Ressources linguistiques et identification automatique d'expressions polylexicales

Mathieu Constant

Ressources linguistiques et identification automatique d'expressions polylexicales

Mathieu Constant

Université de Lorraine, CNRS, ATILF, F-54000 Nancy, France

Mathieu.Constant@univ-lorraine.fr

1 Introduction

Le traitement automatique des langues est un domaine cherchant à concevoir des méthodes et outils pour analyser et générer automatiquement des données langagières. Les chercheurs dans ce domaine sont généralement confrontés à de nombreux problèmes liés à la langue, et notamment le traitement des expressions polylexicales qui représentent un ensemble de phénomènes linguistiques que l'on retrouve fréquemment dans les textes. Les *expressions polylexicales* sont généralement définies comme des combinaisons de mots qui montrent une certaine irrégularité de composition à un ou plusieurs niveaux linguistiques. Les exemples prototypiques sont en général des expressions idiomatiques totalement opaques sémantiquement : ex. *cordon bleu* (signifiant « excellent cuisinier »), *prendre le taureau par les cornes* (signifiant « s'attaquer sérieusement à une difficulté »). Mais ces expressions sont en réalité très variées et montrent une grande disparité comme cela est montré dans (Sag et al. 2002).

Leur identification automatique dans les textes est une étape cruciale pour le traitement automatique des langues, et notamment des applications telles que la traduction automatique. Par exemple, la phrase *Luc avait un coup dans le nez à cette soirée* où l'expression idiomatique *avoir un coup dans le nez* signifie « être ivre », est traduite littéralement par *Luc has a punch in the nose* par le traducteur automatique de Google (https://translate.google.fr, consultation : 31-07-2020), ce qui est clairement erroné. L'indication au système que *avait un coup dans le nez* est une occurrence d'expression idiomatique, avec idéalement sa traduction dans la langue cible, peut être extrêmement utile pour de tels systèmes, comme cela est montré dans (Constant et al. 2017).

Dans le cadre de nos travaux de recherche, nous nous intéressons à l'identification automatique de telles expressions dans les textes. Les approches que nous mettons en œuvre sont génériques, dans le sens où elles sont applicables à la majorité des langues. Nous portons néanmoins une attention particulière au français. Nos méthodes s'appuient néanmoins sur des ressources linguistiques telles que des corpus annotés ou des ressources lexicales, propres à chaque langue. Dans cet article, nous

montrons différentes approches d'identification à partir de ces deux types de ressources, sur lesquels nous avons travaillé. Pour un état de l'art plus complet concernant le domaine, nous invitons les lecteurs à se référer notamment à (Constant et al. 2017).

Cet article est organisé comme suit. Nous présentons, dans un premier temps, le traitement automatique des langues et l'exploitation des ressources linguistiques dans ce cadre (section 2). Puis, dans la section 3, nous nous intéresserons aux expressions polylexicales : nous en donnerons une définition, puis nous indiquerons des critères linguistiques d'identification et les principaux défis pour leur détection automatique dans des textes. Nous décrirons ensuite trois méthodes d'identification automatique : la section 4 sera dédiée à l'identification à partir d'un apprentissage sur un corpus annoté, la section 5 sera dédiée à l'identification à partir de ressources lexicales ; enfin la section 6 s'intéressera au couplage des deux méthodes précédentes. Nous terminerons par la section 7 qui présentera brièvement le projet PARSEME-FR dédié au traitement automatique des expressions polylexicales.

2 Le traitement automatique des langues

Le traitement automatique des langues (TAL) est un domaine de recherche dont l'objectif est de concevoir des méthodes et outils permettant d'analyser et de générer des données en langage naturel. C'est un domaine pluridisciplinaire qui se trouve à la croisée des chemins entre plusieurs disciplines comme la linguistique, l'informatique ou l'intelligence artificielle. Certaines applications du TAL sont très connues du grand public comme la traduction automatique, popularisée par les services de traduction de Google par exemple. Mais il y a aussi des applications telles que le résumé automatique de textes, ou la recherche d'informations spécialisées dans de grandes bases textuelles. Le TAL est historiquement lié au domaine de la linguistique. Pendant très longtemps, les informaticiens et les linguistes ont travaillé main dans la main. De multiples outils d'analyse linguistique automatique ont vu le jour grâce à cette collaboration. Nous citerons en particulier le découpage et l'étiquetage lexical qui identifie et catégorise les unités lexicales, l'étiquetage morphologique (analysant la forme de ces unités), l'analyse syntaxique produisant la structure des phrases ou même l'analyse sémantique qui calcule le sens des mots et des phrases… tout cela automatiquement. Nous proposons, dans la figure 1, un exemple d'analyse linguistique potentiellement produite par un outil du TAL. L'exemple décrit l'étiquetage morphosyntaxique et l'analyse syntaxique en dépendances de la phrase *ceci remet sa garde à vue en cause* . L'étiquetage morphosyntaxique est indiqué sous la phrase. Chaque mot est associé à une étiquette morphosyntaxique. Par exemple, *remet* est associé à la catégorie VERBE. La structure syntaxique est décrite par des relations de dépendance entre les mots. Par exemple, *remet* est associé à *garde* par une relation étiquetée OBJ indiquant le relation complément d'objet.

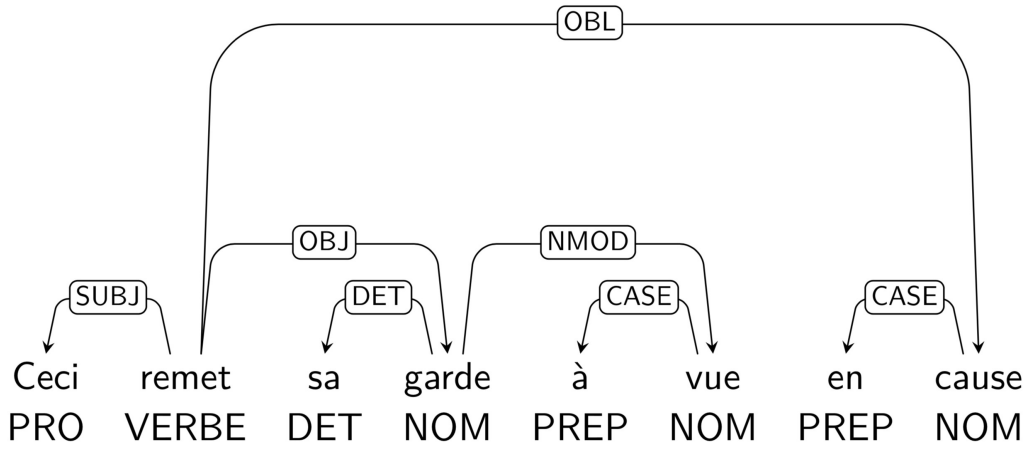

Figure 1: Exemple d'analyse syntaxique

La recherche en TAL est confrontée à de multiples obstacles inhérents à la langue naturelle. Le premier obstacle, le plus connu, est l'ambiguïté : l'ambiguïté des mots ou des attachements entre les mots. Dans notre exemple, le mot *cause* a, dans l'absolu, deux analyses morphosyntaxiques possibles : soit un verbe (*causer*), soit un nom (*cause*). C'est le contexte d'occurence qui va aider à lui associer la bonne analyse.

Une autre difficulté consiste à mettre au point un mécanisme pertinent de composition du sens à partir du sens des mots, l'un des buts majeurs du TAL étant notamment de calculer le sens des textes donnés en entrée des outils. Par ailleurs, nous avons à faire face au phénomène de l'idiomaticité qui est extrêmement fréquent dans les textes et, qui quelque part, représente une irrégularité à prendre en compte dans le mécanisme de composition : ex. le sens de l'expression idiomatique *casser du sucre sur le dos (de quelqu'un)* signifiant « dire du mal (de quelqu'un) en son absence » ne peut être dérivé à partir du sens des mots *casser, sucre* et *dos*. Notre exemple de la Figure 1 compte deux expressions idiomatiques : *garde à vue* et *remet en cause*.

Le TAL connaît actuellement une petite révolution avec la ré-émergence de l'apprentissage profond plus connu sous le terme « deep Learning » en anglais, et avec le développement de réseaux de neurones pour modéliser les différentes tâches du TAL. Les performances des outils ont fait des bonds spectaculaires grâce à cela. Mais, l'un des inconvénients de ce type d'approche et d'être extrêmement gourmand en données, en plus d'avoir un cout computationnel important (et donc énergivore). Les données prises en compte par ces outils prennent plusieurs formes. Tout d'abord, il y a les données annotées qui servent d'exemples pour l'apprentissage des modèles. Les annotations qui correspondent aux analyses que doivent produire le modèle, sont réalisées manuellement, voire automatiquement avec une validation manuelle par des linguistes. A partir des exemples, les méthodes d'apprentissage automatique tentent d'inférer un modèle pour la tâche cible. De gros corpus bruts contenant uniquement du texte sont également utilisés et servent à connaître la

distribution des mots et leurs contextes d'occurrence en général, et ainsi à améliorer l'apprentissage des modèles. Enfin, il y a les ressources lexicales qui sont produites par des linguistes et qui peuvent apporter des informations complémentaires des deux précédents types de ressources. Pour plus de détails sur différentes approches utilisées en TAL, nous proposons aux lecteurs de se référer à (Eisenstein 2019).

3 Les expressions polylexicales

Dans cet article, nous nous intéressons au traitement automatique des expressions poylexicales et plus précisément à leur identification automatique dans des textes. Dans un premier temps, nous allons donner une définition de ce que nous entendons pas le terme *expression polylexicale*. Puis nous donnerons quelques critères linguistiques permettant de les identifier. Enfin, nous indiquerons les défis majeurs pour l'identification automatique de telles expressions dans des textes.

Définition. Les expressions polylexicales (EP) – ou « multiword expressions » en anglais – sont des expressions formées de plusieurs mots, qui affichent une irrégularité de composition (i.e. une idiosyncrasie). Cette irrégularité peut être située à un ou plusieurs niveaux linguistiques. Comme exemple d'idiosyncrasie au niveau sémantique, nous pouvons citer *pomme de terre* qui est un mot composé dont on ne peut pas connaître le sens à partir du sens de ses composants. Une *pomme de terre* n'est pas une pomme (ou un fruit) qui vient de la terre. De même, pour l'expression idiomatique *mettre les voiles* dans le sens de « partir ».

La définition ci-dessus reprend en grande partie les définitions de référence que l'on retrouve dans la littérature, cf. (Sag et al. 2002), (Baldwin ; Kim 2010) ou (Savary et al. 2017). Comme dans la littérature sur le sujet, elle est cependant relativement floue. Le contour exact des expressions polylexicales dépend beaucoup du cadre théorique linguistique dans lequel on se trouve. Par ailleurs, cette définition couvre des phénomènes linguistiques très variés. Par exemple, *bien que* et *en dépit de* -- ce dernier voulant dire « malgré » – représentent des mots grammaticaux complexes. La séquence *à reculons* est une expression adverbiale qui présente des irrégularités au niveau morphosyntaxique et syntaxique. En effet, *reculons* correspond au verbe *reculer* à la première personne du pluriel au présent de l'indicatif. Or, un verbe précédé de la préposition *à* doit être à l'infinitif, d'où l'irrégularité de composition de la séquence. Il existe également des expressions nominales comme *moulin à paroles* qui désigne une personne qui parle beaucoup ou comme *Los Angeles* qui forme un nom propre. Il existe aussi des expressions verbales dont nous donnons des exemples provenant des trois catégories que nous avons traitées dans nos travaux. Tout d'abord, l'expression idiomatique *lancer des fleurs (à quelqu'un)* qui veut dire « flatter (quelqu'un) ». L'expression *se souvenir* correspond à un verbe intrinsèquement pronominal : le verbe *souvenir* en tant que tel n'existe pas seul, le pronom réflexif est obligatoire. Les expressions verbales peuvent

aussi être des constructions à verbe support dont le sens est porté par un nom prédicatif en position de sujet ou de complément. Dans l'expression *faire un choix*, c'est le nom *choix* qui porte le sens de la phrase. Le sens du verbe est neutralisé et la sélection de ce verbe dépend du nom.

Critères linguistiques d'identification. Les expressions polylexicales ont un lien très fort avec le domaine de la phraséologie en linguistique, où les différentes classes d'expressions ont déjà été largement étudiées dans la communauté. Un certain nombre de critères linguistiques d'identification ont été mis au point notamment dans le cadre de recensement systématique au sein de grandes bases lexicales – *inter alia* (Gross 1982) – ou de campagnes d'annotation de corpus – entre autres (Abeillé et al. 2003) –. Afin d'avoir une méthodologie rigoureuse de construction des ressources, il est important de mettre en place des critères opératoires d'identification linguistique des expressions.

Dans un certain nombre de projets de recherche tels que ceux décrits dans (Gross 1982) pour la constitution de lexiques syntaxiques ou (Savary et al. 2017) pour l'annotation d'expressions verbales dans des corpus, les critères fonctionnent de la manière suivante: étant donné une séquence candidate, on lui applique une opération linguistique simple. Le critère est satisfait si l'opération produit une séquence interdite ou une séquence dont le sens comporte une modification qui va au-delà de l'opération initiale. Nous donnons maintenant deux exemples de critères pour illustrer ce principe. Le premier est très classique. On remplace l'un des éléments pleins de l'expression par un mot sémantiquement lié tel qu'un synonyme ou un hyperonyme. Le critère est satisfait si l'opération produit une séquence interdite ou une séquence dont le sens comporte une modification qui va au-delà de la substitution initiale. Par exemple, le critère est satisfait lorsque l'on substitue le mot *eau* par *boisson* dans la séquence *eau de vie* correspondant à une boisson alcoolisée : **boisson de vie*. Comme deuxième exemple, nous donnons un critère morphosyntaxique où l'on modifie un trait morphologique d'un des éléments de l'expression. Par exemple, le mot au pluriel *voiles* de la séquence *met les voiles* est transformé en son équivalent au singulier. En suivant les règles de la grammaire, la séquence devient *met la voile* qui n'a rien plus à voir avec le sens initial. Le critère est donc satisfait.

Les défis du traitement automatique des expressions polylexicales. La prise en compte de tels phénomènes est crucial pour le TAL, comme l'ont montré d'illustres chercheurs en linguistique et en traitement automatique des langues – par ex. (Gross 1986), (Sag et al. 2002) –. Ce constat est particulièrement vrai pour la traduction automatique, comme le montre cet exemple d'erreur flagrante de traduction par le service de traduction automatique de Google : la phrase « A cette soirée, Luc avait un coup dans le nez » est traduite littéralement par « At this evening, Luc had a

blow in the nose » [1] (consultation : 31-07-2020), alors que l'expression idiomatique *avoir un coup dans le nez* signifie « être ivre » en français.

Dans cet article, nous nous intéressons à l'identification des expressions polylexicales qui consiste à annoter automatiquement les occurrences de ces expressions dans des textes donnés en entrée. Cette tâche fait face à de nombreux défis. Tout d'abord, le premier d'entre eux est bien sûr la détection de la non-compositionalité inhérente à la définition des expressions polylexicales. La discontinuité est aussi une difficulté importante. En effet, les composants d'une expression ne sont pas toujours juxtaposés les uns à côté des autres car il peut exister des éléments extérieurs à l'expression au sein de la séquence. Par exemple, dans la phrase *Marie prend très souvent part aux conversations* , il existe deux mots *très* et *souvent* entre les composants de l'expression *prend part* . Une autre difficulté est la variation potentielle des expressions que ce soit au niveau flexionnel ou syntaxique. Par exemple, le nom composé *pomme de terre* possède une forme au singulier (*pomme de terre*) et une autre au pluriel (*pommes de terre*). Autre exemple, la construction à verbe support *faire un choix* autorise des transformations syntaxiques, comme ici la relativisation: *Le choix que Luc fait*.

Enfin, l'ambiguïté pose aussi des problèmes. En effet, une expression peut aussi prendre son sens littéral suivant le contexte. Par exemple, *prendre la porte* dont le sens idiomatique veut dire « sortir », peut très bien être pris dans son sens littéral comme dans « est-ce que tu peux prendre la porte et la fenêtre, puis les ramener chez moi avec ton camion ? ». Il peut arriver que les composants de l'expression apparaissent ensemble de manière accidentelle. Par exemple, les deux mots *bien* et *que* peuvent se trouver juxtaposés au sein d'une phrase alors qu'il n'ont rien à voir avec le mot grammatical complexe *bien que* . Dans la phrase *j'aime bien que tu répondes à mes messages* , l'adverbe *bien* peut être substitué très naturellement par *beaucoup*, et le mot *que* correspond à une conjonction de subordination qui introduit la complétive servant de complément au verbe *aimer*.

Dans les trois prochaines sections, nous allons décrire des méthodes d'identification des expressions polylexicales à partir de deux types de ressources linguistiques (les corpus annotés et les ressources lexicales).

4 Identification par apprentissage sur corpus annoté

Une première méthode d'identification automatique des expressions polylexicales dans les textes consiste à utiliser un modèle d'annotation qui a été appris automatiquement à partir d'un corpus annoté en expressions polylexicales. Les exemples d'annotations vont permettre d'inférer le modèle.

Nous allons maintenant montrer un exemple simple de méthode d'identification qui est très utilisée dans la communauté. Dans cette approche, l'identification des expressions peut être vue comme une

1 La phrase est traduite par *At that party, Luc had a punch in the nose* par le traducteur en ligne DeepL (https://www.deepl.com/translator, consultation : 31-07- 2020).

tâche d'étiquetage de la phrase comme l'est l'étiquetage morphosyntaxique. Nous avons notamment abordé cette approche dans (Constant ; Sigogne 2011) et (Constant et al. 2013) pour le français, mais aussi dans (Constant et al. 2018) pour le Serbe . À partir de données d'apprentissage, il est possible d'apprendre un modèle qui permet d'associer à chacun des mots de la phrase une étiquette indiquant si le mot appartient à une expression ou pas, et, si tel est le cas, à quelle position dans l'expression, comme on le montre dans l'exemple ci-dessous.

Le	**premier**	**ministre**	**met**	souvent	**les**	**voiles**	vers	Bayonne
O	B	I	B	O	I	I	O	O

Dans la phrase *Le premier ministre met souvent les voiles vers Bayonne* où *premier ministre* et *met les voiles* sont des expressions, le mot *souvent* sera associé à la catégorie O voulant dire que le mot ne se trouve pas dans une expression (O pour « Outside » en anglais). L'étiquette B associée au mot *met* indique que ce dernier se trouve au début d'une expression (B pour « Begining » en anglais). Enfin, la catégorie I associée aux mots *les* et *voiles* dans l'exemple montre que ces mots se trouvent en position non-initiale d'une expression polylexicale (I pour « Inside » en anglais). On pourra noter que l'utilisation distincte des catégories B et I permet d'étiqueter deux expressions juxtaposées dans la même phrase. Après l'étiquetage des mots, une procédure automatique permet de récupérer les occurrences des expressions. Il existe des variantes de ce type d'annotation séquentielle comme le montre l'exemple ci-dessous.

Le	**premier**	**ministre**	**met**	souvent	**les**	**voiles**	vers	Bayonne
O	B-NOM	I-NOM	B-VERBE	O	I-VERBE	I-VERBE	O	O

Dans cet exemple, on complexifie le jeu d'étiquettes en juxtaposant les étiquettes B et I avec la catégorie grammaticale de l'expression à laquelle appartient le mot. Par exemple, *premier ministre* est une expression nominale (étiquette NOM) et *met les voiles* est une expression verbale (étiquette VERBE). C'est la variante que nous avons utilisé dans (Constant et al. 2013) pour pré-identifier les expressions polylexicales continues non-verbales avant de réaliser une analyse syntaxique automatique.

Dans ce type d'approche, il existe deux phases : (1) une phase d'apprentissage du modèle à partir de données annotées dans l'une des variantes montrées ci-dessus ou équivalente ; (2) une phase d'étiquetage d'une nouvelle séquence de mots donnée en entrée et qui va produire une séquence d'étiquettes associées aux différents mots à partir du modèle appris en phase (1). A partir de cette séquence d'étiquettes, il sera alors possible de reconstituer l'ensemble des expressions polylexicales se trouvant dans la séquence en entrée. Dans les expériences classiques, les jeux de données annotées sont généralement découpées en deux parties disjointes : une grande partie (ex. 80 %) qui

va servir de corpus pour l'apprentissage du modèle ; une plus petite partie (ex. 20%) pour l'évaluation du modèle. Les annotations des données d'évaluation sont considérées comme les annotations de référence auxquelles seront comparées les annotations obtenues automatiquement sur les mêmes données (non annotées) à l'aide du modèle appris sur le corpus d'apprentissage.

Les méthodes d'identification par étiquetage séquentiel sont extrêmement populaires dans la communauté du traitement des expressions polylexicales car elles sont à la fois relativement simple à mettre en œuvre et efficaces. Mais il existe d'autres approches parfois plus sophistiquées qui permettent de contrecarrer certaines limitations – cf. (Constant et al. 2017), (Al Saied 2019) pour un panorama détaillé –. En particulier, il existe des méthodes d'identification se fondant sur l'interaction des expressions polylexicales avec l'analyse syntaxique. L'une des hypothèses sous-jacentes à ce type d'approche est que l'identification des expressions polylexicales peut aider l'analyse syntaxique et inversement. Dans le cadre de l'étiquetage séquentiel à la mode IOB, nous avons par exemple intégré l'étiquetage morphosyntaxique avec la reconnaissance des expressions polylexicales (Constant ; Sigogne 2011). En plus de l'étiquetage des expressions polylexicales avec leur catégorie morphopsyntaxique, nous avons ajouté l'étiquetage morphosyntaxique des mots n'appartenant pas à une expression dans le schéma d'annotation comme dans l'exemple ci-dessous.

Le	**premier**	**ministre**	**met**	souvent	**les**	**voiles**	vers	Bayonne
DET	B-NOM	I-NOM	B-VERBE	ADV	I-VERBE	I-VERBE	PREP	NOM

Les données annotées les plus souvent utilisées pour les expériences d'annotation de ce type d'expressions sont le corpus arboré de Paris 7 (Abeillé et al. 2003) pour le français , le corpus DiMSUM (Schneider et al. 2016) pour l'anglais , et le corpus multilingue PARSEME éditions 1.0 (Savary et al. 2017) et 1.1 (Ramisch et al. 2018) . Le corpus arboré de Paris 7 inclut l'annotation des expressions principalement non-discontinues et non-verbales. Le corpus DiMSUM contient des annotations de tous types. Le corpus PARSEME contient des données annotées en expressions verbales pour une vingtaine de langues. Il existe bien évidemment bien d'autres corpus annotés en expressions polylexicales que nous ne citerons pas par manque de place. Nous invitons par exemple le lecteur à se référer à (Rosén et al. 2015) qui fait un état des lieux des corpus annotés en syntaxe contenant des annotations d'expressions polylexicales. Pour le français, on notera l'existence de plusieurs autres corpus annotés : voir, par exemple, (Laporte et al. 2008), (Candito ; Seddah 2012), (Tutin et al. 2015) ou (Nivre et al. 2016).

L'identification au moyen d'une approche par apprentissage du mod èle d'annotation à partir d'un corpus annoté a plusieurs avantages. En particulier, la mise en contexte des expressions permet d'apprendre au modèle à gérer au moins partiellement les ambiguïtés et les variations. Cependant,

elle se heurte à un problème majeur : les performances se dégradent fortement pour les expressions qui n'ont jamais été vues dans le corpus d'apprentissage (Savary et al. 2019). La faible couverture des corpus annotés en expressions polylexicales est un frein pour ce type d'approche. L'utilisation de ressources lexicales, couvrant notamment des expressions polylexicales rarement présentes en corpus, peut se montrer intéressante.

5 Identification avec des ressources lexicales

Nous nous intéressons maintenant à l'identification des expressions polylexicales à partir de ressources lexicales. Comme nous l'avons évoqué dans la section précédente, les corpus annotés ont une couverture limitée des expressions polylexicales, ce qui n'est pas le cas pour les ressources lexicales. Une identification au moyen de ressources lexicales peut donc se révéler intéressante. C'est d'ailleurs une approche qui a fait ses preuves depuis longtemps comme dans (Silberztein 1993).

Afin d'être utilisées pour l'identification automatique d'expressions dans des textes, les ressources lexicales ont besoin d'être explicitement formalisées et/ou être l'objet de procédures automatiques. Prenons l'exemple du DELACF (Courtois et al. 1997) qui est un dictionnaire électronique de mots composés pour le français. Une entrée lexicale contient non seulement sa forme de base, mais aussi sa catégorie grammaticale, et facultativement une étiquette indiquant sa structure interne et des traits sémantiques. Ces informations ne sont pas suffisantes pour repérer les occurrences des expressions correspondantes dans des textes. Ainsi, les entrées lexicales du DELACF indiquent aussi les formes fléchies associées, ainsi que leurs traits morphologiques. Ces formes fléchies peuvent être obtenues de manière semi-automatique en assignant une classe flexionnelle associée à différentes règles de flexion (Savary 2009). Ainsi, pour l'expression *pomme de terre* , nous aurons les informations suivantes :

pomme de terre,pomme de terre.N+NDN+Conc:fs

pommes de terre,pomme de terre.N+NDN+Conc:fp

Dans cet exemple, la forme avant la virgule correspond à la forme fléchie, la forme entre la virgule et le point correspond à la forme lemmatisée. Les étiquettes N, NDN et Conc correspondent respectivement à la catégorie grammaticale (N pour nom), à la structure syntaxique (NDN pour le patron Nom de Nom) et à un trait sémantique (Conc pour concret) du mot composé *pomme de terre*. Les lettres f, s et p après le double point représentent les traits morphologiques féminin, singulier et pluriel.

Pour repérer les occurrences des expressions codées dans le dictionnaire, il existe plusieurs outils possibles. Par exemple, le logiciel Unitex (https://unitexgramlab.org/fr) repère toutes les

occurrences potentielles dans le texte en essayant de trouver une correspondance directe avec les formes fléchies du dictionnaire. Cela permet en particulier de récupérer toutes les analyses lexicales possibles pour une phrase donnée. Dans nos travaux, nous avons repris ce principe tout en ajoutant une méthode simple permettant de générer une seule segmentation lexicale pour la phrase en entrée : nous sélectionnons la segmentation la plus courte de la phrase favorisant ainsi les segments lexicaux identifiés les plus longs (une expression polylexicale comptant pour un segment lexical). Si l'on prend l'exemple ci-dessous en supposant que les expressions *premier ministre*, *en effet* et *à partir de* sont dans notre dictionnaire, la consultation de ce même dictionnaire proposerait huit segmentations possibles de la phrase (nous représentons ces segmentations de manière factorisée):

Le	premier	ministre	**construit**	en	effet	**son**	**discours**	à	partir	de	**statistiques**	**économiques**
	premier_ministre			**en_effet**				**à_partir_de**				

La segmentation la plus courte (en gras) est alors sélectionnée comme segmentation de la phrase. L'inconvénient d'une telle approche est qu'elle ne permet pas de gérer les discontinuités dans les expressions. Pour les expressions verbales, c'est particulièrement problématique car elles sont sujettes à de multiples variations syntaxiques impliquant des discontinuités : *Luc fait face à cette situation*, *Luc fait souvent face à cette situation.* La gestion de la coordination pour les expressions nominales possédant des éléments communs est également problématique : la séquence *acides aminé et chlorhydrique* contient les noms composés *acide aminé* et *acide chlorhydrique*. Il est donc nécessaire de mettre en place des heuristiques pour gérer ces différents cas de figure. Par ailleurs, il existe d'autres types de ressources lexicales d'expressions polylexicales : par exemple, des lexiques syntaxiques – ex. (Gross 1982) –. Pour chaque type de ressource lexicale, il est nécessaire de mettre au point des méthodes spécifiques de projection des ressources sur les textes : ex. (Savary ; Waszczuk 2017) pour les lexiques syntaxiques. Nous vous invitons à lire (Constant et al. 2017) pour avoir un panorama plus complet de telles approches.

Malgré l'intérêt d'utiliser des ressources lexicales pour l'identification des expressions polylexicales, de telles approches ont cependant certains défauts. Tout d'abord, une expression non présente dans la ressource lexicale ne pourra pas être identifiée dans un texte, contrairement aux méthodes d'identification par apprentissage de modèles à partir d'un corpus annoté. En effet, ces dernières permettent parfois de repérer des expressions non présentes dans le corpus d'apprentissage, bien que les performances puissent être assez dégradées dans ce cas de figure (Savary et al. 2019).

Par ailleurs, l'ambiguïté n'est pas vraiment gérée non plus. Par exemple, elle ne l'est pas du tout avec la méthode par consultation simple de dictionnaires présentée ci-dessus. A titre d'illustration, en supposant que l'expression adverbiale *sur ce* (signifiant « là-dessus ») se trouve dans le dictionnaire,

on identifierait de manière erronée cette expression dans la phrase *Sur ce dessin, la couleur est jolie.* En effet, les deux mots de l'expression sont juxtaposés de manière accidentelle dans le texte. Ils se composent de manière régulière dans le cadre du groupe nominal prépositionnel *sur ce dessin*.

Sur	ce	**dessin**	**,**	**la**	**couleur**	**est**	**jolie**
Sur_ce							

6 Identification hybride

Les données annotées sont l'élément incontournable pour apprendre un modèle TAL moderne. Dans notre cas, ces données annotées sont des corpus annotés en expressions polylexicales (cf. section 4). Le principal problème d'utiliser uniquement des corpus annotés pour apprendre un modèle d'identification est leur couverture en expressions qui est assez limitée. En effet, la littérature sur le sujet tend à montrer que les expressions non vues à l'apprentissage sont très mal identifiées par ces modèles, même si cela peut dépendre du type d'expression. On pourra se référer par exemple aux résultats de la compétition internationale PARSEME en 2018 sur l'identification des expressions verbales dans une vingtaine de langues (Ramisch et al. 2018). Le meilleur système utilisant juste des données annotées, atteint péniblement les 20% de succès sur les expressions non couvertes par les données d'apprentissage. Une solution pour palier ce problème est de combiner les données annotées avec des ressources lexicales à large couverture. Dans nos travaux, nous avons mis au point une méthode générique pour combiner ces deux types de ressources (Constant ; Sigogne 2011). Le principe est le suivant : lors de l'apprentissage du modèle qui se fait à partir des données annotées, les ressources lexicales sont consultées à l'aide de la méthode décrite dans la section 5 et permettent d'indiquer où se trouvent les expressions potentielles. Lors de l'apprentissage, les modèles vont apprendre à faire confiance ou pas à la ressource lexicale selon le contexte.

Par exemple, cette approche est particulièrement utile dans les cas d'ambiguïté. Supposons que l'expression *bien que* se trouve dans un dictionnaire de formes composées fléchies. Il existe des contextes où la séquence *bien que* correspond à une co-occurrence accidentelle comme dans l'exemple *J'aime bien que tu viennes* où l'adverbe *bien* pourrait être remplacé par l'adverbe *beaucoup* sans changer le sens. Alors que dans d'autres contextes, il faut faire confiance au dictionnaire, comme dans l'exemple *J'aime le chorizo bien que cela soit épicé*.

Cette approche hybride est aussi utile pour identifier les expressions non vues à l'apprentissage, mais qui se trouve dans la ressource lexicale. En effet, suivant le contexte, le modèle apprend à faire confiance ou pas aux ressources lexicales. Cette méthode a notamment été appliquée avec succès sur le français (Constant ; Sigogne 2011) et le serbe (Constant et al 2018) avec un modèle dit

linéaire (champs aléatoires conditionnels) pour des expressions polylexicales continues. Ce qu'il en ressort est que les performances des outils sont bien meilleures avec des ressources lexicales que sans, notamment dans des textes appartenant à des domaines thématiques éloignés des données annotées d'apprentissage, où la proportion d'expressions non vues dans le corpus d'apprentissage est importante. Une future direction de recherche pourrait consister à explorer la robustesse de cette approche en passant à des modèles neuronaux plus récents.

7 Le projet PARSEME-FR

Toutes les techniques testées pour combiner des ressources lexicales et des corpus annotés en expressions polylexicales se sont révélées très intéressantes, mais elles étaient limitées à quelques types d'expressions. L'étape suivante était de passer à l'échelle. C'est l'objectif principal du projet PARSEME-FR[2] financé par l'Agence Nationale de la Recherche. Il a commencé en 2016 et doit se terminer en mars 2021. Il regroupe cinq partenaires académiques: deux laboratoires de linguistique (Analyse et Traitement Informatique de la Langue Française [ATILF] et Laboratoire de Linguistique Formelle [LLF]) et trois laboratoires d'informatique (Laboratoire d'informatique Fondamentale et Appliquée de Tours [LIFAT], Laboratoire d'Informatique Fondamentale d'Orléans [LIFO] et Laboratoire d'Informatique et Systèmes [LIS]). A noter que PARSEME-FR est un projet dérivé de l'action européenne COST PARSEME[3] (2013 - 2017) sur ces mêmes thèmes. Ce projet s'intéresse à la fois à des aspects informatiques et linguistiques du problème. Le premier objectif est de mettre au point des méthodes et des outils d'identification, (1) en explorant et appliquant les approches neuronales récentes, (2) en étudiant son articulation avec l'analyse syntaxique automatique, tout en essayant de tenir compte au mieux des variations. Le deuxième objectif est plus linguistique et consiste à produire des ressources linguistiques à vaste couverture pour le français: des corpus annotés en expressions polylexicales en couvrant tous les types d'expressions, et l'extraction d'un lexique structuré pour le TAL.

Concernant les données annotées, deux campagnes d'annotation ont été réalisées durant le projet en s'appuyant sur des critères formels opérationnels. La première campagne a consisté à annoter les données françaises de la compétition internationale PARSEME sur l'identification des expressions polylexicales verbales. Le corpus comporte près de 20 000 phrases et est formé de deux sous-corpus: le premier correspond au corpus initial du français de *Universal Dependencies* (Nivre et al. 2016) qui est une initiative internationale pour constituer des corpus annotés en syntaxe de dépendance; le second sous-corpus correspond au corpus Sequoia (Candito ; Seddah 2012) comportant un peu plus de 3 000 phrases. A noter que ce deuxième sous-corpus est un corpus de référence du français, annoté en syntaxe profonde. Sur le corpus global, ont été annotées près de 5

2 https://parsemefr.lis-lab.fr
3 https://typo.uni-konstanz.de/parseme/

000 occurrences d'expressions verbales en suivant scrupuleusement le guide d'annotation basé sur des arbres de décisions fondées elles-mêmes sur des critères opératoires (Savary et al. 2017) . Il est intéressant de noter qu'en moyenne il y a une occurrence d'expression verbale toutes les quatre phrases. Chacune de ces annotations ont été catégorisées en quatre classes : les expressions idiomatiques, les verbes intrinsèquement pronominaux, les constructions à verbe support et les autres expressions verbales[4]. Pour plus de détails, nous invitons le lecteur à se référer à (Candito et al. 2017). P our la deuxième campagne d'annotation, tous les types d'expressions polylexicales ont été annotées, mais sur un plus petit corpus, le Sequoia. Les annotations verbales ont été reprises de la campagne précédente. A noter aussi que nous avons également annoté les entités nommées contenant un seul ou plusieurs éléments tel que *Bilbao* ou *Los Angeles*. Le corpus contient un peu plus de 6 500 annotations pour environ 3 000 phrases. Parmi ces annotations, un peu moins de la moitié sont des entités nommées. Les données et guides d'annotation sont librement disponibles sur le site web du projet PARSEME-FR.

8 Conclusion

Le traitement des expressions polylexicales est fondamental pour le traitement automatique des langues. Les ressources y jouent un rôle crucial: en particulier, les corpus annotés et les ressources lexicales produites par des linguistes. Dans cet article, nous avons décrit plusieurs approches simples que nous avons développées dans nos travaux et qui utilisent ces deux types de ressources linguistiques montrant une complémentarité certaine. Nous avons terminé l'article en présentant rapidement le projet PARSEME-FR dédié aux expressions polylexicales pour le français, qui a produit de nouvelles données linguistiques et des outils d'identification automatique librement disponibles.

9 Remerciements

Ce travail a été partiellement financé par le projet PARSEME-FR lui-même financé par l'Agence Nationale de la Recherche (*ANR-14-CERA-0001*). L'auteur de cet article remercie très chaleureusement les coordinateurs locaux du projet pour leur investissement : Marie Candito, Yannick Parmentier, Carlos Ramisch et Agata Savary.

10 Bibliographie

ABEILLÉ, Anne, CLÉMENT, Lionel, TOUSSENEL, François (2003) « Building a Treebank for French », in *Treebanks: Building and Using Parsed Corpora.* Dordrecht: Kluwer Academic Publishers, 165-187.

AL SAIED, Hazem (2019) *Analyse automatique par transitions pour l'identification des*

4 La proportion des autres expressions verbales dans le corpus est négligeable par rapport aux trois autres classes.

expressions polylexicales. Thèse de doctorat, Université de Lorraine.

BALDWIN, Timothy, KIM, Su Nam (2010) « Multiword Expressions », in *Handbook of Natural Language Processing. Second Edition.* Boca Raton: CRC Press , 267–292.

CANDITO, Marie, SEDDAH, Djamé (2012) « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical », in *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, Grenoble: ATALA/AFCP, 321-334.

CANDITO, Marie, CONSTANT, Mathieu, RAMISCH, Carlos, SAVARY, Agata, PARMENTIER, Yannick, PASQUER, Caroline, ANTOINE, Jean-Yves (2017) « Annotation d'expressions polylexicales verbales en français », in *Actes de la 24e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017) : articles courts.* Orléans: France, 1-9.

CONSTANT, Matthieu, SIGOGNE, Anthony, WATRIN, Patrick (2013) « Stratégies discriminantes pour intégrer la reconnaissance des mots composés dans un analyseur syntaxique en constituants », *Traitement Automatique des Langues,* 54:1 (2013).

CONSTANT, Mathieu , KRSTEV, Cvetana, VITAS, Dusko (2018) « Lexical Analysis of Serbian with Conditional Random Fields and Large-Coverage Finite-State Resources », in *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2015* , Lecture Notes in Computer Science, 10930. Cham: Springer.

CONSTANT, Mathieu, ERYIGIT, Gülşen, MONTI, Johanna, VAN DER P LAS, Lonneke, RAMISCH, Carlos, ROSNER, Michael, TODARISCU, Amalia (2017) « Multiword Expression Processing: A Survey », *Computational Linguistics,* 43:4 (2017), 837–892.

CONSTANT, Matthieu, SIGOGNE, Anthony (2011) « MWU-aware Part-of-Speech Tagging with a CRF model and lexical resources », in *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World.* Portland: Association for Computational Linguistics, 49-56.

COURTOIS, Blandine, GARRIGUES, Mylène, GROSS, Gaston, GROSS, Maurice, JUNG, René, MATHIEU-COLAS, Michel, MONCEAUX, Anne, PONCET-MONTANGE, Anne, SILBERZTEIN, Max, VIV ÈS, Robert (1997) *Dictionnaire électronique DELAC : les mots composés binaires,* Technical Report, 56. LADL.

EISENSTEIN, Jacob (2019) *Introduction to Natural Language Processing*. MIT Press.

GROSS, Maurice (1982) « Une classification des phrases "figées" du français », *Revue québécoise de linguistique,* 11:2 (1982), 151-185.

GROSS, Maurice (1986) « Lexicon Grammar. The Representation of Compound Words », in *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*. Bonn: Association for Computational linguistics, 1-6.

LAPORTE, Éric, NAKAMURA, Takuya, VOYATZI, Stavroula (2008) « A French Corpus Annotated for Multiword Nouns », in *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*. Marrakech, 27-30.

NIVRE, Joakim, DE MARNEFFE, Marie-Catherine, GINTER, Filip, GOLDBERG, Yoav, HAJI Č, Jan, MANNING, Christopher D., MCDONALD, Ryan, PETROV, Slav, PYYSALO, Sampo, SILVEIRA, Natalia, TSARFATY, Reut, ZEMAN, Daniel (2016) « Universal Dependencies v1: A Multilingual Treebank Collection », in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož: European Language Resources Association (ELRA), 1659-1666.

RAMISCH, Carlos, CORDEIRO, Silvio R., SAVARY, Agata, VINCZE, Veronika, BARBU MITITELU, Verginica, BHATIA, Archna, BULJAN, Maja, CANDITO, Marie, GANTAR, Polona, GIOULI, Voula, GÜNGÖR, Tunga, HAWARI, Abdelati, IÑURRIETA, Uxoa, KOVALEVSKAITE, Jolanta, KREK, Simon, LICHTE, Timm, LIEBESKIND, Chaya, MONTI, Johanna, PARRA ESCARTÍN, Carla, QASEMIZADEH, Behrang, RAMISCH, Renata, SCHNEIDER, Nathan, STOYANOVA, Ivelina, VAIDYA, Ashwini, WALSH, Abigail (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)* . Santa Fe:Association for Computational Linguistics, 222-240.

ROSÉN, Victoria, SMØRDAL LOSNEGAARD, Gyri, DE SMEDT, Koenraad, BEJČEK, Eduard, SAVARY, Agata, PRZEPIÓRKOWSKI, Adam, OSENOVA, Petya, Verginica BARBU MITETELU (2015) « A survey of multiword expressions in treebanks », in *Proc. of the 14th International Workshop on Treebanks & Linguistic Theories Conference*.

SAG, Ivan A., BALDWIN, Timothy, BOND, Francis, COPESTAKE, Ann A., FLICKINGER Dan (2002) « Multiword Expressions: A Pain in the Neck for NLP », in *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002*. Springer, 1-15.

SAVARY, Agata (2009) « Multiflex: a Multilingual Finite-State Tool for Multi-Word Units », in *Proceedings of the Conference on Implementation and Application of Automata.* Sydney, 237-240.

SAVARY, Agata, CORDEIRO, Silvio R., RAMISCH, Carlos (2019) « 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) . Florence: Association for Computational Linguistics, 79-91.

SAVARY, Agata, R AMISCH, Carlos, CORDEIRO, Silvio R., SANGATI, Federico, VINCZE, Veronika, QASEMIZADEH, Behrang, CANDITO, Marie, CAP, Fabienne, GIOULI, Voula, STOYANOVA, Ivelina, DOUCET, Antoine (2017) « The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions », in *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017): shared task track* . Valencia: Association for Computational Linguistics, 31-47.

SAVARY, Agata, WASZCZUK, Jakub (2017) « Projecting Multiword Expression Resources on a Polish Treebank », in *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017).* Valencia: Association for Computational Linguistics, 20-26.

SCHNEIDER, Nathan, HOVY, Dirk, JOHANNSEN, Anders, CARPUAT, Marine (2016) « SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM) », in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* , San Diego: Association for Computational Linguistics, 546-559.

SILBERZTEIN, Max (1993) *Dictionnaires électroniques et analyse automatique de textes - Le système Intex*. Masson.

TUTIN, Agnès, ESPERANÇA-RODIER, Emmanuelle, IBORRA, Manolo, REVERDY, Justine (2016) « Annotation of multiword expressions in French », in *EUROPHRAS 2015 - Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives.* Geneva: Editions Tradulex, 60–67.