le `test` d'un rappel de 85. Seuls le farsi et le hongrois ont le rappel maximal, nous pourrions donc envisager une prise en compte d'un nombre plus élevé de *POSnorm* afin d'augmenter le rappel lors de l'extraction de candidats. De plus, la performance sur les variantes telles que définies pour

9. Implémentation de la bibliothèque NLTK : `http://www.nltk.org`

Lang.	F per-EP	$F_{\text{var-of-train}}$ per-EP	Var-of-train % TEST1.1	Candidats de TRAIN pour l'entraînement du classifieur			$R_{\text{var-of-train}}$ avant classif.	$R_{\text{var-of-train}}$ après classif.	format UD	dépend. syntax. fournies	*Filtre 20*
				# EP	# non-EP	% EP					
ES	25,30	18,83	52%	1580	3414	32	89,66	83,45	x	x	x
FR	55,4	57,22	50%	4303	5089	46	92,86	89,68	x	x	
IT	32,50	32,26	62%	2755	5721	32	62,50	57,07		x	x
PT	60,84	65,74	59%	4171	4014	51	94,42	70,82	x	x	x
RO	71,15	26,13	12%	4636	5501	46	96,92	89,23	x	x	x
DE	15,30	26,14	59%	2437	1114	69	75,68	15,54	x	x	
EN	24,17	52,50	53%	316	336	48	94,74	55,26	x	x	x
BG	62,52	58,42	36%	5031	6637	43	96,25	85,62	x	x	x
HR	12,57	24,47	73%	1381	843	62	96,98	14,57	x	x	
LT	1,96	5,15	83%	301	**96**	**76**	99,46	2,69	x		x
PL	11,25	22,05	60%	3954	2119	65	95,07	12,56	x	x	
SL	42,34	39,08	73%	2281	13330	15	98,12	96,24		x	x
EL	34,77	46,76	68%	1270	1341	49	92,39	32,99	x	x	
EU	52,31	34,82	39%	2499	5147	33	94,51	92,68		x	x
FA	44,95	58,06	53%	2437	1707	59	100	43,11	x	x	
HE	18,62	21,57	41%	932	820	53	84,72	15,28	x	x	
HI	56,80	72,24	49%	526	463	53	95,00	67,86	x	x	x
HU	18,69	6,49	21%	6187	516	**92**	100	3,36		x	
TR	7,87	25,98	60%	5802	**156652**	**4**	97,33	97,33		x	x

TABLE 12.3 – Pour chaque famille de langues (romanes, germaniques, slaves et autres) : performances d'identification d'EP de VarIDE sur l'ensemble des EP (vues + non-vues), sur les *variant-of-train*, proportion de variantes dans le `test`, résultat d'extraction de candidats (quantité par étiquette 'EP' vs. 'non-EP', proportion d'"EP'), rappel avant et après classification pour les *variant-of-train*. Les trois dernières colonnes fournissent des détails sur les spécificités des corpus (présence d'étiquettes au format UD, relations de dépendances syntaxiques disponibles) ainsi que le recours (ou non) à *Filtre20* pour limiter la longueur des discontinuités, ce choix étant effectué d'après les performances observées sur le `dev`.

la ST (*variant-of-train*) dépend de leur proportion en corpus qui varie de 12% (roumain) à 83% (lituanien). Malgré un nombre restreint de *variant-of-train* en roumain, la *F-per-EP* tire bénéfice des occurrences *identical-to-train* correctement identifiées.

12.4.0.2 `Classif` : Classification de candidats

La performance de classification (F et R) est sensible à la fiabilité des corpus annotés, elle peut être pénalisée à la fois par des faux positifs ((`EN`) *UV lights*$_{\text{NOUN}}$ *up*$_{\text{VERB}}$ *the temperature* $\neq$ *to* ***light***$_{\text{VERB}}$ ***up***$_{\text{ADP}}$) et des faux négatifs.

Le déséquilibre entre EP et non-EP, qu'il s'agisse de la surreprésentation d'EP, comme en hongrois (92%), ou l'inverse en turc (4%), peut également avoir un impact néfaste. Il aurait peut-être fallu ici procéder à un équilibrage des classes dans les données d'entraînement : en ayant un nombre équivalent d'exemples dans les deux classes 'EP' et 'non-EP', on éviterait qu'une classe sur-représentée soit systématiquement prédite par le modèle de classification. Il ne faut d'ailleurs pas seulement prendre en compte les proportions : en lituanien, seuls 96 candidats extraits du `train` sont classés comme 'non-EP'. Dans ce cas, un classifieur risque de ne pas avoir assez de contre-exemples pour acquérir des connaissances sur les traits.

Après classification, 8 langues demeurent avec un rappel supérieur à 70, mais la F-mesure ne dépasse 50 que pour six langues : français, portugais, roumain, bulgare, basque et hindi. C'est d'ailleurs pour trois de ces langues que VarIDE occupe une place parmi le top-3 dans le classement de la ST : premier pour le bulgare, deuxième pour le français et le portugais (auxquels s'ajoute le slovène), ce qui souligne la capacité de généralisation multilingue.

Les traits devraient être améliorés pour optimiser à la fois le rappel et la précision de classification. D'autres problèmes sont dus à des jeux d'étiquettes n'étant pas au format UD (ce qui a nécessité des ajustements pour quelques langues), des erreurs de segmentation de phrases (gérées avec *Filtre20*) et des cas de lemmatisation manquantes (p. ex. en turc avec la *LemmNorm* ⟨_ ;fotoğraf⟩ pour ***prendre photographie***, dans laquelle le lemme du verbe est absent). De même en espagnol, les IRV sont lemmatisés avec le pronom *él* au lieu de *se* (`ES`) (⟨empeñar ;él⟩ vs. ⟨empeñar ;se⟩), ce qui crée une ambiguïté particulièrement préjudiciable en raison de la fréquence du pronom personnel sujet *él*. Cela peut expliquer pourquoi, malgré des similarités entre le français et l'espagnol, toutes deux des langues romanes, ayant également une proportion similaire de variantes (Table 12.3) et un rappel similaire des variantes après classification, la F-mesure est nettement plus faible pour l'espagnol (57 vs. 19) en raison d'une précision plus faible.

12.4.0.3 Performance vis-à-vis des autres systèmes de la ST

Performances pour l'ensemble des langues Au classement général (EP vues et non vues) des systèmes de la ST, VarIDE se situe en 5$^{\text{e}}$ position (macro-moyenne $F = 45{,}97$) sur les 13 systèmes ayant utilisé uniquement les corpus fournis durant la compétition (dit mode fermé [10]). Le système le plus performant tous modes confondus (ouvert, fermé) est

10. Les classements figurant dans cette section se restreignent aux seuls systèmes de ce mode.

SHOMA* [11] ($F = 58$). Les quatre systèmes les plus performants en mode fermé sont : TRAVERSAL ($F = 54$), TRAPACC_S (49,74), TRAPAAC (49,57) et CRF-Seq-nocategs (46,11) (Waszczuk, 2018; Stodden *et al.*, 2018; Moreau *et al.*, 2018). CRF-Seq-nocategs tire profit de lemmes et d'informations morphosyntaxiques à l'aide d'un modèle de séquences de type CRF (fonctionnement décrit en section 6.1.2.2) ; TRAVERSAL s'appuie également sur le principe des CRF mais en l'appliquant à des arbres de dépendance syntaxique au lieu de séquences ; TRAPACC_S et TRAPAAC reposent sur l'utilisation de réseaux de neurones (combinés avec un SVM pour TRAPACC_S) pour créer des représentations denses qui sont ensuite données à un modèle de transition adapté de TRANSITION, le modèle le plus performant lors de la compétition de 2017 (Al Saied *et al.*, 2017).

Concernant l'identification d'EP vues, donc de variantes selon notre définition (section 7.3), VarIDE obtient une F-mesure de 62,8 (7^{e} position) : 53,33 pour les *variant-of-train* (6^{e} position) et 67,69 pour les *identical-to-train* (8^{e} position). On remarque ici que malgré un score moins élevé pour les *variant-of-train* que pour les *identical-to-train*, VarIDE est moins affecté que les autres systèmes par l'existence de variantes ayant une forme de surface différente. C'est d'ailleurs le deuxième système, après TRAVERSAL (37,4 vs. 44,36), le plus performant pour la prise en compte des discontinuités (et le troisième si l'on ajoute les systèmes en mode ouvert).

VarIDE obtient par ailleurs des résultats très contrastés pour la Tâche selon les différentes langues :

— la F-mesure dépasse 50 pour le roumain (71 - 6^{e}), le bulgare (62 - 1^{e}), le portugais (61 - 2^{e}), le hindi (57 - 7^{e}), le basque (52 - 7^{e}), le français (50 - 2^{e}),

— la F mesure se situe entre 20 et 50 pour le farsi (45 - 9^{e}), le slovène (42 - 2^{e}), le grec (35 - 9^{e}), l'italien (32 - 6^{e}), l'espagnol (25 - 5^{e}), l'anglais (25 - 6^{e}), l'allemand (25 - 10^{e}),

— la F mesure est inférieure à 20 pour le hongrois (19 - 9^{e}), l'hébreu (25 - 5^{e}), le croate (13 - 9^{e}), le polonais (11 - 10^{e}), le turc (8 - 7^{e}), le lituanien (1 - 6^{e}).

Ces niveaux de performances et classements mettent en évidence qu'une F mesure élevée n'est pas toujours synonyme d'un meilleur classement : malgré une F-mesure de 42, VarIDE est en deuxième position pour le slovène (après TRAVERSAL qui obtient 64) mais, avec 57 pour l'hindi, il ne figure qu'en septième position (CRF-DepTree-categs se situe en tête avec 72). Cela traduit le fait que les systèmes ont tendance à présenter des niveaux de performance similaires pour une langue donnée : F-mesure élevée pour le roumain (5 systèmes au-delà de 80), faible pour l'hébreu (au maximum 23 pour TRAVERSAL). On remarque par ailleurs que les meilleures performances de VarIDE s'obtiennent essentiellement sur des langues romanes (roumain, français, portugais), mais on retrouve également le bulgare, l'hindi et le basque, sans que l'on puisse expliquer précisément pourquoi VarIDE réussit mieux la Tâche sur ces langues. Il est également difficile d'analyser en profondeur, pour chaque langue, les raisons pour lesquelles VarIDE échoue. Pour le français, l'analyse d'erreur du `test` serait limitée à 266 faux négatifs et 188 faux positifs [12]. Quelques éléments de réponse sont cependant avancés dans la section 12.5.1.

11. Nous ajoutons un astérisque pour signaler les systèmes fonctionnant en mode ouvert, c'est-à-dire utilisant des ressources externes aux seuls corpus fournis durant la compétition.

12. Cette difficulté a d'ailleurs été illustrée en section 11.3.2, page 151.

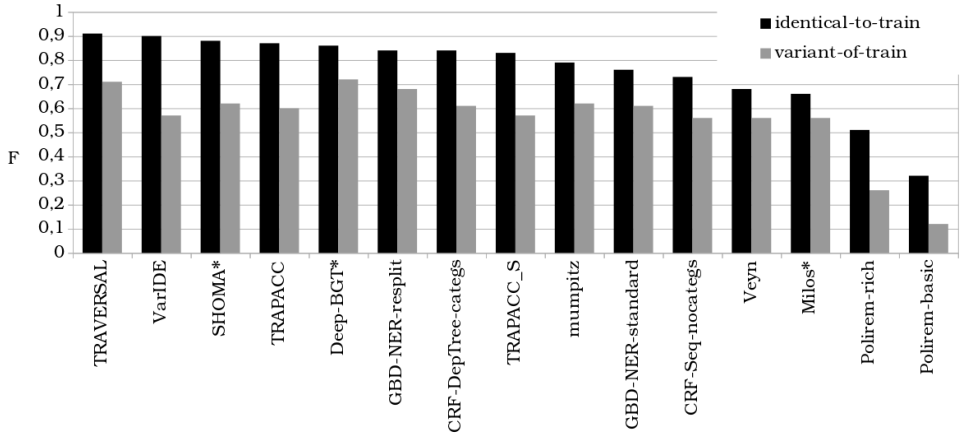


FIGURE 12.1 – Performances des 15 systèmes ayant participé à la ST pour le français avec une focalisation sur les données vues (*seen-in-train*) selon qu'il s'agit de *identical-to-train* ou de *variant-of-train*.

Performances pour le français Si l'on compare les performances d'identification de variantes d'EP de VarIDE à celles des autres systèmes de la ST pour le français (Fig. 12.1), on remarque que c'est essentiellement sur les variantes ayant une forme de surface identique qu'il se distingue : la F-mesure sur les *identical-to-train* est alors inférieure d'un point à celle du meilleur système (TRAVERSAL). Sur les *variant-of-train*, il n'est plus qu'en neuvième position (15 points de F-mesure d'écart avec TRAVERSAL). L'optimisation de VarIDE, faisant l'objet du chapitre 13, aura pour objectif d'améliorer ces performances.

12.5 Bilan

Le développement de ce système a donné lieu à plusieurs réflexions, portant sur la nature des données d'entrée et leur modélisation, sur les modalités de mise en oeuvre de la phase de `Classif` ainsi que sur la pertinence des traits définis.

12.5.1 Nature des données

Notre méthodologie de représentation des données peut introduire certains biais lors des normalisations mises en œuvre (*LemmNorm* et *POSnorm*). La qualité d'annotation des données peut aussi induire des biais préjudiciables. Les limites de la *LemmNorm*, évoquées dans la section 10.2.3 (1 EP associée à plusieurs *LemmNorm* ou l'inverse), peuvent conduire à comparer des occurrences d'EP distinctes lors de l'utilisation des traits relatifs, mais cela nous semble un problème mineur compte tenu du peu d'EP concernées. Si l'utilisation des permutations de POS autorisées est particulièrement efficace pour certaines langues ($F = 94,94$ pour le hongrois sur les EP vues), elle implique parfois des généralisations injustifiées. Prenons l'exemple des 2 EP ***la page tournée*** ⇒ 'commencer une nouvelle étape d'une vie en oubliant le passé' et ***le bât blesse*** ⇒ 'cela fait mal'. Comme elles partagent la même

POSnorm ⟨DET ;NOUN ;VERB⟩, nous leur attribuerons des ordres possibles de composants similaires, mais parfois inappropriés (***tournerai la page*** vs. ? ? *blesser le bât*). L'impact de cette approximation pourrait être minoré en associant l'ordre observé à la voix (active ou passive), mais non supprimé car il existe des EP pour lesquelles l'ordre NOUN − VERB ne sera jamais toléré avec maintien du sens idiomatique initial : ***faire partie*** vs. # *partie faite*.

L'analyse globale des performances de VarIDE est difficile à mener sans tenir compte des particularités de chaque langue. La qualité d'annotation se révèle particulièrement problématique : en turc, 37% des EP du `TR-train1.1` ont des lemmatisations partiellement ou complètement absentes. Ce problème a également été rencontré sur les données du corpus `FR-test1.1` issues du corpus PUD : 25% (127/498) des occurrences d'EP étaient concernées. Cela pénalise VarIDE puisque `ExtractCands` s'appuie sur les lemmes : 93 des 127 occurrences mal lemmatisées de `FR-test1.1` étaient ainsi des variantes d'EP *vues* durant l'entraînement et non des *non vues*.

Par ailleurs, si l'on se focalise sur la précision de l'identification de variantes, le risque d'intégrer des lectures littérales semble mineur dans 5 langues différentes (Savary *et al.*, 2019b), toutes ayant pris part à la ST (basque, allemand, grec, polonais et portugais). Restent alors les co-occurrences fortuites. L'une des pistes à explorer pourrait être de mesurer le risque de les voir subsister après `ExtractCands`. Plus les composants des EP sont fréquents dans des expressions libres, plus le risque d'erreur est important. En français par exemple, les verbes *avoir* et *être*, fréquents en dehors d'EP, constituent à eux seuls 15% des tokens d'EP du `FR-train1.1`, et 8,6% des types d'EP. En ajoutant le verbe *prendre*, cela couvre presque 20% des tokens. Parmi les autres spécificités intéressantes, on peut également citer le nombre de composants des EP : en français 77% des occurrences du `FR-train1.1` des EP comportent deux composants, alors qu'en hongrois 74% n'en ont qu'un seul. Cela peut expliquer pourquoi `ExtractCands` est si performante pour le hongrois : le risque de co-occurrences fortuites y est nettement moindre.

12.5.2 Remarques sur la phase de `Classif`

Nos essais préliminaires de classification en établissant une distinction selon les catégories d'EP (c. à-d. un classifieur pour les idiomes, un autre pour les constructions à verbe support, etc.) ou les familles de traits (morphologiques, linéaires d'après les discontinuités, ou syntaxiques) ne s'étaient pas avérés concluants, et ont donc été écartés de notre méthodologie. Il existe toutefois une corrélation *a priori* assez forte entre certaines catégories et par exemple les séquences d'insertions tolérées, ce qui peut créer un biais au niveau de la classification. Nous supposons que c'est la grande variété de profils syntaxiques des $\text{VID}_{1.1}$ qui avait pu nuire à cette approche par catégorie, et qu'il conviendrait de s'y intéresser de nouveau en se concentrant dans un premier temps sur les catégories très distinctes que sont les $\text{LVC}_{1.1}$ et les $\text{IRV}_{1.1}$ afin d'évaluer la pertinence de classifieurs par catégories d'EP.

Notre hypothèse de départ était que la similarité de caractéristiques d'une EP candidate avec une forme attestée de la même EP (trait REL) ou avec d'autres types d'EP (trait ABS), pourrait permettre de déterminer si ce candidat devait être considéré comme EP ou non. Or, comme le souligne la Table 12.4, l'ajout de cette étape de classification après l'extraction n'est pas toujours avantageuse. Si, pour 11 langues, cette étape s'accompagne

	ExtractCands		Classif		Différence Classif vs. ExtractCands	
Langue	$F_{seen\text{-}in\text{-}train}$	$F_{variant\text{-}of\text{-}train}$	$F_{seen\text{-}in\text{-}train}$	$F_{variant\text{-}of\text{-}train}$	$F_{seen\text{-}in\text{-}train}$	$F_{variant\text{-}of\text{-}train}$
SL	20,74	16,14	46,12	39,08	**25,38**	22,94
HI	58,34	42,09	79,48	72,24	21,14	**30,15**
FR	54,55	38,55	70,03	57,22	15,48	18,67
BG	59,86	39,9	74,95	58,42	15,09	18,52
RO	60,08	16,34	72,43	26,13	12,35	9,79
PT	60,79	48,35	72,80	65,74	12,01	17,39
EN	44,70	34,20	56,09	52,50	11,39	18,30
IT	29,79	22,35	40,24	32,26	10,45	9,91
ES	19,73	11,94	28,54	18,83	8,81	6,89
TR	29,38	20,39	35,95	25,98	6,57	5,59
EU	48,88	29,16	55,27	34,82	6,39	5,66
EL	53,80	45,90	52,30	46,76	-1,50	0,86
FA	70,35	57,29	62,74	58,06	-7,61	0,77
HE	72,64	57,28	40,82	21,57	-31,82	-35,71
DE	65,69	56,14	28,09	26,14	-37,60	-30,00
HR	71,99	65,20	21,52	24,47	-50,47	-40,73
PL	69,57	59,47	15,23	22,05	-54,34	-37,42
LT	75,93	72,55	4,27	5,15	-71,66	-67,40
HU	94,94	79,89	20,41	6,49	**-74,53**	**-73,40**

TABLE 12.4 – Évaluation de l'impact de l'ajout d'une phase de classification `Classif` après la phase d'extraction de candidats `ExtractCands` pour chaque langue : on précise la F-mesure – sur les *seen-in-train* et les *variant-of-train* – obtenue après chacune de ces étapes et les deux dernières colonnes mettent en évidence la différence $F_{\texttt{Classif}}$ - $F_{\texttt{ExtractCands}}$.

d'un gain de F-mesure, pour 8 autres elle engendre au contraire une baisse significative. Pour ces langues il aurait été judicieux, dans l'optique de performances de la ST, de se restreindre à la seule phase d'extraction, dont la vocation initiale était de limiter les co-occurrences fortuites. L'origine de telles divergences sur l'intérêt de la classification des candidats est à rechercher dans les spécificités desdites langues. Les langues qui bénéficient le plus de l'étape `Classif` sont le slovène pour les EP *seen-in-train* (F augmentée de 26 points) et le hindi pour les *variant-of-train* (+ 30 points). A l'inverse, en hongrois, cela occasionne une baisse considérable (baisse de plus de 70 points à la fois sur les *seen-of-train* et sur les *variant-of-train*), allant jusqu'à presque annihiler l'identification pour cette dernière catégorie ($F = 7$). En français, avant `Classif`, nous obtenons : $P = 32{,}49$, $R = 49$ et $F = 39{,}07$, tandis qu'après cette étape le rappel diminue légèrement ($R = 46{,}59$), mais la précision augmente de façon notable ($P = 55{,}24$), d'où une F-mesure finale améliorée ($F = 50{,}54$).

12.5.3 Pertinence des traits

Les *word embeddings* ne font pas partie de l'ensemble de traits initial. Cela se justifie par l'inflexibilité lexicale des EP : il est rare qu'une EP puisse être découverte par similarité sémantique de ses composants par rapport à une EP attestée (cela ne fonctionne par

exemple pas pour ***prendre*** vs. *appréhender*[13] *une* ***décision***). Cette observation est confirmée par la ST, dans laquelle les systèmes ayant les meilleures performances d'identification ne dépassent jamais une F-mesure de 28 pour des EP non vues au préalable. D'après nous, ce très faible pouvoir de généralisation pourrait être pris en compte, à l'instar de Savary *et al.* (2019a), en couplant de façon systématique l'identification d'EP avec leur découverte automatique, ce qui représente une perspective pour de futurs travaux.

Notons par ailleurs qu'une étude préliminaire de l'intérêt des traits ABS et REL avait été menée sur la fraction des EP de patron VERB – (DET) – NOUN (Pasquer *et al.*, 2018c)[14]. Nous avions alors montré que le trait REL portant sur les séquences de POS des discontinuités et les traits ABS portant sur l'existence de discontinuités spécifiques (VERB, PUNCT) étaient particulièrement discriminants pour attribuer respectivement l'étiquette 'EP' ou 'non-EP'. Cette analyse des traits pertinents, effectuée de façon manuelle, gagnerait à être automatisée, ce qui permettrait de déterminer quels traits sont les plus pertinents pour le français, tout en tenant compte de l'intégralité des catégories d'EP, et pas uniquement de celles de patron VERB – (DET) – NOUN.

12.5.4 Disponibilité de VarIDE

VarIDE a été mis en ligne sous la forme d'un démonstrateur mis au point par le laboratoire ATILF[15] : il suffit de choisir la langue, le système, de copier/coller ou d'importer un texte (brut ou bien au format CoNLL-U[16]) pour obtenir le texte annoté en EP verbales en sortie avec les étiquettes de catégorie ($\text{LVC.full}_{1.1}$, $\text{VID}_{1.1}$, etc.) (Fig. 12.2).

La recherche automatisée de traits pertinents permettant de distinguer 'EP' et 'non-EP', et exploitée pour la nouvelle version de VarIDE (VarIDE+) fait l'objet du chapitre 13.

13. Alors que ces deux verbes partagent une similarité contextuelle : *Les voleurs ont été pris/appréhendés en flagrant-délit.* La séquence *appréhender une décision* serait comprise comme le fait de *redouter une décision*, ce qui traduit une modification sensible du sens.

14. Cette étude ne figure pas ici en raison de la restriction au français et à un seul patron de POS. La F-mesure alors obtenue de 92 pour la Tâche (évaluée sur `FR-test1.0`) nous avait semblé prometteuse, d'où l'extension de ce système vers VarIDE.

15. `https://mwedemonstrator.atilf.fr/mwetools`

16. `https://universaldependencies.org/format`

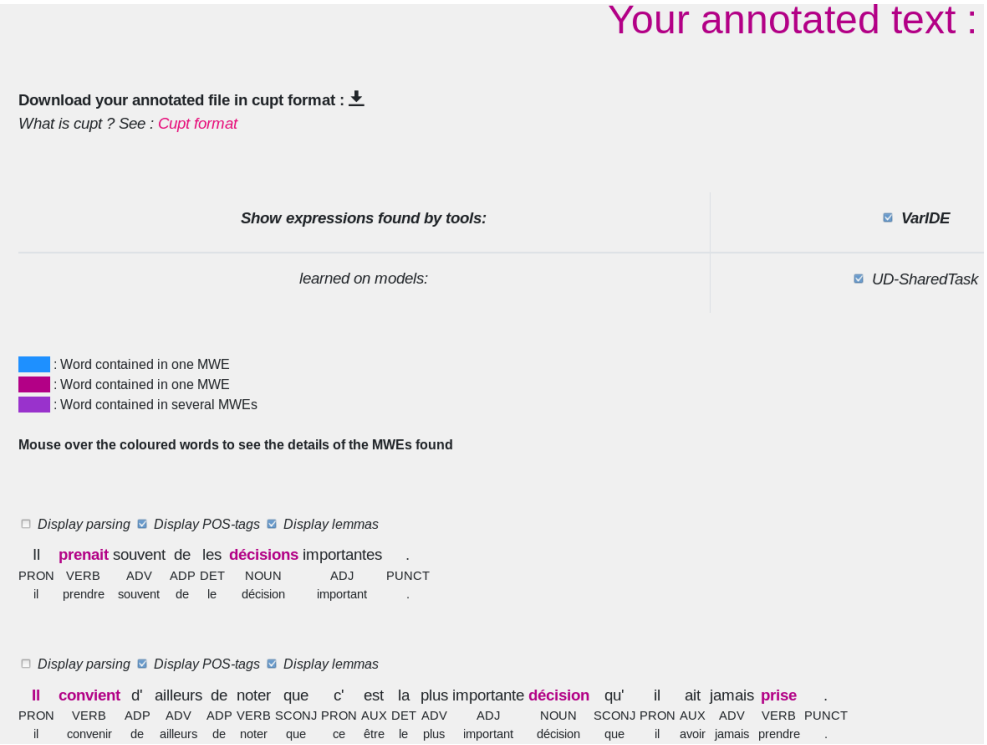


FIGURE 12.2 – Démonstrateur mis au point par le laboratoire ATILF : exemple d'identification par notre système VarIDE. Les éléments lexicalisés des EP ***prendre décision*** et ***il convenir*** sont mis en évidence.

Chapitre 13

VarIDE+ avec sélection automatique de traits : *comment séparer le bon grain de l'ivraie ?*

Dans la classification supervisée, chaque exemple est décrit au moyen d'un ensemble de traits. Tout d'abord, nous proposons dans ce chapitre une nouvelle procédure d'extraction de candidats, plus rapide que celle présentée dans la section 12.2.2, et donc applicable à l'extraction de variantes d'un corpus de grande taille (section 13.1). Ensuite, la contribution principale de ce chapitre se concentre sur la sélection de traits. Dans la classification supervisée, le choix de ces traits est susceptible d'influencer le processus de classification, c'est pourquoi nous cherchons à déterminer les traits pertinents pour la Tâche suivant la méthode présentée dans la section 13.2. Cette sélection sera uniquement opérée sur la partie française du corpus de la ST afin de bénéficier d'une meilleure capacité d'analyse des traits mis en évidence. La performance de VarIDE+ intégrant une sélection de traits est présentée dans la section 13.3 avant d'entamer une discussion sur la pertinence de ces traits (section 13.4).

A la différence de VarIDE qui n'utilisait qu'un classifieur Naïve Bayes pour la classification de candidats, plusieurs méthodes sont exploitées pour VarIDE+ : Naïve Bayes, SVM et arbres de décision. Nous proposons également une méthode de classification s'appuyant sur l'exploitation de la similarité d'EP `SIM` décrite dans la section 11.2.1.

Dans ce chapitre, nous explorons les hypothèses suivantes :

H3a Utilisation d'autres types de corpus : en comparant les performances d'identification de variantes d'EP dans un corpus de `test` de nature similaire à celui de l'entraînement (à l'origine du modèle de classification) et dans un corpus différent, on peut s'attendre à ce que les performances se dégradent pour une utilisation sur un nouveau corpus.

H4 Enrichir un corpus manuellement annoté en EP de taille restreinte par un autre corpus annoté automatiquement en EP de la façon la plus fiable possible devrait permettre d'enrichir le profil de variabilité. Nous nous attendons donc à un gain de performances par rapport à la seule utilisation du corpus d'entraînement initial de taille réduite.

H5 Certains traits sont plus pertinents que d'autres : certains traits peuvent notamment s'avérer plus intéressants selon qu'il s'agit d'EP fréquentes ou rares.

H5a Diminuer le nombre de traits rendrait la classification plus performante et plus facilement interprétable.

13.1 Extraction de candidats (ExtractCands2)

Compte tenu des résultats obtenus durant la ST, et dans l'optique de mener l'extraction de candidats représentant des variantes d'EP déjà vues sur un corpus de très grande taille, la méthode d'extraction de candidats `ExtractCands` a été modifiée par rapport à celle précédemment utilisée pour VarIDE. Notons que l'optimisation de VarIDE (nommée VarIDE+) concerne exclusivement le français.

`ExtractCands2` sera employée pour :

— l'entraînement du classifieur afin de bénéficier d'exemples positifs et négatifs, ensuite utilisés pour la sélection de traits.

— la phase de test, afin de pré-identifier des candidats d'EP qui serviront d'entrée au classifieur.

Pour chaque EP e attestée dans le corpus d'entraînement, `ExtractCands2` extrait des expressions candidates c d'après la co-occurrence des mêmes composants que ceux de e lorsque trois conditions soient satisfaites.

Tout d'abord, l'ensemble des lemmes et POS de c doit correspondre à celui de e qui nous sert de référence. Par exemple, pour e dans les exemples 13.1 (EP = ***avoir lieu***) et 13.2 (EP = ***avoir pouvoir***), les candidats de (Ex. 13.3-Ex. 13.4) seraient extraits mais pas celui de (13.5), en raison de la POS différente de *avoir* (nom au lieu de verbe). La limite de cette règle est illustrée par (13.6) comportant les lemmes *avoir* et *pouvoir* de l'EP ***avoir***$_{\text{VERB}}$ ***pouvoir***$_{\text{NOUN}}$: la *POSnorm* est bien ⟨NOUN ;VERB⟩ mais les POS sont en réalité inversées dans cet exemple puisqu'*avoir* y est un nom et *pouvoir* un verbe. Le système VarIDE décrit dans le chapitre 12 procédait à une vérification systématique des POS de chaque composant, mais nous supposons que de telles ambiguïtés sont suffisamment rares pour autoriser l'approximation reposant sur le respect de la *POSnorm*, ceci dans le but d'augmenter la rapidité d'extraction de candidats par VarIDE+ [1].

(13.1) *Le phénomène* ***eut***$_{\text{VERB}}$ ***lieu***$_{\text{NOUN}}$ *en décembre 1989.* (`FR-train1.1`)

(13.2) *Ils* ***ont***$_{\text{VERB}}$ *le* ***pouvoir***$_{\text{NOUN}}$ *de voter.* (`FR-train1.1`)

(13.3) *La finale n'****aura***$_{\text{VERB}}$ *donc pas* ***lieu***$_{\text{NOUN}}$.

(13.4) *Bruxelles n'****aura***$_{\text{VERB}}$ *pas le* ***pouvoir***$_{\text{NOUN}}$ *de bloquer cette opération.*

(13.5) *Le vendeur propose un* $\underline{\text{avoir}}_{\text{NOUN}}$ *au* $\underline{\text{lieu}}_{\text{NOUN}}$ *d'un remboursement.*

(13.6) *Le commerçant* $\underline{\text{peut}}_{\text{VERB}}$*-il vous imposer un* $\underline{\text{avoir}}_{\text{NOUN}}$ *?*

Par ailleurs, VarIDE+ ne prend plus en compte les séquences de POS autorisées pour une *POSnorm* donnée comme le faisait VarIDE, ni la restriction aux 10 patrons de *POS-*

1. La classification des candidats pour l'ensemble des langues nécessitait environ trois heures.

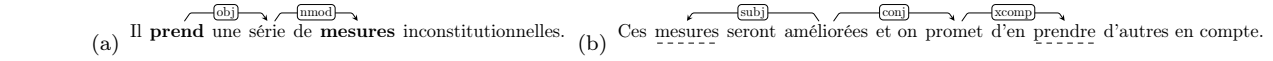


FIGURE 13.1 – Candidats d'EP ayant une chaîne de dépendance discontinue : (a) extrait, (b) non-extrait.

norm les plus fréquents (ceci pour augmenter le rappel de la phase d'extraction de candidats). On suppose que les discontinuités, qui reflètent justement les permutations autorisées, peuvent être avantageusement utilisées.

Deuxièmement, si c contient deux composants ou bien s'il ne comporte qu'un seul verbe et un seul nom, alors nous imposons une distance syntaxique $distSyn \leq 1$. Cette condition est remplie dans l'exemple (13.2) et dans le cas de déterminants complexes, mais généralement pas dans le cas de co-occurrences fortuites comme dans les Fig. 13.1(a vs. b) de la page 179. Aucune contrainte sur les dépendances n'a été formulée pour les candidats ayant plus de deux composants en raison de leur faible fréquence dans le corpus français (moins de 10% des occurrences) et parce qu'une recherche des liens de dépendances directs ou indirects entre tous les composants d'EP – pouvant aller jusqu'à sept – alourdit considérablement le temps d'exécution.

Troisièmement, d'après les discontinuités observées dans les candidats c du `train`, nous restreignons le nombre d'éléments externes insérés entre les composants de c afin d'éviter un grand nombre de candidats incorrects en raison de lemmes identiques (déterminant ou pronom), comme dans (Ex. 13.7).

(13.7) *Il faut rappeler que, jusqu'en 1983, il n'y avait pas ...*

Pour ce faire, comme il est rare qu'une même EP apparaisse plusieurs fois dans la même phrase, nous restreignons la longueur des discontinuités à la seconde [2] plus faible valeur observée dans les données d'entraînement pour l'EP en question. Dans (Ex. 13.8), ces longueurs valent ainsi {0,2,0,1}, la seconde valeur la plus faible est 1 (la première serait ici de zéro). Par conséquent, seuls les candidats (a)(c)(d) ayant une longueur maximale d'une insertion seraient extraits.

(13.8) ***Il faut*** *...* ***il*** *va encore* ***falloir****...* ***faut-il*** *...* ***il*** *ne* ***faut****.*
(a) (b) (c) (d)

Enfin, à l'instar d'`ExtractCands`, `ExtractCands2` n'extrait que les candidats dont l'EP attestée apparaît au moins deux fois dans le `train`, ce qui représente 78% des EP annotées du corpus (3582 tokens).

En résumé, `ExtractCands2` s'appuie sur les conditions suivantes :
— même *LemmNorm* qu'une expression e de référence,
— même *POSnorm* que e,
— si le candidat comporte deux composants : DISTSYN_2COMP ≤ 1 entre les composants,
— si le candidat comporte un seul nom et un seul verbe : DISTSYN_V-N ≤ 1 entre le nom et le verbe. Précisons que cette condition est souvent (mais pas de façon

2. Si une EP n'apparaît qu'une seule fois, c'est la longueur de discontinuités observée sur ce seul exemple qui servira de longueur maximale tolérée.

Corpus	Candidats extraits			P	R	F
	Tous	**Positifs**	**Négatifs**			
`FR-train1.1`	4 596	3 582 (78%)	1 014 (22%)	78	98	87
`FR-test1.1`	368	210 (57%)	158 (43%)	57	100	73
`CoNLL17`	32 789 815	n/a	n/a	n/a	n/a	n/a

TABLE 13.1 – Performance de `ExtractCands2`, en se fondant sur les EP attestées vues au moins deux fois dans `FR-train1.1`. La F-mesure sur le corpus `FR-test1.1` serait alors de 73.

systématique) redondante avec la précédente. Une EP comportant un verbe, un déterminant et un nom satisfera en effet cette condition mais pas la précédente. A l'inverse, toute EP ayant deux composants mais différents d'un nom et d'un verbe ne satisfera pas ce critère.

— longueur de discontinuités limitée d'après la seconde valeur minimale observée pour e

— restriction aux EP vues au moins 2 fois dans le `train`.

Quand `ExtractCands2` est utilisé pour l'entraînement, les candidats extraits sont marqués comme positifs s'ils ont été manuellement annotés comme EP dans le `train`, et négatifs sinon. La Table 13.1 montre les résultats de l'extraction de candidats dans le corpus. Comme attendu, la méthode est optimisée pour un rappel élevé et une précision raisonnable. Ce dernier facteur devrait s'améliorer grâce à la classification, en se fondant sur un ensemble de traits choisi avec soin, ce que nous abordons dans le chapitre 13.2.

13.2 Méthodologie de sélection de traits

13.2.1 Motivation

Le nombre initial de traits absolus et relatifs définis dans le chapitre 12 est extrêmement important : avec le jeu d'étiquettes utilisé dans notre corpus, nous atteignons un total d'environ 18 000 traits générés automatiquement pour le français par combinaison d'informations (p. ex. POS + nombre). Cependant, cette méthode rencontre des écueils. Tout d'abord, certains traits sont en réalité invalides, comme ADV_NOMBRE, les adverbes étant invariables. Un premier filtrage de traits, dénommé `Activefeat` aura pour effet de supprimer ces traits invalides grâce à un échantillon du corpus `CoNLL17`. D'autres traits, pourtant valides, s'avèrent superflus et augmentent inutilement le nombre de traits. Nous avions par exemple défini un trait prenant en compte la présence de POS doublonnées dans une EP, ce qui pouvait servir pour les EP du `train` comportant plusieurs noms, verbes, etc. Le trait généré pour la gestion de deux conjonctions de subordination se révèle en revanche inutile car aucune EP du `train` n'en contient deux. Enfin, un trait peut exister sans pour autant être pertinent. La flexion des verbes nous semble ainsi rarement discriminante car la quasi-totalité des EP y est insensible, contrairement à la flexion personnelle discriminante pour les tournures impersonnelles (***il y a*** vs. *ils y ont*). C'est justement cette recherche de pertinence de traits vis-à-vis de la Tâche qui justifie notre intérêt pour la sélection de traits.

	Traits ABSolus et RELatifs (binaires)	Systèmes de la compétition PARSEME 1.1 (partie FR)										VarIDE	VarIDE+
		(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	Traits initiaux	Traits. sélect. auto.
ABS	Catégorie d'EP (p. ex. LVC)	✓			✓	✓				✓	✓	✓	✓
	Forme de surface de l'EP	✓	✓	✓	✓			✓					
	Ensemble de lemmes de l'EP									✓		✓	✓
	Lemmes individuels de l'EP	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
	POS des composants de l'EP	✓	✓	✓	✓		✓	✓	✓	✓		✓	
	Longueur de l'EP				✓								
	Étiquettes morphosyntaxiques			✓	✓	✓	✓			✓		✓	
	Dépendances syntaxiques de l'EP												
	Connexion	✓				✓				✓		✓	
	Quasi-connexion									✓		✓	
	Labels	✓			✓	✓		✓	✓	✓		✓	
	Discontinuités :												
	Séquences de POS									✓		✓	✓
	Longueur									✓		✓	
REL	Lemmes individuels de l'EP									✓		✓	✓
	POS de l'EP									✓		✓	
	Étiquettes morphosyntaxiques de l'EP									✓		✓	
	Dépendances syntaxiques de l'EP												
	Connexion									✓		✓	
	Quasi-connexion									✓		✓	
	Étiquettes									✓		✓	
	Discontinuités :												
	Séquences de POS									✓		✓	✓
	Longueur									✓		✓	
	Word embeddings externes		✓					✓					
	$F_{seen\text{-}in\text{-}train}$	**81,72**	76,57	76,19	75,06	73,71	72,90	72,86	70,84	70,03	62,24		**82,07**

TABLE 13.2 – Traits pertinents pour l'identification d'EP et performances pour les EP préalablement vues pour notre systèmes vs. les systèmes de la compétition PARSEME. (Ramisch *et al.*, 2018) : (a) (Waszczuk, 2018), (b) (Taslimipoor et Rohanian, 2018), (c) (Boros et Burtica, 2018), (d) (Stodden *et al.*, 2018), (e) (Moreau *et al.*, 2018) (CRF-DepTree-categs system), (f) (Moreau *et al.*, 2018) (CRF-Seq-nocategs system), (g) (Berk *et al.*, 2018), (h) (Ehren *et al.*, 2018), (i) (Pasquer *et al.*, 2018a), (j) (Zampieri *et al.*, 2018)

La plupart des systèmes de la ST utilisent des traits – répertoriés dans la Table 13.2 – réputés être linguistiquement pertinents, et en outre relativement simples à obtenir à partir de corpus annotés (Waszczuk, 2018; Taslimipoor et Rohanian, 2018; Boros et Burtica, 2018; Stodden *et al.*, 2018; Moreau *et al.*, 2018; Berk *et al.*, 2018; Ehren *et al.*, 2018; Pasquer *et al.*, 2018a; Zampieri *et al.*, 2018). Une approche alternative consiste à fournir un ensemble de traits morphosyntaxiques potentiellement pertinents au classifieur et à le laisser attribuer les pondérations adéquates pour chaque trait selon leur pertinence. Cela devrait théoriquement minimiser ou supprimer les traits les moins pertinents (Pasquer *et al.*, 2018a). Si l'on exploite les connaissances du domaine pour définir l'ensemble initial de traits, notre méthode de sélection de traits a pour vocation d'être généralisable à d'autres langues, d'où une procédure automatique répondant à un triple objectif. Tout d'abord, un faible nombre de traits méticuleusement choisis devrait permettre une classification de meilleure qualité, réduire le surapprentissage et le temps d'entraînement. Ensuite, on s'attend à découvrir des traits dont la pertinence vis-à-vis de la Tâche n'est pas établie linguistiquement (comme l'influence de la longueur des discontinuités). Troisièmement, cette procédure automatique est indépendante de la langue, à condition que l'ensemble initial de traits soit suffisamment générique. Quatrièmement, un ensemble réduit de traits devrait favoriser leur interprétabilité.

Après suppression des traits invalides (`ActiveFeat`), le processus de sélection de traits se déroule selon deux phases successives : les traits sont tout d'abord triés par pertinence (phase dite `Rankfeat`), puis une sélection est opérée parmi ceux les mieux classés (phase dite `Selectfeat`).

13.2.2 Suppression de traits invalides : `ActiveFeat`

En appliquant `ExtractCands2` à un échantillon du corpus `CoNLL17`, cela fournit 2,6 millions de candidats représentant des variantes potentielles d'EP du `train`. Leur classe (c.-à-d. 'EP' ou 'non-EP') est inconnue car ce corpus n'est pas annoté en EP. Lors de la définition de l'ensemble initial de traits, des propriétés telles que le genre ou la POS étaient agglomérées, ce qui donnait ADJ_GENRE, NOUN_GENRE, ADV_GENRE, etc. Pour chaque trait absolu t généré, nous extrayons la fréquence d'apparition de chaque valeur v. Le trait ABS_ADJ_GENRE apparaît par exemple x_1 fois associé à v=féminin et x_2 fois à v=masculin. L'analyse des candidats fournit la fréquence de chaque paire $t = v$ sous la forme suivante :

— ABS_ADJ_GENRE = *féminin* (x_1 occ.)
— ABS_ADJ_GENRE = *masculin* (x_2 occ.)
— ABS_NOUN_GENRE = *féminin* (y_1 occ.)
— ABS_NOUN_GENRE = *masculin* (y_2 occ.)
— ABS_ADV_GENRE : 0 occ.
— ...

`ActiveFeat` a pour vocation de supprimer les traits invalides en se fondant sur leur non-activation, autrement dits ceux dont la fréquence est nulle. De nombreux traits, à l'instar de ABS_ADV_GENRE, sont ainsi supprimés, et il ne subsiste alors que 400 traits absolus. Chaque trait absolu se voit alors assorti du trait relatif correspondant, sauf pour deux d'entre eux : ABS_LEMMNORM et ABS_CATEP. Le pendant relatif de ces traits est en

effet inintéressant car, comme les traits relatifs servent à comparer des tokens d'un même type d'EP, leurs valeurs seraient toujours *vrai*. A l'issue d'`ActiveFeat`, nous obtenons donc un ensemble de 798 traits (400 absolus et 398 relatifs). C'est sur ce nouvel ensemble de traits que seront opérés le classement (`RankFeat`) puis la sélection de traits (`SelectFeat`).

13.2.3 Classement des traits : `RankFeat`

Durant la phase `RankFeat`, quatre méthodes de classement de traits ont été évaluées. Toutes, hormis la première, sont appliquées sur les candidats (positifs et négatifs) du corpus `train` car elles tirent parti des informations qui leur sont associées.

— `FREQ` Il s'agit d'une méthode non standard fondée sur la fréquence des couples traits-valeurs activés au moins une fois durant `Activefeat`. Supposons que les couples traits-valeurs activés durant cette étape soient, par valeurs décroissantes de fréquences : $x_1 > y_2 > y_1 > x_1$, ces couples sont alors triés d'après le rang de fréquence le plus élevé pour un trait donné, dans ce cas pour le trait correspondant au genre, on aurait : ADJ_GENRE > NOUN_GENRE. Pour obtenir un classement qui tienne seulement compte des traits et non des valeurs associées, nous conservons le classement en tronquant les couples traits-valeurs de façon à ne plus conserver que les traits. Nous attribuons à cette méthode de troncature de la valeur des paires traits-valeurs le nom de `strip`.

— `CHI2` Le test de χ^2 (Kumbhar et Mali, 2016) est un test statistique qui cherche à déterminer si la distribution d'exemples en classes Y pour un trait donné t relève du hasard ou non. D'après la répartition par classe de candidats décrits par une paire trait-valeur $t = v$, le test de χ^2 compare leurs fréquences observées Obs par rapport à celles qui sont attendues Att en supposant l'indépendance de $t = v$ vis-à-vis de Y :

$$\chi^2 = \frac{(Obs - Att)^2}{Att}$$

Plus l'écart entre les fréquences observées et attendues est élevé pour un trait donné, plus ce trait est discriminant dans le choix de la classe, ce qui se traduit par une valeur χ^2 plus élevée. Nous utilisons la version de Pearson de ce test, et le classement initial trait-valeur correspond à des valeurs décroissantes de χ^2. Le classement de traits définitif résulte du `strip` appliqué au classement de ces paires $t = v$.

— `GAIN` Les traits sont classés par valeur décroissante de gain d'information tel que défini en section 6.1.2.2. Ce gain d'information est calculé d'après la différence entre l'entropie initiale de l'ensemble des observations et celle obtenue après partition selon un trait donné. Cette différence est d'autant plus élevée que le trait est discriminant.

— `FOREST` L'algorithme des forêts d'arbres aléatoires (Breiman, 2001), combine plusieurs arbres de décision (ici 10 arbres sans limite de profondeur) en un unique modèle par un vote majoritaire. Pour construire l'arbre, la qualité de chaque partition est mesurée par l'impureté de Gini. Les traits sont ensuite classés d'après leur pertinence pour le choix des partitions.

13.2.4 Optimisation du choix de traits : SelectFeat + Classif

Chacune des méthodes FREQ, CHI2, GAIN et FOREST classe l'ensemble initial de traits d'après leur pertinence estimée vis-à-vis de la Tâche. Il faut désormais procéder à une sélection des traits les plus pertinents. Dans cette perspective, pour chaque méthode de Rankfeat, nous sélectionnons un sous-ensemble de traits (SelectFeat) et nous évaluons la performance de classification (Classif) d'après ce sous-ensemble. Dans VarIDE+, Classif intègre désormais différents modèles de classification supervisée au lieu d'un seul pour VarIDE (Naïve Bayes) : Naive Bayes, un SVM linéaire et un arbre de décision, d'après leurs implémentations disponibles [3].

Concrètement, pour chaque classifieur $Classif$, et pour chacune des 4 listes de traits classés par pertinence décroissante R [4] résultant de FeatRank, nous procédons de la façon suivante :

— (i) nous sélectionnons les i traits les mieux classés dans R,
— (ii) nous entraînons et évaluons $Classif$, sur la base de ces i traits sélectionnés, dans une configuration de validation croisée non stratifiée (10 partitions) avec une partition du corpus train en 90%-10%, les candidats pour l'entraînement et le test étant extraits par ExtractCands2,
— (iii) nous calculons la F-mesure moyenne F_{moy} sur les 10 partitions,
— (iv) nous répétons les étapes (i–iii) pour chaque valeur i de 1 à $taille(R)$ et nous sélectionnons la valeur de i pour laquelle F_{moy} est la plus haute. En raison de la décroissance rapide et de la stabilisation des performances au-delà de quelques dizaines de traits, il n'est toutefois pas nécessaire de couvrir l'intégralité du spectre des valeurs possibles de i.

13.3 Résultats

Cette section présente les performances du processus RankFeat → SelectFeat → Classif, les corpus utilisés étant FR-train1.1 pour l'entraînement et FR-test1.1 pour l'évaluation. L'ensemble de traits ainsi obtenu est ensuite évalué d'après deux méthodes : par comparaison avec le système le plus performant de la ST et par une évaluation manuelle complémentaire sur un second corpus de test à la fois externe à la ST et comportant 20 fois plus de tokens de variantes que le corpus FR-test1.1.

13.3.1 Ensembles optimum de traits

La Table 13.3 présente le choix optimal de paramètres pour chaque étape du processus parmi l'ensemble des possibilités SelectFeat = {FREQ, CHI2,...}, RankFeat = {i =1,2,...}, Classif = {Naïve Bayes, SVM,...}. L'évaluation des différentes combinaisons SelectFeat+RankFeat mettra en évidence, lors la phase de Classif, plusieurs ensembles

3. NLTK (Loper et Bird, 2002) pour Naive Bayes, et Scikit-learn (Pedregosa *et al.*, 2011) pour le SVM et les arbres de décision.

4. Si une méthode attribue une pertinence nulle à certains traits, alors elle réalise un classement R doublé d'une première sélection de traits.

SelectFeat	RankFeat	Ensemble de traits	Classif	*P*	*R*	*F*	σ
FREQ	4	①	NB	88,1	96,5	92,1	1,2
	6	②	SVM	89,6	98	**93,6**	1,5
			Arbre de décision	91,3	93,4	92,3	0,9
CHI2	8	③	NB	81,8	97,5	88,9	1,1
			SVM	82,6	97,3	89,3	1,3
			Arbre de décision	83,3	96,4	89,4	1,2
GAIN	7	④	NB	82,6	97,4	89,4	1,3
			SVM	89,7	98	**93,6**	1,5
			Arbre de décision	87,5	96,2	91,6	2
FOREST	4	⑤	NB	78,7	97,5	87,1	1,6
			SVM	78,7	97,5	87,1	1,6
			Arbre de décision	78,7	97,5	87,1	1,6

TABLE 13.3 – Meilleure performance moyenne sur 10% de FR-train1.1 (validation croisée) : P, R et F sont les scores moyens obtenus sur les 10 partitions, et σ la variance correspondante. Les ensembles de traits pris en compte – dont le nombre figure dans la colonne RankFeat – sont représentés sous la forme de chiffres entourés d'un cercle dans la colonne SelectFeat. Ils sont détaillés dans la Table 13.4.

de traits. Nous y ferons référence en attribuant à chacun d'entre eux un nombre entouré d'un cercle (le premier ensemble de trait sera donc ①[5]).

P, R et F sont les scores moyens obtenus sur les 10 partitions, et σ la variance correspondante. Notons que ces valeurs ne sont pas comparables avec les résultats de la ST, car obtenus par cross-validation sur le train et non sur le test. La faible valeur de variance (2 au maximum) confirme la validité d'une simple partition du corpus en 90%-10% et le fait qu'aucune stratification n'est nécessaire. Quoique le nombre de candidats positifs/négatifs du corpus FR-train1.1 ne soit pas équilibré (65% de tokens d'EP), cela n'a pas eu d'impact négatif par comparaison avec des corpus plus équilibrés mais, de ce fait, pourvus de moins d'exemples.

Étonnement, la méthode non standard de sélection de traits FREQ, dont la vocation initiale était uniquement d'éliminer les traits invalides de l'ensemble initial de traits, obtient des résultats identiques voire meilleurs que des méthodes standards, avec $F = 93{,}6$ lorsqu'elle est associée à un SVM linéaire tirant parti de 6 traits.

Les ensembles de traits optimum correspondant à ces résultats, décrits dans la Table 13.4, mettent en évidence une certaine cohérence : pour CHI2, GAIN et FOREST, chaque classifieur sélectionne toujours le même ensemble de traits (resp. ③, ④, ⑤). Pour FREQ, les deux ensembles ① et ② varient seulement d'un trait (le lemme du verbe). Toutefois, les ensembles de traits sont assez variés d'une méthode de sélection à l'autre, ce qui prouve la complémentarité de ces méthodes.

Le nombre optimal de traits varie de 4 à 8, ce qui conforte l'hypothèse H_5 sur le fait que certains traits sont plus pertinents que d'autres. Comme attendu, les traits relatifs sont davantage sélectionnés. Or, ils représentent la similarité d'une variante vis-à-vis de réalisations

5. L'ordre dans la numérotation des ensembles de traits est purement arbitraire.

Traits ABSolus et RELatifs	**Ensembles de traits de la Table 13.3** ①	②	③	④	⑤	Total
ABS_LemmNorm (p. ex. ⟨décision ;prendre⟩	✓	✓		✓		3
ABS_catEP (p. ex. LVC)	✓	✓				2
ABS_DiscontSeq (p. ex. DET-ADJ)	✓	✓		✓		**3**
ABS_lemme_VERB (p. ex. '***prendre***')		✓				1
REL_DiscontSeq	✓	✓	✓	✓	✓	**5**
REL_Discont_ADP			✓	✓	✓	3
REL_Discont_CCONJ					✓	1
REL_Discont_DET			✓	✓		2
REL_Discont_NOUN			✓	✓	✓	3
REL_Discont_PUNCT			✓			1
REL_Discont_VERB			✓			1
REL_LongueurDiscont_0à5			✓	✓		2
REL_Lemme_VERB		✓				1
REL_DistSyn			✓			1
Nombre de traits	4	6	8	7	4	

Table 13.4 – Ensembles de traits optimum mentionnés dans la Table 13.3

de référence de la même EP, ce qui implique que leur validité est conditionnée par la présence d'un nombre suffisant de ces tokens de référence. `CHI2` et `FOREST` sélectionnent même exclusivement des traits relatifs. Le trait REL_DiscontSeq est sélectionné par toutes les méthodes, tandis que 6 autres traits ne sont sélectionnés qu'une seule fois. Ceci est cohérent avec notre hypothèse H_1 sur les profils de variabilité. Les traits absolus sont toutefois également pertinents et confortent l'hypothèse H_5, les discontinuités sont par exemple une information intéressante mais dont on ne peut apprécier l'étendue que d'après un grand nombre d'exemples, d'où l'intérêt de transcender les types d'EP. A titre d'exemple, dans le corpus `FR-train1.1`, les 211 tokens d'EP de patron VERB − DET − NOUN couvrent au total 7 patrons de discontinuités [6], pourtant aucun type – même très fréquent comme ***faire l'objet*** qui apparaît 25 fois – n'est associé à plus de 3 patrons.

13.3.2 Évaluation comparative

On peut évaluer indirectement la qualité de l'ensemble de traits sélectionnés en comparant les performances obtenues avec la configuration optimale [7] identifiée précédemment (② + SVM) par rapport aux 15 systèmes de la ST appliqués à la partie française du corpus [8] d'après la chronologie suivante :

— (i) le SVM est entraîné sur l'intégralité (et non pas 90% comme précédemment) des

6. ∅, ADJ, ADP-DET-NOUN, ADP-NOUN, ADV, DET et PRON-ADV-PUNCT-ADP-DET-NOUN-PUNCT
7. D'après la valeur maximale de F-mesure non arrondie.
8. `https://gitlab.com/parseme/sharedtask-data/tree/master/1.1/system-results`

candidats extraits avec `ExtractCands2` du corpus `FR-train1.1`, ces candidats étant décrits au moyen de l'ensemble de traits (2)),
— (ii) ce modèle de classification est appliqué aux candidats extraits avec `ExtractCands2` dans `FR-test1.1`,
— (iii) les EP de `FR-test1.1` omises par `ExtractCands2` ou mal classées sont simplement comptées comme étant des faux négatifs. Cela diminue évidemment la F-mesure par comparaison avec la Table 13.3, où seules les EP vues au moins deux fois sont considérées.

Comme le met en évidence la Table 13.5, la F-mesure pour (2)+SVM diminue effectivement de plus de 11 points dans la configuration de la ST (jusqu'à 82,24) par rapport à l'évaluation sur 10% de `FR-train1.1`. Ce score de VarIDE+ est cependant bien plus élevé que celui obtenu par VarIDE (Pasquer *et al.*, 2018a) ($F = 70,03$), avec une méthode relativement similaire hormis le fait de se restreindre à un clasifieur Naïve Bayes et de n'opérer aucune sélection de traits. Cela souligne à quel point un nombre réduit de traits facilite la classification, ce nombre étant désormais nettement inférieur à la quantité d'exemples (hypothèse H_{5a}). En outre, notre F-mesure est légèrement plus élevée que la meilleure F-mesure de la ST (81,72) par le système TRAVERSAL (Waszczuk, 2018), dans lequel une régression logistique était appliquée à des séquences de nœuds dans des arbres de dépendances. De plus, la gestion des EP de type *variant-of-train* (c.-à-d. les EP de `FR-test1.1` apparaissant sous une forme de surface différente de `FR-train1.1`) est meilleure avec notre méthode qu'avec TRAVERSAL ($F = 73,47$ vs. 71,23) et bien plus élevée qu'avec VarIDE (57,22). On obtient également de meilleures F-mesures pour les *seen-in-train* que le deuxième et le troisième meilleurs systèmes, fondés sur des réseaux de neurones : GBD-NER-resplit (Boros et Burtica, 2018), qui utilise uniquement les corpus fournis durant la ST, et SHOMA (Taslimipoor et Rohanian, 2018), qui fait également appel à des *word embeddings* de *Wikipedia*.

Lorsque la configuration (2)+SVM est évaluée sur l'intégralité des EP du corpus `FR-test1.1`, et pas uniquement sur les vues, la F-mesure décroît à 55, ce qui est toujours relativement proche de la valeur de 56 obtenue par le meilleur système (TRAVERSAL), bien que les EP non-vues soient totalement en dehors de notre champ d'investigation.

13.3.3 Évaluation manuelle sur un corpus externe

13.3.3.1 Classification par un SVM linéaire

Il est intéressant d'évaluer la qualité des traits sélectionnés sur un corpus externe, de nature différente de `FR-train1.1` utilisé pour la sélection des traits. C'est cet objectif d'estimer la capacité de généralisation de notre système qui nous a conduit à constituer le corpus `WebSample`. Nous appliquons à ce corpus, désormais considéré comme corpus de `test`, la configuration optimale de choix des traits et de classifieur précédemment identifiée, c'est-à-dire : (2)+SVM. Comme montré dans la Table 13.6, la F-mesure globale de 92,2 est comparable avec les 93,6 obtenus sur 10% du `FR-train1.1`. Cela met en évidence la bonne reproductibilité de la méthode quelque soit le corpus : notre système ne souffre ni de surapprentissage, ni de sensibilité aux données source. En effet, `WebSample` provient à moitié d'exploration du Web, donc de données de moindre qualité textuelle, et qui sont de plus

Évaluation sur 10% de `FR-train1.1` (EP vues deux fois)			
Méthode	P	R	F
VarIDE+ = Ensemble de traits (2) + SVM	89,65	98,05	**93,64**
Évaluation sur `FR-test1.1` (toutes les EP vues)			
VarIDE+ = Ensemble de traits (2) + SVM	82,40	82,07	**82,24**
TRAVERSAL	88,79	75,70	81,72
GBD-NER-resplit	88,42	66,93	76,19
SHOMA	91,67	65,74	76,57
VarIDE	56,82	91,24	70,03
Évaluation sur `WebSample`			
VarIDE+ = Ensemble de traits (2) + SVM	87,32	97,70	**92,22**

TABLE 13.5 – Résultats d'évaluation dans la configuration d'évaluation de la compétition PARSEME, et avec un corpus externe manuellement annoté

Catégorie d'EP	P	R	F
$\text{VID}_{1.1}$	91,7	96,1	93,9
$\text{IRV}_{1.1}$	87,3	98,5	92,6
$\text{LVC}_{1.1}$	83	99	90,3
$\text{MVC}_{1.1}$	75	75	75
Total	87,3	97,7	92,2

TABLE 13.6 – Résultats de VarIDE+ par catégorie sur `WebSample` (EP vues au moins deux fois)

automatiquement annotées en morphosyntaxe, donc là encore sujettes à davantage d'erreurs qu'avec l'annotation manuelle du corpus français de PARSEME. Quant aux résultats par catégorie, en écartant les $\text{MVC}_{1.1}$ (dont le nombre est négligeable), nous remarquons que les $\text{LVC}_{1.1}$ sont les plus difficiles à classer ($F = 90,3$) contre 93,9 pour les $\text{VID}_{1.1}$, probablement en raison de la forte variabilité de ces premiers, comme illustré précédemment (page 49).

13.3.3.2 Classification par mesure de similarité

La classification par mesure de similarité se déroule comme exposé dans la section 11.2.1. Pour cette mesure de similarité, nous avons fait appel à la mesure de Jaccard pondérée [9] qui représente la similarité entre deux ensembles A et B de la façon suivante :

$Jaccard_{\text{pondéré}(A,B)} = \frac{\sum_{i \in A \cap B} w_i}{\sum_{i \in A \cup B} w_i}$, où w_i est le poids associé à chaque couple trait-valeur i.

Le corpus d'entraînement est constitué de l'intégralité du `FR-train1.1`. Le corpus `FR-dev1.1` n'est pas utilisé comme un corpus de développement standard mais comme un

9. D'autres mesures de similarité ont été testées, mais ensuite écartées en raison de performances similaires voire inférieures : Dice, Sneath&Sokal, Anderberg, Czekanowski, 3wJaccard, Nei&Li, Driver&Kroeber, Sorensen, Tanimoto (Choi *et al.*, 2010)

corpus externe afin d'optimiser le choix du seuil de similarité au-delà duquel nous considérons qu'il s'agit d'une EP. Enfin, `WebSample` est ici utilisé comme corpus de `test`. Contrairement à nos attentes, l'utilisation de cette mesure n'est pas concluante, la F-mesure sur `WebSample` étant plus faible que celle obtenue avec un SVM (F = 87,2). Pour l'expérience suivante d'enrichissement du corpus `FR-train1.1` par le corpus `CoNLL17` automatiquement annoté en EP (technique de *bootstrap*), nous ne retenons donc que la classification par un SVM.

13.3.4 *Bootstrap*

L'une des limites de notre corpus d'entraînement `FR-train1.1` est sa taille réduite qui ne permet pas de bénéficier d'un large éventail de réalisations de chaque EP, ce qui implique des profils de variabilité potentiellement incomplets. A titre d'exemple, 37% des types d'EP sont illustrés par seulement deux tokens et 74% par moins de sept exemples. Or, compte-tenu des performances satisfaisantes d'identification de variantes obtenues sur le corpus `WebSample`, il semble judicieux d'annoter automatiquement en EP un vaste corpus (`CoNLL17`) pour enrichir le corpus d'entraînement initial `FR-train1.1`. L'hypothèse sous-jacente H_4 est que cela offrirait une vision plus juste de la variabilité de chaque EP et que cela pourrait de ce fait améliorer les performances sur le `FR-test1.1`.

Nous avons extrait du corpus `CoNLL17` un maximum de 500 000 candidats par EP annotée deux fois dans le corpus `FR-train1.1`. Nous avons fixé cette limite pour éviter la sur-représentation de quelques EP, l'une d'entre elles fournissant 3,8 millions de candidats. Le fait d'imposer cette limite n'affecte en réalité que les 10 types les plus fréquents. De cette façon, 583 types d'EP sont représentés par un nombre d'occurrences allant de 8 à 500 000. Seuls 8 types sont illustrés par moins de 100 exemples. Les candidats ont ensuite été étiquetés comme étant des '(non-)EP' par VarIDE+ (configuration (2) + SVM). Seuls les 26,5 millions de candidats classés comme étant des 'EP' ont été ajoutés au corpus d'entraînement `FR-train1.1` afin de l'enrichir et de servir de nouvelles données d'entraînement à VarIDE+. L'évaluation a ensuite été effectuée sur le corpus `FR-test1.1` dans les conditions de la ST.

Cette expérience n'a toutefois pas eu le résultat escompté, les performances étant dégradées par rapport à la seule utilisation d'un corpus réduit : la F-mesure diminue en effet de 13 points (69 vs. 82). Cela semble donc aller à l'encontre de l'idée répandue selon laquelle il est préférable de bénéficier de données volumineuses, bien qu'une analyse d'erreurs future soit requise pour confirmer ce résultat.

13.4 Bilan

En conclusion, les contributions principales du système intégrant la sélection de traits sont les suivantes :

— Un niveau de performances satisfaisant, comparable pour le français avec le meilleur système de la ST,

— Son interprétabilité et sa performance de classification en raison du nombre limité de traits (6),

— Sa capacité de généralisation, confirmée par l'évaluation sur un corpus à la fois externe et plus large,
— Sa capacité à mettre en lumière les propriétés linguistiques des EP par une combinaison de traits inédite.

On peut s'interroger sur la pertinence des traits sélectionnés vis-à-vis du phénomène de variabilité. Tout d'abord, comme souligné dans la Table 13.4, la majorité des traits sélectionnés par les différentes méthodes repose sur les discontinuités, telles que DET-ADJ dans (Ex. 13.11), qui se situent de plus toujours en tête des classements de traits (en tant que trait REL et/ou ABS), alors que ce trait n'est généralement pas pris en compte dans d'autres travaux. Le trait REL_LONGUEURDISCONT_0À5, considéré comme pertinent par les méthodes `GAIN` et `FOREST` subsume les insertions individuelles en spécifiant une comparaison sur la longueur des discontinuités. Plusieurs plages étaient définies (0-5, 6-10, 11-15, supérieures à 15), or seule la première est discriminante, ce qui renvoie à la tendance observée sur les représentations graphiques des profils de variabilité (p. 139-140), les EP tendant à avoir – pour les 2 EP françaises décrites – en moyenne des discontinuités de longueur égale à 2 contre 10 pour les non-EP.

Les discontinuités peuvent, en effet, modéliser plusieurs phénomènes pertinents tels que la variabilité du déterminant dans (Ex. 13.9) vs. (Ex. 13.10), des modifieurs autorisés comme l'adjectif dans (Ex. 13.11), la passivation (Ex. 13.12) ou la relativation (Ex. 13.13). Les discontinuités incluant un verbe, un nom ou un signe de ponctuation apparaissent comme étant les plus discriminantes, peut-être car elles peuvent suggérer des non-EP comme dans *Je jette l'eau*$_{\text{NOUN}}$, *tu nettoies l'*$_{\text{VERB}}$ *éponge*.

(13.9) *Je* ***prends*** *des*$_{\text{DET}}$ ***décisions***

(13.10) *Je* ***prends*** *deux*$_{\text{NUM}}$ ***décisions***

(13.11) *Je* ***prends*** *une*$_{\text{DET}}$ *grande*$_{\text{ADJ}}$ ***décision***

(13.12) *Ma* ***décision*** *est*$_{\text{AUX}}$ ***prise***

(13.13) *C'est la* ***décision*** *que*$_{\text{PRON}}$ *je*$_{\text{PRON}}$ ***prends***

On observe par ailleurs que la moitié des 6 traits sélectionnés sont de nature lexicale : ils renvoient aux lemmes de l'EP. Le trait absolu portant sur le lemme du verbe apparaît comme pertinent car plusieurs EP partagent un même composant (p. ex. ***adresser refus / remerciement / reproche***). Sans les traits lexicaux, la F-mesure diminuerait de 4 points. Cela confirme la force du phénomène d'inflexibilité lexicale et la difficulté de généralisation vers des données non-vues, comme ce que l'on a observé dans la ST [10].

Certains traits, considérés comme pertinents pour la variabilité des EP, n'apparaissent cependant pas dans la Table 13.4. Plusieurs EP interdisent par exemple la modification du nom (*jeter l'éponge verte*) ou sa flexion en nombre (*jeter les éponges*). Étant donné que de telles variations sont souvent tolérées par les $\text{LVC}_{1.1}$, mais pas par les $\text{VID}_{1.1}$, cela laisse supposer que les candidats $\text{VID}_{1.1}$ ne devaient pas être assez nombreux pour acquérir cette connaissance sur les contraintes de $\text{VID}_{1.1}$. Il faudrait dans ce cas envisager un classifieur par catégorie d'EP pour prendre en compte leurs spécificités. Ceci représente l'une des perspectives que nous entrevoyons au terme de cette thèse pour améliorer notre système

10. La meilleure performance pour des données non vues est inférieure à $F = 29$, même pour des systèmes utilisant des réseaux de neurones et des *word embeddings* (Ramisch *et al.*, 2018).

d'identification de variantes d'EP, perspectives sur lesquelles nous revenons dans le chapitre 14.

Chapitre 14

Perspectives

Concernant les perspectives, une analyse détaillée des traits pourrait être menée en prenant en compte les poids attribués par le classifieur aux différents couples trait=valeur. On s'aperçoit par exemple que deux $\text{IRV}_{1.1}$ (***s'emparer, s'avérer***) ont des poids similaires dans le modèle de classification (respectivement 7,1 et 6,8 pour le trait ABS_LEMMNORM). Or ces deux EP ont en commun d'inclure des cranberry words (**emparer, *avérer*). De façon générale, une analyse d'erreurs approfondie devrait être conduite afin de mieux comprendre l'influence de chaque trait vis-à-vis du phénomène de variabilité, voire d'en intégrer de nouveaux. En effet, quoique la prépondérance des traits lexicaux dans le classement des traits pertinents souligne la difficulté de généraliser sur des données non vues, nous supposons que des traits complémentaires s'appuyant par exemple sur la sémantique distributionnelle pourraient être utiles pour considérer comme vues des données non vues en réalité. Connaissant par exemple les EP ***il pleuvoir***, ***mener enquête*** ou ***lancer appel*** dans `FR-train1.1`, on pourrait ainsi obtenir des EP sémantiquement proches (***il neiger*** / ***bruiner*** / etc.), voire synonymes (***conduire étude***) ou antonymes (***recevoir appel***). Ce couplage de l'identification et de la découverte d'EP permettrait d'enrichir le corpus d'entraînement manuellement annoté. Nous pourrions également évaluer la généricité de notre méthode sur d'autres langues de la ST. Il se peut que la sélection de traits par langue ou famille de langues révèle des caractéristiques universelles des EP.

Par ailleurs, nos premiers essais de classification par catégorie d'EP ne s'étaient pas avérés concluants mais cela provient peut-être de l'hétérogénéité de la catégorie des $\text{VID}_{1.1}$, qui se distinguent sur ce point des $\text{LVC}_{1.1}$ et des $\text{IRV}_{1.1}$. Or, il est probable que les traits pertinents soient dépendants des catégories d'EP en raison de leurs propriétés syntaxiques. Le patron syntaxique des $\text{MVC}_{1.1}$ (VERB − VERB) va ainsi être associé à des séquences d'insertions distinctes des $\text{IRV}_{1.1}$ (PRON − VERB) : ***s'est***$_{\text{AUX}}$ ***emparé*** mais a_{AUX} ***laissé tomber***. Les $\text{MVC}_{1.1}$, $\text{IRV}_{1.1}$ et $\text{LVC}_{1.1}$ constituent des classes relativement homogènes : toutes les MVC sont de patron VERB − VERB, la quasi-totalité des $\text{LVC}_{1.1}$ est de patron VERB − (ADP) − (DET) − NOUN (97,5% des cas), et les $\text{IRV}_{1.1}$ de patron PRON − VERB (99,8% des cas). Les $\text{VID}_{1.1}$ en revanche sont bien plus hétérogènes, il faudrait à notre avis scinder cette catégorie en trois sous-catégories d'après leurs patrons syntaxiques : VERB − (ADP) − (DET) − NOUN (56,9%), impersonnels de patron PRON − (PRON) − VERB

(29,8%) et autres patrons (13,3%) [1]. De plus, la distance syntaxique ne couvre pas de façon homogène l'ensemble des catégories d'EP : en effet, elle n'a jamais été calculée pour les EP comportant plus de 2 composants ou plusieurs noms ou verbes, ce qui reste rare pour les $IRV_{1.1}$ et $LVC_{1.1}$ (respectivement 0,2% et 2,1%), mais pas pour les $VID_{1.1}$ (24%). Cette catégorie d'EP mérite donc qu'on lui consacre un traitement spécifique.

Une autre perspective de travail consisterait à trouver une stratégie pour les EP vues une seule fois, ici écartées des EP vues, ce qui diminue de ce fait le vivier d'EP connues et limite le rappel lors de l'identification de leurs variantes.

1. Dont le plus fréquent, PRON-VERB-NOUN comme dans ***il est question***, ne couvre que 12% des cas. D'autres patrons observés sont : VERB-ADP-VERB (p. ex. ***trouver à redire*** ⇒ 'trouver des défauts'), VERB-DET-NOUN-ADP-DET-NOUN (p. ex. ***donner sa langue au chat*** ⇒ 'renoncer à deviner la solution').

Conclusion

Notre question de recherche était la suivante : *La variabilité restreinte des EP verbales favorise-t-elle leur identification dans un texte donné* ? Pour cela, nous avons cherché à décrire un ensemble d'EP connues *via* des traits portant sur des informations d'ordre morphologique ou syntaxique, ainsi que sur les parties de discours des éléments éventuellement insérés (p. ex. un déterminant et un adjectif dans ***prendre*** *une*$_{\text{DET}}$ *importante*$_{\text{ADJ}}$ ***décision***) afin qu'une méthode de classification puisse ensuite déterminer si un nouvel exemple, d'après la valeur des traits qui lui sont associés, est – ou non – une EP. Cette méthodologie implique que cette thèse se focalise exclusivement sur des EP déjà vues, et que par conséquent le défi majeur est une prise en compte adéquate de la variabilité auxquelles sont soumises les EP verbales. En effet, comme exposé dans la partie I, cette variabilité plus ou moins importante (p. ex. *je* ***prends*** *une* ***décision*** vs. *les* ***décisions prises***) les distingue des EP non verbales et pose problème aux applications de TAL, notamment en raison de l'ambiguïté de séquences, en apparence identiques, mais pouvant relever de lectures idiomatiques, de lectures littérales ou bien encore de co-occurrences fortuites. La fréquence des EP rend par ailleurs cette distinction cruciale pour garantir la qualité de tâches applicatives telles que la traduction automatique.

Dans la partie II, nous nous sommes intéressée aux données et outils disponibles pour le recensement des EP, tout en montrant leurs limites en terme de couverture d'EP. Les tâches de découverte – pour combler les lacunes des lexiques existants –, d'identification – pour favoriser des tâches ultérieures comme la traduction automatique – et plus précisément d'identification de variantes ont été décrites du point de vue de l'exploitation de la variabilité des EP car certaines approches (comme la prise en compte de blocages morpho-syntaxiques) nous ont paru transférables vers la tâche d'identification de variantes d'EP. Notre choix de focalisation sur cette tâche particulière se justifie par le fait que l'optimisation de l'identification d'expressions connues représente un premier pas pour l'amélioration de l'identification des EP en général. Nous avons pour cela tiré profit des données fournies dans le cadre des campagnes d'annotation PARSEME.

Nos contributions portent en premier lieu sur une description fine de la variabilité des EP verbales en français (partie III), avec analyse statistique en corpus des différents modes de variabilité selon les catégories d'EP (Pasquer, 2017), ce qui nous a permis de mettre au jour des traits susceptibles d'être pertinents pour la mise en œuvre de la classification. Si certains de ces traits sont fréquemment cités dans l'état de l'art, d'autres le sont rarement, comme la typologie des discontinuités, souvent restreinte à la variabilité du déterminant, alors que l'insertion d'un autre verbe (non modal) par exemple peut s'avérer particulièrement discriminante. De plus, ce recours aux POS des discontinuités permettrait de se

contenter d'un étiquetage en partie de discours, plus facilement disponible qu'un étiquetage en relations de dépendances syntaxiques. Certains modes de variabilité apparaissent spécifiques à certaines catégories d'EP, mais des EP d'une même catégorie peuvent avoir des comportements parfois très différents, ce qui nous a conduit à considérer que chaque type d'EP bénéficiait d'un *profil de variabilité* (hypothèse H_1). Notre capacité à mettre en lumière des profils distincts par catégorie d'EP et par type d'ambiguïté (hypothèse H_2) tend à conforter H_1. C'est cette notion de profil de variabilité qui sous-tend notre méthode d'identification : nous avons émis l'hypothèse que plus un nouvel exemple ressemble à un emploi attesté de la même EP (propriétés dites *relatives*) ou d'une autre EP (propriétés *absolues*), plus il est susceptible d'être à son tour une EP.

Nos autres contributions portent sur la validation d'une mesure de variabilité des EP de patron VERBE-(DET)-NOM correspondant à deux catégories majeures d'EP (constructions à verbe support et idiomes) (Pasquer *et al.*, 2018b) et sur une méthode d'identification de variantes d'EP possédant ce patron grâce à l'extraction de candidats décrits avec des traits absolus et relatifs suivie d'une classification supervisée (Pasquer *et al.*, 2018c). Ayant obtenu un niveau de performance d'extraction supérieur à la méthode de Savary et Cordeiro (2017) nous ayant servi de baseline, nous avons généralisé ce système à des EP de patrons syntaxiques variés et dans différentes langues, ce qui s'est concrétisé sous la forme d'un système d'identification de variantes d'EP présenté à la compétition PARSEME 1.1 en 2018 (Pasquer *et al.*, 2018a).

Ce système, nommé VarIDE et décrit en partie IV, s'est alors classé en première place pour le bulgare et en deuxième position pour le français (F = 70,03), le portugais et le slovène. Ces résultats démontrent sa capacité de généralisation multilingue et corroborent l'hypothèse H_3 sur l'utilité des profils de variabilité pour l'identification de variantes d'EP. Notre système fonctionnait selon deux étapes (extraction de candidats d'après des motifs de variabilité de patrons syntaxiques puis classification pour écarter les cas non pertinents), mais on s'aperçoit qu'en menant une approche différenciée selon les langues (en se contentant de la seule phase d'extraction de candidats pour certaines langues), cela conduirait à des performances bien meilleures, avec par exemple une F-mesure de 95 au lieu de 20 pour les EP vues en hongrois. Ce système, en plus d'être librement accessible sur Gitlab [2], bénéficie également d'une vitrine grâce au démonstrateur mis en ligne par le laboratoire ATILF [3].

Nous avons également présenté une optimisation de VarIDE, nommée VarIDE+, obtenant un score F = 82,24 sur les expressions vues pour le français, soit 12 points supérieur au système initial. Cette optimisation repose sur une extraction plus ciblée et surtout sur une sélection de traits qui permet non seulement d'accélérer la tâche d'identification de variantes et rend notre modèle intelligible, mais également de mettre en évidence le fait que certains traits sont davantage pertinents que d'autres. Ici encore, les séquences de parties de discours des discontinuités jouent un rôle crucial, quoique ce type de traits ne figure pas parmi ceux qui sont traditionnellement pris en compte pour l'identification d'EP. On s'aperçoit par ailleurs que les traits relatifs prédominent parmi les traits pertinents lors de la construction du modèle de classification, en accord avec notre hypothèse sur l'existence

2. `https://gitlab.com/cpasquer/SharedTask2018_varIDE`
3. `https://mwedemonstrator.atilf.fr/mwetools`

de profils de variabilité des EP verbales. La capacité d'adaptation de VarIDE+ à des données de nature très différente de celles ayant servi à construire le modèle de classification a également été validée, et une évaluation sur d'autres langues pourrait être menée pour déterminer si la sélection de traits opérée se révèle aussi avantageuse.

Au terme de cette thèse, la réponse à la question de recherche doit être nuancée en fonction des différentes catégories d'EP. Comme le montrent les résultats de classification obtenus (partie IV), les catégories les moins variables (verbes intrinsèquement réflexifs ou IRV) tendent à être mieux prises en compte que les autres (notamment les constructions à verbe support ou LVC), ce qui souligne l'intérêt de poursuivre les travaux sur la variabilité des EP verbales. On note par ailleurs que la catégorie des idiomes verbaux (VID) recouvre des EP de patrons syntaxiques très variés (VERB-(DET)-NOUN) à l'instar des LVC, mais également des structures impersonnelles, etc.). Il nous semblerait pertinent, au vu des résultats obtenus, de tenir compte de cette diversité des patrons en scindant la catégorie VID en sous-catégories de façon à favoriser la tâche de classification automatique. Les configurations impersonnelles ont en effet des profils particulièrement restreints (flexion à la 3$^{\text{ème}}$ personne, impossibilité de relative, etc.).

Les perspectives à court terme consistent à élargir notre champ d'investigation des EP vues vers celles non vues au préalable, mais qui partagent certaines similarités sémantiques avec celles-ci, en ayant recours aux *word embeddings* bien que l'on s'attende à une portée limitée de l'apport, celui-ci pourrait peut-être s'avérer plus ou moins important selon les catégories d'EP. La découverte de constructions à verbe support nous semble davantage prometteuse : beaucoup constituent en effet des paraphrases susceptibles d'apparaître dans des contextes similaires à celui du verbe dérivé du nom qui les composent (p. ex. ***venir en aide*** = *aider*, ***prendre une décision*** = *décider*, etc.). Une autre perspective intéressante serait la prise en compte des nominalisations (p. ex. ***prise de décision***). La capacité de nominaliser des EP verbales pourrait d'ailleurs être considérée comme un trait supplémentaire de variabilité (cela peut en effet signaler des lectures littérales comme dans *la prise de taureau par les cornes*). On pourrait aussi s'inspirer de la retokénization pour modéliser certains phénomènes tels que les déterminants complexes ***prendre*** DET_NOUN_ADP ***décisions***, de façon à limiter la dispersion des traits : si une EP tolère un déterminant, on s'attend en effet à ce qu'elle accepte aussi un déterminant complexe. Certains cas d'ambiguïté en français gagneraient également à être traités de façon spécifique : c'est le cas pour ***il y a*** pouvant être une EP verbale ou non (***il y a*** trois ans). Cette EP représentait 7% de l'ensemble des occurrences d'EP dans le corpus de `test` proposé pour l'identification d'EP, or 18% d'entre elles étaient de nature adverbiale et non verbale. On pourrait par exemple intégrer une règle exploitant la présence de marqueurs numériques et temporels (***il y a*** NUM DUREE) ou bien, là encore, tirer profit des *word embeddings*.

A plus long terme, une réflexion plus approfondie sur la polylexicalité pourrait être menée, en intégrant par exemple une dimension multilingue fondée sur le fait que certaines EP bénéficient d'équivalents dans d'autres langues (`(FR)` ***faire référence*** = `(ES)` ***hacer referencia***). On pourrait peut-être ainsi mettre en évidence des caractéristiques 'universelles' de variabilité. Enfin, les mécanismes de (dé)figement d'après une perspective diachronique pourraient également être intéressants, notamment pour savoir s'ils s'opèrent à un rythme spécifique et dans un ordre particulier.

Annexe

Annexe A

Liste d'EP de `WebSample`

Fréquence dans `FR-train1.1`	$IRV_{1.1}$	$LVC_{1.1}$	$VID_{1.1}$
Élevée	*s'étendre*	*faire apparition*	*il s'agir*
	se voir	*avoir droit*	*il convenir*
	se trouver	*jouer match*	*mettre fin*
	se retrouver	*poser question*	*avoir lieu*
	s'élever	*faire appel*	*il y avoir*
	se situer	*avoir besoin*	*faire partie*
	s'engager	*rendre hommage*	*il falloir*
	se produire	*avoir chance*	*tenir compte*
	se dérouler	*jouer rôle*	*faire l'objet*
	se rendre		*prendre part*
Médiane	*s'efforcer*	*conclure accord*	*prendre au piège*
	se servir	*mener opération*	*en finir*
	se baser	*porter choix*	*avoir affaire*
	s'agir	*rendre service*	*venir à bout*
	s'envoler	*disputer épreuve*	*ne pas payer de mine*
	se lever	*avoir perception*	*vouloir dire*
	s'interroger	*livrer bataille*	*s'en aller*
	s'adonner	*présenter signe*	*faire écho*
	s'élancer	*donner concert*	*mettre sur pied*
			mettre en lumière
Basse	*s'acquitter*	*faire traduction*	*tourner mal*
	s'engouffrer	*avoir ennui*	*couper du monde*
	se ressentir	*prendre sanction*	*être au rendez-vous*
	se donner	*faire entrée*	*tenir responsable*
	se partager	*inscrire but*	*faire les frais*
	s'empresser	*effectuer tournée*	*être l'occasion*
	se renseigner	*dresser horoscope*	*mettre sur la table*
	se signaler	*signer victoire*	*prendre le dessus*
	s'investir	*réserver accueil*	*tel être le cas*
	se promener		*faire le plein*

TABLE A.1 – Types d'EP des catégories $IRV_{1.1}$, $LVC_{1.1}$ et $VID_{1.1}$ utilisés pour constituer le corpus `WebSample`.

$MVC_{1.1}$
entendre parler
faire remarquer
faire savoir
laisser tomber

TABLE A.2 – Types d'EP de la catégorie $MVC_{1.1}$ utilisés pour constituer le corpus `WebSample`.

Bibliographie

Abeillé, A. et Clément, L. (2003). Annotation morpho-syntaxique : Les mots simples-les mots composés, corpus Le Monde.

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for French. *In Treebanks*, pages 165–187. Springer.

Abraham, R. D. (1950). Fixed order of coordinates : A study in comparative lexicography. *The Modern Language Journal*, 34(4):276–287.

Al Saied, H., Candito, M. et Constant, M. (2017). The ATILF-LLF system for parseme shared task : A transition-based verbal multiword expression tagger.

Aronoff, M. (1976). *Word Formation in Generative Grammar*. The MIT Press.

Ashraf, N. et Ahmad, M. (2015). Machine translation techniques and their comparative study. *International Journal of Computer Applications*, 125(7).

Augenstein, I., Derczynski, L. et Bontcheva, K. (2017). Generalisation in named entity recognition. *Comput. Speech Lang.*, 44(C):61–83.

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.

Baldwin, T. et Kim, S. N. (2010). Multiword expressions. *In* Indurkhya, N. et Damerau, F. J., éditeurs : *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 édition.

Banerjee, S. et Pedersen, T. (2003). The Design, Implementation, and Use of the Ngram Statistics Package. *In Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'03, pages 370–381, Berlin, Heidelberg. Springer-Verlag.

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. *In Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 1–8. Association for Computational Linguistics.

Bauer, L. et Laurie, B. (1983). *English word-formation*. Cambridge University Press.

Berk, G., Erden, B. et Güngör, T. (2018). Deep-BGT at PARSEME Shared Task 2018 : Bidirectional LSTM-CRF model for verbal multiword expression identification. *In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 248–253.

Blanco, X. et Bogacki, K. (2014). *Introduction à l'histoire de la langue française*. Documents. Universidat Autònoma de Barcelona, Servei de Publicacions.

Boros, T. et Burtica, R. (2018). GBD-NER at PARSEME Shared Task 2018 : Multiword expression detection using bidirectional long-short-term memory networks and graph-based decoding. *In* Savary, A., Ramisch, C., Hwang, J. D., Schneider, N., Andresen, M., Pradhan, S. et Petruck, M. R. L., éditeurs : *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions, LAW-MWE-CxG@COLING 2018, Santa Fe, New Mexico, USA, August 25-26, 2018*, pages 254–260. Association for Computational Linguistics.

Boukobza, R. et Rappoport, A. (2009). Multi-word expression identification using sentence surface features. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2*, pages 468–477. Association for Computational Linguistics.

Breidt, E., Segond, F. et Valetto, G. (1996). Formal description of multi-word lexemes with the finite-state formalism idarex. *In Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 1036–1040. Association for Computational Linguistics.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R. et Stone, C. (1984). Classification and regression trees.

Candito, M., Constant, M., Ramisch, C., Savary, A., Parmentier, Y., Pasquer, C. et Antoine, J.-Y. (2017). Annotation d'expressions polylexicales verbales en français. *In TALN 2017*, Actes de TALN 2017, Orléans, France. Association pour le Traitement Automatique des Langues.

Candito, M. et Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (The Sequoia Corpus : Syntactic Annotation and Use for a Parser Lexical Domain Adaptation Method)[in French]. *In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, pages 321–334.

Cap, F., Fraser, A., Weller, M. et Cahill, A. (2014). How to produce unseen teddy bears : Improved morphological processing of compounds in SMT. *In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 579–587.

Carpuat, M. et Diab, M. (2010). Task-based evaluation of multiword expressions : a pilot study in statistical machine translation. *In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245.

Caseli, H. M., Ramisch, C., das Graças Volpe Nunes, M. et Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1):59–77.

Chaulet, G. (1975). *Appelez Fantômette*. Hachette jeunesse.

Choi, S.-S., Cha, S.-H. et Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. *In RIAO 88 (Recherche d'Information Assistée par Ordinateur). Conference*, pages 609–623.

Church, K. W. et Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M. et Todirascu, A. (2017). Multiword expression processing : A survey. *Computational Linguistics*, 43(4):837–892.

Constant, M. et Nivre, J. (2016). A transition-based system for joint lexical and syntactic analysis. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 161–171, Berlin.

Constant, M., Sigogne, A. et Watrin, P. (2012). La reconnaissance des mots composés à l'épreuve de l'analyse syntaxique et vice-versa : évaluation de deux stratégies discriminantes. *In Conférence sur le Traitement Automatique des Langues Naturelles*, pages 57–70.

Constant, M. et Tellier, I. (2012). Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger. *In 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 646–650.

Cordeiro, S., Ramisch, C. et Villavicencio, A. (2016a). mwetoolkit+ sem : Integrating Word Embeddings in the mwetoolkit for Semantic MWE Processing. *In LREC*.

Cordeiro, S., Ramisch, C. et Villavicencio, A. (2016b). UFRGS&LIF at SemEval-2016 task 10 : rule-based MWE identification and predominant-supersense tagging. *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 910–917.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716.

Daille, B. (2017). *Term Variation in Specialised Corpora : Characterisation, automatic discovery and applications*, volume 19. John Benjamins Publishing Company.

Delaunay, B. et Laurent, N. (2012). *Bescherelle-La conjugaison pour tous*. Hatier.

Diab, M. T. et Bhutada, P. (2009). Verb noun construction MWE token supervised classification. *In Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications*, pages 17–22. Association for Computational Linguistics.

Dias, G., Guilloré, S. et Lopes, J. P. (2000). Normalisation of association measures for multiword lexical unit extraction. *In International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications*, pages 207–216. Citeseer.

do Amaral, D. O. F., Buffet, M. et Vieira, R. (2015). Comparative Analysis between Notations to Classify Named Entities using Conditional Random Fields. *In Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, pages 27–31.

Drouin, P. et Dury, P. (2009). When terms disappear from a specialized lexicon : A semi-automatic investigation into necrology. *In Actes de la conférence internationale "Language for Special Purposes"(LSP 2009)*. Citeseer.

Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., Marcellesi, J.-B. et Mével, J.-P. (1997). *Dictionnaire de linguistique*. Larousse.

Durrani, N., Haddow, B., Koehn, P. et Heafield, K. (2014). Edinburgh's phrase-based machine translation systems for WMT-14. *In Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 97–104.

Ehren, R., Lichte, T. et Samih, Y. (2018). Mumpitz at PARSEME Shared Task 2018 : A Bidirectional LSTM for the Identification of Verbal Multiword Expressions. *In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 261–267.

Eryiğit, G., Ilbay, T. et Can, O. A. (2011). Multiword expressions in statistical dependency parsing. *In Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 45–55. Association for Computational Linguistics.

Fazly, A., Cook, P. et Stevenson, S. (2009). Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1):61–103.

Fillmore, C. J., Kay, P. et O'connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions : The case of let alone. *Language*, pages 501–538.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Firth, J. R. (1961). *Papers in Linguistics 1934-1951 : Repr.* Oxford University Press.

Green, S., de Marneffe, M.-C. et Manning, C. D. (2013). Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Gross, G. (1988a). Degré de figement des noms composés. *Langages*, 90:57–71. Paris : Larousse.

Gross, M. (1975). Méthodes en syntaxe, Hermann, Paris.

Gross, M. (1987). The use of finite automata in the lexical representation of natural language. *In LITP Spring School on Theoretical Computer Science*, pages 34–50. Springer.

Gross, M. (1988b). Les limites de la phrase figée. *Langages*, 23(90):7–22.

Gross, M. et Senellart, J. (1998). Nouvelles bases statistiques pour les mots du français. *In Journée d'Analyse Statistique des Données Textuelles (JADT)*, page 335–349, Nice.

Hachey, B., Radford, W., Nothman, J., Honnibal, M. et Curran, J. R. (2013). Evaluating entity linking with Wikipedia. *Artificial intelligence*, 194:130–150.

Halliday, M. et Hasan, R. (1976). Cohesion in English. London : Longman. *English language series*.

Hammerton, J. (2003). Named entity recognition with long short-term memory. *In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 172–175. Association for Computational Linguistics.

Hassler, G. et Hümmer, C. (2005). Figement et défigement polylexical : l'effet des modifications dans des locutions figées. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (53):103–119.

Hazem, A. et Daille, B. (2014). Semi-compositional method for synonym extraction of multi-word terms. *In 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*.

Hilma, R. (2011). Literal Translation using Google Translate in Translating the Text from French to English in Digital Tourism Brochure "Bienvenue À Paris". *Binus Business Review*, 2(1):502–509.

Hutchins, J. (2005). Towards a definition of example-based machine translation. *In Machine Translation Summit X, Second Workshop on Example-Based Machine Translation*, pages 63–70.

Jackendoff, R. (1997). The architecture of the language faculty. *Linguistic Inquiry Monographs*.

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing*. The MIT Press.

Katiyar, A. et Cardie, C. (2018). Nested named entity recognition revisited. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 861–871.

Katz, G. et Giesbrecht, E. (2006). Automatic Identification of Non-compositional Multi-word Expressions Using Latent Semantic Analysis. *In Proceedings of the Workshop on Multiword Expressions : Identifying and Exploiting Underlying Properties*, MWE '06, pages 12–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

Khalid, M. A., Jijkoun, V. et De Rijke, M. (2008). The impact of named entity normalization on information retrieval for question answering. *In European Conference on Information Retrieval*, pages 705–710. Springer.

Korkontzelos, I. et Manandhar, S. (2010). Can recognising multiword expressions improve shallow parsing ? *In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 636–644. Association for Computational Linguistics.

Kumar, S., Behera, P. et Jha, G. N. (2017). A classification-based approach to the identification of Multiword Expressions (MWEs) in Magahi Applying SVM. *Procedia Computer Science*, 112:594–603.

Kumbhar, P. et Mali, M. (2016). A survey on feature selection techniques and classification algorithms for efficient text classification. *International Journal of Science and Research*, 5(5):1267–1275.

Lafferty, J., McCallum, A. et Pereira, F. C. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data.

Lapata, M. et Lascarides, A. (2003). Detecting novel compounds : The role of distributional evidence. *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 235–242. Association for Computational Linguistics.

Laporte, E. (1988). Reconnaissance des expressions figées lors de l'analyse automatique. *Langages*, 23(90):117–126.

Larrivée, P. et Moline, E. (2012). Un parcours de subjectification. Bel et bien : du redoublement de la manière au renforcement de l'assertion. *Travaux de linguistique*, 2(65):45–64.

Le Marchant, J. et Kunstmann, P. (1973). *Miracles de Notre-Dame de Chartres*, volume 1. Éditions de l'Université d'Ottawa.

Le Roux, J., Rozenknop, A. et Constant, M. (2014). Syntactic parsing and compound recognition via dual decomposition : Application to French. *In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, pages 1875–1885.

Leblanc, M. (1907). *Arsène Lupin gentleman-cambrioleur*. Éditions Pierre Lafitte.

Li, W., Zhang, X., Niu, C., Jiang, Y. et Srihari, R. (2003). An expert lexicon approach to identifying English phrasal verbs. *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 513–520. Association for Computational Linguistics.

Lichte, T. et Kallmeyer, L. (2016). Same syntax, different semantics : A compositional approach to idiomaticity in multi-word expressions. *In* Piñón, C., éditeur : *Empirical Issues in Syntax and Semantics 11*, pages 111–140.

Lichte, T., Petitjean, S., Savary, A. et Waszczuk, J. (2017). *Lexical encoding formats for multi-word expressions : The challenge of "irregular" regularities"*, pages 79–111. Language Science Press.

Lipka, L., Handl, S. et Falkner, W. (2004). Lexicalization & institutionalization : the state of the art in 2004. *SKASE Journal of Theoretical Linguistics*, 1(1):2–19.

Liu, X., Zhou, M., Wei, F., Fu, Z. et Zhou, X. (2012). Joint inference of named entity recognition and normalization for tweets. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers - Volume 1*, ACL '12, pages 526–535, Stroudsburg, PA, USA. Association for Computational Linguistics.

Loper, E. et Bird, S. (2002). NLTK : The Natural Language Toolkit. *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia : Association for Computational Linguistics.*

Losnegaard, G. S., Sangati, F., Escartín, C. P., Savary, A., Bargmann, S. et Monti, J. (2016). PARSEME Survey on MWE Resources. *In* Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. et Piperidis, S., éditeurs : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Malkiel, Y. (1959). Studies in irreversible binomials. *Lingua*, 8:113–160.

Mann, H. B. et Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Mathieu-Colas, M. (1995). Un dictionnaire électronique des mots à trait d'union. *Langue française*, pages 76–85.

Mejri, S. (2009). Figement, défigement et traduction. problématique théorique.

Mel'čuk, I. (1998). Collocations and lexical functions. *Phraseology. Theory, analysis, and applications*, pages 23–53.

Metin, S. K. (2018). Feature selection in multiword expression recognition. *Expert Systems with Applications*, 92:106–123.

Mollin, S. (2013). Pathways of change in the diachronic development of binomial reversibility in Late Modern American English. *Journal of English Linguistics*, 41(2):168–203.

Monti, J., Sangati, F. et Arcan, M. (2015). TED-MWE : a bilingual parallel corpus with MWE annotation. *In Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it*, pages 193–197.

Moreau, E., Alsulaimani, A., Maldonado, A. et Vogel, C. (2018). CRF-Seq and CRF-DepTree at PARSEME Shared Task 2018 : Detecting Verbal MWEs Using Sequential and Dependency-Based Approaches. *In Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018) at the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 241–247.

Morin, E. et Daille, B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44(1-2):79–95.

Morin, E. et Daille, B. (2012). Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique (Compositionality and Context for Bilingual Lexicon Extraction from Comparable Corpora)[in French]. *In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, pages 141–154.

Nakagawa, H. et Mori, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(2):201–219.

Nasr, A., Ramisch, C., Deulofeu, J. et Valli, A. (2015). Joint dependency parsing and multiword expression tokenization. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, volume 1, pages 1116–1126.

Nerima, L., Seretan, V. et Wehrli, E. (2016). Un outil multilingue d'extraction de collocations en ligne. *In Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 5 : Démos*, pages 34–36.

Nissim, M. et Zaninello, A. (2013). Modelling the internal variability of multiword expressions through a pattern-based method. *In ACM Transactions on Speech and Language Processing , Special issue on Multiword Expressions*, volume 10.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N. *et al.* (2016). Universal Dependencies v1 : A Multilingual Treebank Collection. *In LREC*.

Nivre, J. et Nilsson, J. (2004). Multiword units in syntactic parsing. *Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

Nunberg, G., Sag, I. A. et Wasow, T. (1994). Idioms. *Language*, 70:491–538.

Parmentier, Y. et Waszczuk, J. (2019). Representation and parsing of multiword expressions : Current trends.

Pasquer, C. (2017). Expressions polylexicales verbales : étude de la variabilité en corpus. *In TALN-RECITAL 2017*.

Pasquer, C., Ramisch, C., Savary, A. et Antoine, J.-Y. (2018a). VarIDE at PARSEME Shared Task 2018 : Are Variants Really as Alike as Two Peas in a Pod ? *In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 283–289. Association for Computational Linguistics.

Pasquer, C., Savary, A., Antoine, J.-Y. et Ramisch, C. (2018b). Towards a Variability Measure for Multiword Expressions. *In Proceedings of NAACL*. Accepted paper.

Pasquer, C., Savary, A., Ramisch, C. et Antoine, J.-Y. (2018c). If you've seen some, you've seen them all : Identifying variants of multiword expressions. *In Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*. The COLING 2018 Organizing Committee.

Pausé, M.-S. (2017). *Structure lexico-syntaxique des locutions du français et incidence sur leur combinatoire*. Thèse de doctorat, Université de Lorraine (Nancy).

Pearce, D. (2001). Synonymy in collocation extraction. *In Proceedings of the workshop on WordNet and other lexical resources, second meeting of the North American Chapter of the Association for Computational Linguistics*, pages 41–46.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. et Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger, J., Yangarber, R. *et al.* (2017). The First Cross-Lingual Challenge on Recognition, Normalization and Matching of Named Entities in Slavic Languages. *In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics.

Polguère, A. (2003). Collocations et fonctions lexicales : pour un modèle d'apprentissage. *Revue Française de Linguistique Appliquée*, pages 117–133.

Privat, M. (2008). Les jeux de mots dans la presse française et leur traduction en espagnol. *In La culture de l'autre [Recurso electrónico] : l'enseignement des langues à l'Université : Deuxième Rencontre Hispano-français de Chercheurs (SHF/APFUE), École Normale Supérieure Lettres et Sciences Humaines (26 au 29 novembre 2008)= Segundo Encuentro Hispanofrancés de Investigadores (APFUE/SHF)*, page 7. La clé des langues.

Quillard, G. (2001). La traduction des jeux de mots dans les annonces publicitaires. *TTR : traduction, terminologie, rédaction*, 14(1):117–157.

Ramisch, C. (2015). Multiword expressions acquisition. *A Generic and Open Framework. Cham : Springer International Publishing*.

Ramisch, C. (2017). Putting the Horses Before the Cart : Identifying Multiword Expressions Before Translation. *In International Conference on Computational and Corpus-Based Phraseology*, pages 69–84. Springer.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Mititelu, V. B., 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., Escartín, C. P., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova,

I., Vaidya, A. et Walsh, A. (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. ACL. https://aclweb.org/anthology/W18-4925.

Ramisch, C., De Araujo, V. et Villavicencio, A. (2012). A broad evaluation of techniques for automatic acquisition of multiword expressions. *In Proceedings of ACL 2012 Student Research Workshop*, pages 1–6. Association for Computational Linguistics.

Ramisch, C., Villavicencio, A. et Kordoni, V. (2013). Introduction to the special issue on multiword expressions : From theory to practice and use. *In ACM Transactions on Speech and Language Processing*, volume 10.

Rastier, F. (2005). Enjeux épistémologiques de la linguistique de corpus. *La linguistique de corpus*, pages 31–45.

Rebourcet, S. (2008). Le français standard et la norme : l'histoire d'une «nationalisme linguistique et littéraire» à la française. *Mot de l'équipe éditoriale*, page 107.

Rey, A. et la Peste, D. (2007). *Lexik des cités : lexik des cités illustré*. Fleuve noir.

Riedl, M. et Biemann, C. (2016). Impact of MWE resources on multiword recognition. *In Proceedings of the 12th Workshop on Multiword Expressions*, pages 107–111.

Riehemann, S. Z. (2001). *A constructional approach to idioms and word formation*. Thèse de doctorat, stanford university Stanford, CA.

Rish, I. *et al.* (2001). An empirical study of the naive Bayes classifier. *In IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Rosén, V., De Smedt, K., Losnegaard, G. S., Bejček, E., Savary, A. et Osenova, S. (2016). MWEs in treebanks : From survey to guidelines. *In Tenth International Conference on Language Resources and Evaluation (LREC 2016)* .

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.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *In 7th international conference on Language Resources and Evaluation (LREC 2010)*.

Salehi, B., Cook, P. et Baldwin, T. (2015). A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions. *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Sang, E. F. et De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Savary, A. (2008). Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, pages 1–53.

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S., Eryiğit, G., Giouli, V., Maarten, G., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Escartín, C. P., Plas, L., QasemiZadeh, B., Ramisch, C., Sangati, F., Stoyanova, I. et Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. *In* Markantonatou, S., Ramisch, C., Savary, A. et Vincze, V., éditeurs : *Multiword expressions at length and in depth : Extended papers from the MWE 2017 workshop*. Language Science Press, Berlin, Germany. http://langsci-press.org/catalog/view/204/1344/1319-1.

Savary, A., Cordeiro, S. et Ramisch, C. (2019a). Without lexicons, multiword expression identification will never fly : A position statement. *In Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91.

Savary, A. et Cordeiro, S. R. (2017). Literal readings of multiword expressions : as scarce as hen's teeth. *In Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 64–72.

Savary, A., Cordeiro, S. R., Lichte, T., Ramisch, C., nurrieta, U. I. et Giouli, V. (2019b). Literal Occurrences of Multiword Expressions : Rare Birds That Cause a Stir. *The Prague Bulletin of Mathematical Linguistics*, 112:5–54.

Savary, A. et Jacquemin, C. (2003). Reducing Information Variation in Text. *LNCS*, 2705:145–181.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I. et Doucet, A. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. *In Proceedings of the 13th Workshop on MWEs*, pages 31–47, Valencia, Spain. ACL. https://aclweb.org/anthology/W17-1704.

Schlogel, G. (1992). *Les princes du sang : roman.* Fayard.

Schneider, N., Danchik, E., Dyer, C. et Smith, N. A. (2014). Discriminative lexical semantic segmentation with gaps : running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Scholivet, M. et Ramisch, C. (2017). Identification of Ambiguous Multiword Expressions Using Sequence Models and Lexical Resources. *In Proceedings of the 13th Workshop on MWEs*, pages 167–175, Valencia, Spain. ACL. https://aclweb.org/anthology/W17-1723.

Scholivet, M., Ramisch, C. et Cordeiro, S. (2018). Sequence Models and Lexical Resources for MWE Identification in French. *In* Markantonatou, S., Ramisch, C., Savary, A. et Vincze, V., éditeurs : *Multiword expressions at length and in depth : Extended papers from the MWE 2017 workshop*. Language Science Press, Berlin, Germany. http://langsci-press.org/catalog/view/204/1651/1307-1.

Segura Bedmar, I., Martínez, P. et Herrero Zazo, M. (2013). SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIextraction 2013). *In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013)*. Association for Computational Linguistics.

Sekine, S., Sudo, K. et Nobata, C. (2002). Extended Named Entity Hierarchy. *In LREC*.

Seretan, V. (2011). *Syntax-based collocation extraction*, volume 44. Springer Science & Business Media.

Sheinfux, L. H., Greshler, T. A., Melnik, N. et Wintner, S. (2018). *Verbal MWEs : Idiomaticity and flexibility*, pages 5–38. Language Science Press, to appear.

Silberztein, M. (1990). Le dictionnaire électronique des mots composés. *Langue Française*, (87):71–83.

Sinclair, J. (1996). Preliminary recommendations on corpus typology. *EAGLES Document TCWG-CTYP/P (disponible à l'adresse* `http://www.ilc.pi.cnr.it/EAGLES/corpustyp/corpustyp.html`*)*.

Smadja, F. (1993). Retrieving collocations from text : Xtract. *Computational linguistics*, 19(1):143–177.

Spence, N. C. (1969). Composé nominal, locution et syntagme libre. *La linguistique*, 5(Fasc. 2):5–26.

Stodden, R., QasemiZadeh, B. et Kallmeyer, L. (2018). TRAPACC and TRAPACCS at PARSEME Shared Task 2018 : Neural Transition Tagging of Verbal Multiword Expressions. *In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 268–274.

Straka, M. et Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *In Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Stymne, S., Cancedda, N. et Ahrenberg, L. (2013). Generation of compound words in statistical machine translation into compounding languages. *Computational Linguistics*, 39(4):1067–1108.

Taslimipoor, S. et Rohanian, O. (2018). SHOMA at Parseme Shared Task on Automatic Identification of VMWEs : Neural Multiword Expression Tagging with High Generalisation. *CoRR*, abs/1809.03056.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task : Language-independent Named Entity Recognition. *In Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Treffers-Daller, J. (2012). Grammatical collocations and verb-particle constructions in Brussels French : a corpus-linguistic approach to transfer. *International Journal of Bilingualism*, 16(1):53–82.

Tsai, R. T.-H., Sung, C.-L., Dai, H.-J., Hung, H.-C., Sung, T.-Y. et Hsu, W.-L. (2006). Nerbio : using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *In BMC bioinformatics*, volume 7, page S11. BioMed Central.

Tsvetkov, Y. et Wintner, S. (2012). Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(4):549–573.

Tsvetkov, Y. et Wintner, S. (2014). Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, 40(2):449–468.

Tutin, A. (2008). For an extended definition of lexical collocations. *In Euralex*, page 00.

Tutin, A. (2016). Comparing morphological and syntactic variations of support verb constructions and verbal full phrasemes in French : a corpus based study. *In PARSEME COST Action. Relieving the pain in the neck in natural language processing : 7th final general meeting*, Dubrovnik, Croatia.

Tutin, A. et Grossmann, F. (2002). Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée*, 7(1):7–25.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Villavicencio, A., Bond, F., Korhonen, A. et McCarthy, D. (2005). Introduction to the special issue on multiword expressions : Having a crack at a hard nut. *Computer Speech & Language*, 19(4):365–377.

Vincze, V. (2012). Light verb constructions in the Szeged Paralell FX English-Hungarian parallel corpus.

Vincze, V., Nagy, T. I. et Berend, G. (2011). Detecting noun compounds and light verb constructions : a contrastive study. *In Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World*, pages 116–121. Association for Computational Linguistics.

Vincze, V., Nagy T, I. et Zsibrita, J. (2013). Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(2):6.

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 Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 275–282. Association for Computational Linguistics.

Wilcoxon, F. (1946). Individual comparisons of grouped data by ranking methods. *Journal of economic entomology*, 39(2):269–270.

Williams, G. (2003). Les collocations et l'école contextualiste britannique. *Les collocations : analyse et traitement*, pages 33–44.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. *et al.* (2016). Google's neural machine translation system : Bridging the gap between human and machine translation. *arXiv preprint arXiv :1609.08144*.

Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law. *Libr. Trends*, 30(1):53.

Yadav, V. et Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. *In Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.

Yvon, F. (2010). Une petite introduction au traitement automatique des langues naturelles. *In Conference on Knowledge discovery and data mining*, pages 27–36.

Zampieri, N., Scholivet, M., Ramisch, C. et Favre, B. (2018). Veyn at PARSEME Shared Task 2018 : Recurrent Neural Networks for VMWE Identification. *In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 290–296.

Zarriess, S. et Kuhn, J. (2009). Exploiting translational correspondences for pattern-independent mwe identification. *In Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications*, MWE'09, pages 23–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhang, Y., Kordoni, V., Villavicencio, A. et Idiart, M. (2006). Automated multiword expression prediction for grammar engineering. *In Proceedings of the Workshop on Multiword Expressions : Identifying and Exploiting Underlying Properties*, MWE '06, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhou, W., Yu, C., Smalheiser, N., Torvik, V. et Hong, J. (2007). Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 655–662. ACM.

Index

Ambiguïté, 57
  co-occurrence fortuite, 57
  lecture littérale, 57
Argument, 46

Binôme irréversible, 35

Chevauchement, 92
Chi2, 183
Collocation, 41
  base-collocatif, 42
  statistique, 42
Composant, 33
Construction multi-verbe, 114
Corpus
  CoNLL, 136
  PARSEME, 114
Cranberry word, 52
Construction à verbe support, 44
  LVC.cause (PARSEME), 113
  LVC.full (PARSEME), 113

Découverte, 67
Défigement, 52
Dictionnaire, 65
Discontinuité, 40
Sémantique distributionnelle, 69
  plongement de mots, 69

Elément lexicalisé, 33
Entité nommée, 43
Expression polylexicale
  définition, 36
  mono-token, 34
  multi-tokens, 33

Figement, 35
Forme canonique, 50

Identification, 75
  variantes, 101
Idiome, 41
  figuratif, 41
    opaque, 41
    transparent, 41
  PARSEME, 114
Idiosyncrasie, 34
  lexicale, 34
  morphologique, 35
  pragmatique, 35
  sémantique, 35
  statistique, 35
  syntaxique, 34
Verbe intrinsèquement réflexif, 114

LemmNorm, 109
Lexème, 33
Lexique-Grammaire, 66

Machine learning, 78
  Arbre de décision, 85
  CRF, 81
  Naïve Bayes, 81
  perceptron multi-couches, 86
  réseaux de neurones, 86
  SVM, 82
Mesures d'association, 68
Métriques
  F-mesure, 73
  F-per-EP, 89
  F-per-token, 89
  précision, 73
  rappel, 73
  similarité
    Jaccard, 188
    Sørensen–Dice, 147
Modifieur, 46
Mot, 33

Mot composé, 39

Nécrologie, 52
Non-compositionnalité, 34

POSNorm, 161
POSseq, 161

Règles, 77

Similarité
    linéaire, 148
    syntaxique, 147
Soudure, 34

Terme, 40
Token
    vs. mot, 33
    vs. type, 49

Variabilité, 49
    continuum, 54
    diachronique, 52
    diastratique, 52
    diatopique, 52
    lexicale, 71
    morphosyntaxique, 72
    profil, 109, 133
    score, 145
Verbe à particule, 42

Zipf (loi de), 45

Résumé :

L'identification automatique d'expressions polylexicales (EP) est un pré-requis pour de nombreuses applications de traitement automatique des langues. Cette tâche représente un défi car les EP, et en particulier les verbales (EPV) telles que ***casser sa pipe*** (signifiant *mourir*), ont des formes de surface très variables (***cassera-t-il*** *un jour* ***sa pipe*** ?). Cependant, comparée à des constructions libres, cette variabilité est généralement plus restreinte (p. ex. certains noms non modifiables par un adjectif), d'où des profils de variabilité distincts. On se penche ici sur un sous-problème de l'identification d'EPV, à savoir l'identification d'occurrences d'EPV vues dans d'autres contextes, quelque soit leur forme de surface, ce qui nécessite de prendre en compte l'ambiguïté pour éviter des lectures littérales (*casser sa vieille pipe*) ou des co-occurrences fortuites (*casser le tuyau de sa pipe*).

On considère pour cela deux approches : la première se fonde sur une mesure de la variabilité des EPV indépendante de la langue. La seconde consiste à modéliser le problème comme une tâche de classification d'après des traits pertinents pour la variabilité morpho-syntaxique des EPV, ce qui nous a conduit à développer un système (VarIDE), qui a participé à la compétition PARSEME d'identification automatique d'EPV en 2018.

Mots clés :

Expression polylexicale, variabilité, ambiguïté, traitement automatique des langues

Abstract :

Automatic identification of multiword expressions (MWEs) is a pre-requisite for many natural language processing applications. This task is challenging because MWEs, especially verbal ones (VMWEs) like *to* ***kick the bucket*** (which means *to die*), exhibit surface variability (*no* ***buckets*** *were* ***kicked***). However, compared with regular constructions, this variability is usually more restricted (e.g. some nouns cannot be modified by an adjective), hence various variability profiles. We address here a subproblem of VMWE identification, namely the identification of occurrences of VMWEs previously seen in corpora, whatever their surface form, which requires to take ambiguity into account to avoid literal (*he kicked the old bucket*) or coincidental occurrences (*he kicked the ball and the bucket fell down*).

To this end, we considered two main approaches : The first one is based on a language-independent measure of VMWE variability. The second one consists in modeling the problem as a classification task on the basis of features relevant to the VMWE morpho-syntactic variability, which led to a system (VarIDE) that participated in the PARSEME shared task on automatic identification of VMWEs in 2018.

Keywords :

Multiword expression, MWE, variability, ambiguity, natural language processing