Une plate-forme générique et ouverte pour l'acquisition des expressions polylexicales

Carlos Ramisch
LIG-GETALP, Grenoble, France
INF-UFRGS, Porto Alegre, Brésil
Carlos.Ramisch@imag.fr

RÉSUMÉ

Cet article présente et évalue une plate-forme ouverte et flexible pour l'acquisition automatique d'expressions polylexicales (EPL) à partir des corpus monolingues. Nous commençons par une motivation pratique suivie d'une discussion théorique sur le comportement et les défis posés par les EPL dans les applications de TAL. Ensuite, nous décrivons les modules de notre plate-forme, leur enchaînement et les choix d'implémentation. L'évaluation de la plate-forme a été effectuée à travers une applications : la lexicographie assistée par ordinateur. Cette dernière peut bénéficier de l'acquisition d'EPL puisque les expressions acquises automatiquement à partir des corpus peuvent à la fois accélérer la création et améliorer la qualité et la couverture des ressources lexicales. Les résultats prometteurs encouragent une recherche plus approfondie sur la manière optimale d'intégrer le traitement des EPL dans de nombreuses applications de TAL, notamment dans les systèmes traduction automatique.

ABSTRACT

An Open and Generic Framework for the Acquisition of Multiword Expressions

In this paper, we present and evaluate an open and flexible methodological framework for the automatic acquisition of multiword expressions (MWEs) from monolingual textual corpora. We start with a pratical motivation followed by a theoretical discussion of the behaviour and of the challenges that MWEs pose for NLP applications. Afterwards, we describe the modules of our framework, the overall pipeline and the design choices of the tool implementing the framework. The evaluation of the framework was performed extrinsically based on an application : computer-assisted lexicography. This application can benefit from MWE acquisition because the expressions acquired automatically from corpora can both speed up the creation and improve the quality and the coverage of the lexical resources. The promising results of previous and ongoing experiments encourage further investigation about the optimal way to integrate MWE treatment into NLP applications, and particularly into machine translation systems.

MOTS-CLÉS : Expressions polylexicales, extraction lexicale, lexique, mesures d'association, corpus, lexicographie.

KEYWORDS: Multiword expression, lexical extraction, lexicon, association measures, corpus, lexicography.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 137–149,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Le terme *expression polylexicale* (EPL, en anglais *multiword expression*) comprend un grand nombre de phénomènes linguistiques qui engendrent des constructions variées telles que les expressions idiomatiques (*payer les yeux de la tête*), les expressions figées (*a priori*), les noms composés (*appareil photo*), les constructions à verbe support (*rendre visite*), etc. Il n'existe pas une définition unique et communément acceptée pour le terme *expression polylexicale*, car il peut être défini comme une « combinaison arbitraire et récurrente de mots » (Smadja, 1993) ou « une unité syntaxique et sémantique dont le sens exact ou la connotation ne peuvent pas être dérivés directement et sans ambiguïté du sens ou de la connotation de ses composantes » (Choueka, 1988) ou simplement comme une « interprétation idiosyncrasique qui dépasse la limite du mot (ou les espaces) » (Sag *et al.*, 2002), avec la propriété qu'elle « doit être répertoriée dans un lexique » (Evert, 2004, p 17.).

La tendance que les mots ont à s'attirer mutuellement, c'est-à-dire le phénomène clé derrière le concept d'EPL, a lieu dans la zone floue entre le lexique et la grammaire. Cela constitue un véritable défi pour les systèmes de TAL classiques. De plus, les EPL sont omniprésentes dans une langue, figurant fréquemment dans le langage oral et écrit de tous les jours, ainsi que dans les communications spécialisées techniques et scientifiques. Parmi les caractéristiques notables des EPL décrites dans la littérature, les plus importantes sont :

- **Caractère arbitraire** : Parfois, des constructions syntaxiquement et sémantiquement valides ne sont pas du tout naturelles simplement parce que les locuteurs natifs de la langue ne les utilisent pas. Même si ces constructions sont tout à fait compréhensibles, elles paraissent étranges et constituent des marqueurs d'un usage non natif. Cela rend l'apprentissage des EPL difficile pour les apprenants d'une langue qui, malgré leur connaissance du lexique et des règles grammaticales générales, n'ont pas assez d'expérience sur l'usage de cette langue. Smadja (1993, p. 143 à 144) illustre cela en présentant huit façons différentes de se référer à l'indice Dow Jones de la bourse de New York, dont seulement quatre sont acceptables.
- **Institutionnalisation** : Les EPL sont récurrentes, car elles correspondent à des façons habituelles de s'exprimer. Jackendoff (1997) estime qu'elles correspondent à la moitié des entrées du lexique d'un locuteur natif. Sag *et al.* (2002) remarquent que cela pourrait être une sous-estimation si l'on prend en compte les EPLs dans les domaines spécialisées, où elles constituent le noyau des connaissances exprimées et représentées.
- **Non-compositionnalité** : Le sens de l'expression entière ne peut pas toujours être déduit directement des sens de ses parties. Par conséquent, la compositionnalité des EPL varie dans un continuum allant des expressions complètement compositionnelles (*appareil photo*) à celles qui sont complètement opaques/idiomatiques (*casser sa pipe*).
- **Hétérogénéité** : Les EPLs sont difficiles à définir car elles englobent une grande quantité de phénomènes. Cela les rend difficiles à traiter par les applications de TAL, qui ne peuvent pas utiliser une approche unifiée. Souvent, les applications de TAL utilisent une des multiples typologies ou schémas de classification existants [1].
- **Non-substituabilité** : Il n'est pas possible de remplacer une partie d'une EPL par un mot proche ou équivalent (synonyme, hyperonyme, etc). Cette propriété motive la notion d'*anti-collocation* (Pearce, 2001), qui correspond à une combinaison de mots maladroite ou inhabituelle (*café corsé* vs *?café consistant*).

1. Par exemple, Smadja (1993) classe les EPL selon leur fonction syntaxique dans la phrase alors que Sag *et al.* (2002) les classent en fonction de leur degré de flexibilité (syntaxique).

$\texttt{en}_{\text{SRC}}$	*I* **paid** *my poor parents* **a visit**
$\texttt{pt}_{\text{TA}}$	*Eu* **pago** *os meus pais pobres* **uma visita**
$\texttt{pt}_{\text{REF}}$	*Eu* **fiz uma visita** *aos meus pobres pais*
$\texttt{fr}_{\text{TA}}$	*J'ai* **payé** *mes pauvres parents* **une visite**
$\texttt{fr}_{\text{REF}}$	*J'ai* **rendu visite** *à mes pauvres parents*
$\texttt{en}_{\text{SRC}}$	*Students pay* **an arm and a leg** *to park on campus*
$\texttt{pt}_{\text{TA}}$	*Estudantes pagam* **braço e uma perna** *para estacionar no campus*
$\texttt{pt}_{\text{REF}}$	*Estudantes pagam* **os olhos da cara** *para estacionar no campus*
$\texttt{fr}_{\text{TA}}$	*Les étudiants paient* **un bras et une jambe** *pour se garer sur le campus*
$\texttt{fr}_{\text{REF}}$	*Les étudiants paient* **les yeux de la tête** *pour se garer sur le campus*
$\texttt{en}_{\text{SRC}}$	*It shares the* **translation-invariance and homogeneity properties** *with the central moment*
$\texttt{pt}_{\text{TA}}$	*Ele compartilha a* **tradução invariância e propriedades de homogeneidade** *com o momento central*
$\texttt{pt}_{\text{REF}}$	*Ele compartilha as* **propriedades de invariância por translação e de homogeneidade** *com o momento central*
$\texttt{fr}_{\text{TA}}$	*Il partage la* **traduction-invariance et propriétés d'homogénéité** *avec le moment central*
$\texttt{fr}_{\text{REF}}$	*Il partage les* **propriétés d'invariance par translation et d'homogénéité** *avec le moment central*

TABLE 1 – Phrases contenant des EPL qui posent problème pour un système de TA empirique.

– **Lexicalisation** : Quelque part dans les applications de TAL, l'information, indiquant qu'un ensemble ou séquence de mots est « indissociable », doit être disponible. Les concepteurs du système doivent donc choisir l'endroit où chaque EPL sera représentée : on ne peut pas les énumérer une à une dans le lexique (sous-génération), ni toutes les inclure dans les règles de la grammaire comme des combinaisons libres (sur-génération). Identifier le degré de lexicalisation de chaque (classe d') EPL est important pour toutes les tâches d'analyse et de génération en TAL.

Dans ce travail, nous adoptons la définition proposée par Calzolari *et al.* (2002), qui définissent les EPL comme :

> ... différents phénomènes liés qui peuvent être décrits comme une séquence [ou groupe] de mots [2] à voir comme une unité à un certain niveau d'analyse linguistique.

Cette définition générique et volontairement vague peut être instanciée selon les besoins des applications. Par exemple, le tableau 1 [3] montre des erreurs générées par un système de traduction automatique (TA) empirique. Pour ce système, une EPL est une séquence de mots qui doit être traduite comme une unité. Sinon, le système génèrera des erreurs, c'est-à-dire des constructions agrammaticales ou artificielles (phrase 1), des traductions littérales maladroites d'expressions idiomatiques (phrase 2) et des mauvais choix lexicaux et syntaxiques dans les textes spécialisés (phrase 3). Dans un système de TA experte, ces EPL seraient typiquement représentées comme

2. Cette définition est réductrice car elle ne considère que les séquences de mots. Nous étendons la notion de séquence vers des groupes de mots, c'est-à-dire des séquences contigües mais aussi des groupes de mots non adjacents liés syntaxiquement et/ou sémantiquement par leur contexte d'occurrence.

3. Source en anglais ($\texttt{en}_{\text{SRC}}$) à partir du web. Traductions automatiques (TA) en portugais (`pt`) et en français (`fr`) fournies par Google Translate (`http://translate.google.com/`) le 18 février 2012. Traductions de référence (REF) fournies par des locuteur natifs.

une entrée à part entière dans le lexique (ou plus précisément, dans le dictionnaire de tournures), sans quoi les mêmes erreurs pourraient survenir.

Ces exemples illustrent l'importance de traiter les EPL dans les systèmes de TA. Plus généralement, le traitement d'EPL peut accélérer et aider à éliminer des ambiguïtés dans de nombreuses applications de TAL, par exemple :

- **Lexicographie** : Church et Hanks (1990) utilisent un environnement lexicographique comme cadre d'évaluation, en comparant la recherche manuelle et intuitive avec la méthode automatique proposée.
- **Reconnaissance optique de caractères** : Supposons qu'un système de reconnaissance optique de caractères a autant de chances de reconnaître les mots *poule* et *poêle* dans *poule/poêle élevée en plein air*. La connaissance d'une EPL utilisant la première option l'aide à choisir [4].
- **Désambiguïsation lexicale** : Les EPL ont tendance à être moins polysémiques que leurs composantes isolées. Finlayson et Kulkarni (2011) illustrent que le mot *world* possède neuf significations possibles dans WordNet 1.6, *record* en possède quatorze, mais *world record* en a seulement une.
- **Étiquetage morpho-syntaxique et analyse syntaxique** : des publications en analyse et étiquetage morpho-syntaxique indiquent que les EPL peuvent aider à éliminer les ambiguïtés syntaxiques (Seretan, 2008; Constant et Sigogne, 2011).
- **Recherche d'information** : Lorsqu'une EPL telle que *pop star* est indexée comme une unité dans un système de recherche d'information, la précision du système s'améliore (Acosta *et al.*, 2011).
- **Apprentissage de langues étrangères** : Puisque les EPL sont très difficiles à apprendre pour des locuteurs non natifs, les dictionnaires et les ressources pédagogiques contenant des entrées polylexicales peuvent être très utiles dans l'enseignement des langues. Un exemple d'une telle ressource est le dictionnaire italien de collocations décrit par Spina (2010).
- **Traduction automatique** : les expériences montrent que l'inclusion d'EPL dans les systèmes de TA améliore la qualité de la traduction (Carpuat et Diab, 2010; Stymne, 2011).

Cet article porte sur le traitement des EPL, allant de l'acquisition automatique à leur intégration dans des applications. Dans un premier temps, nous présentons un bref tour d'horizon de la bibliographie en acquisition automatique d'EPL (§ 2). Dans un deuxième temps, nous décrivons le modèle conceptuel et la plate-forme logicielle développée pour l'acquisition des EPL (§ 3). Par la suite, nous présentons la validation de cette plate-forme dans le cadre de la lexicographie assistée par ordinateur (§ 4). Les expériences en cours montrent des résultats prometteurs mais des enquêtes supplémentaires seront nécessaires pour mieux comprendre les apports des EPL aux systèmes de TAL et en particulier aux systèmes de TA (§ 5).

2 État de l'art

Plusieurs projets ont eu pour objectif de compiler manuellement des ressources lexicales comprenant des EPL. Par exemple, pour le français, le LADL constitue depuis de nombreuses années des listes de noms composés compatibles avec le formalisme/outil Unitex (Gross, 1986). Ce formalisme, cependant, présente des avantages et des inconvénients en ce qui concerne le

4. En fait, cela se fait souvent à l'aide des modèles de langage à *n*-grammes, mais les *n*-grammes ne peuvent pas modéliser adéquatement tous les types d'EPL. Prenons l'exemple d'une expression très flexible telle que **take** *patient risk factors and comfort* **into account**.

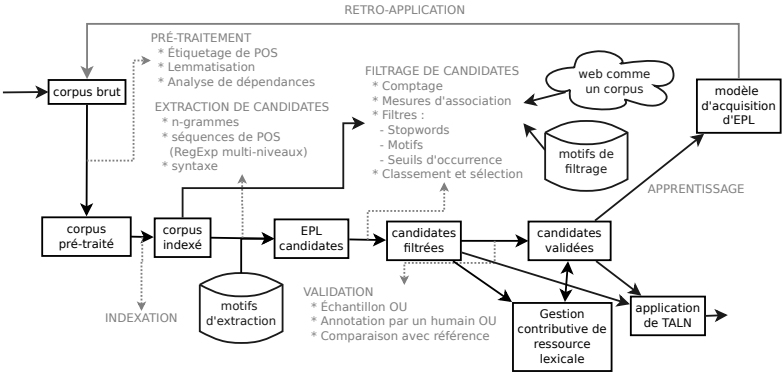

FIGURE 1 – Schéma du mwetoolkit — plate-forme d'acquisition d'EPL à partir de corpus.

compromis entre pouvoir d'expression et temps d'apprentissage (Graliński *et al.*, 2010). La compilation manuelle de ressources lexicales étant très onéreuse, la recherche des dernières années s'est concentrée sur l'acquisition automatique des EPL, dans le but d'accélérer le travail des lexicographes.

Parmi les premiers travaux développant des méthodes automatiques d'identification d'EPL, il y a celui de Smadja (1993). Il a proposé et développé l'outil Xtract pour l'extraction de collocations à partir de textes à travers une combinaison de *n*-grammes et d'une mesure d'information mutuelle. Sur des textes communs, Xtract a une précision de l'ordre de 80%. Depuis lors, de nombreux progrès ont été réalisés, soit sur l'extraction d'EPLs en général (Dias, 2003), soit en se concentrant sur un type d'EPL spécifique, tel que les collocations (Pearce, 2002), les verbes à particule (Ramisch *et al.*, 2008) et les noms composés (Keller et Lapata, 2003).

Une approche indépendante du type devenue très populaire consiste à utiliser des mesures d'association (Evert et Krenn, 2005), qui ont été appliquées avec des degrés variables de succès. Un des avantages de cette approche est qu'elle est indépendante de la langue. Ceci est particulièrement important car les travaux sur l'anglais prédominent (Pearce, 2002; Ramisch *et al.*, 2008), même si des travaux sur les EPL dans plusieurs langues ont été publiés, par exemple Dias (2003) pour le portugais, Evert et Krenn (2005) pour l'allemand et de Cruys et na Villada Moirón (2007) pour le néerlandais.

3 Acquisition d'EPL

Nous proposons une nouvelle plate-forme, décrite dans la figure 1. Elle intègre de multiples techniques parmi celles citées ci-dessus et couvre l'ensemble du pipeline de l'acquisition d'EPL. Cette plate-forme a été mise en œuvre dans un outil libre appelé mwetoolkit [5]. Ici, nous ne résumons que les aspects principaux de la plate-forme, qui a été décrite précédemment dans

5. `http://mwetoolkit.sourceforge.net/`

d'autres publications (Ramisch *et al.*, 2010a,b; Araujo *et al.*, 2011).

Le point de départ de l'extraction est un corpus monolingue de texte brut. Avant l'application du mwetoolkit, il est possible de prétraiter le corpus, à condition que les outils de prétraitement soient disponibles pour la langue cible, en l'enrichissant avec des étiquettes morpho-syntaxiques, des lemmes et de la syntaxe de dépendance. Pour simplifier la représentation des données, nous avons choisi d'utiliser seulement les informations d'analyse syntaxique qui peuvent être représentées comme un attribut du mot, excluant ainsi toute forme d'arborescence. Néanmoins, il est possible de représenter les relations syntaxiques de dépendance à travers une paire ⟨ type de relation, mot père ⟩ qui attribue à chaque mot le mot dont il dépend et l'étiquette qui décrit le type de relation (par exemple, objet direct, modificateur, sujet).

Ensuite, à l'aide des connaissances linguistiques d'experts, de l'intuition, de l'observation empirique et/ou des exemples, il faut décrire les EPL cibles à travers des motifs multiniveaux dans un formalisme similaire aux expressions régulières. Il est possible d'utiliser plusieurs niveaux d'analyse simultanément, par exemple : on veut extraire tous les noms qui sont l'objet direct du verbe *cuisiner*. L'application de ces motifs sur un corpus indexé génère une liste d'EPL candidates.

Pour le filtrage, un grand nombre de méthodes est disponible, allant de simples seuils de fréquence à des listes de mots interdits (*stopwords*), en passant par des mesures d'association plus sophistiquées. Les mesures d'association disponibles (score t, coefficient de Dice, information mutuelle et rapport de vraisemblance) sont, à l'exception de la dernière, applicables à des candidates de longueur arbitraire, ce qui n'est pas toujours le cas dans les outils de linguistique de corpus traditionnels.

Finalement, les candidates résultantes filtrées sont soit directement injectées dans une application de TAL, soit validées manuellement avant l'application. Il est possible de réutiliser les candidates pour l'apprentissage d'un modèle. Ce modèle sera appliqué sur de nouveaux corpus afin d'identifier et d'extraire automatiquement de nouvelles EPL selon les caractéristiques de celles acquises auparavant.

À ce jour, il n'y a pas de consensus sur une méthode optimale d'acquisition d'EPL. Il n'est donc pas possible de déterminer s'il existe une méthode unique pour toutes les EPL, ou alors s'il faudrait chercher une combinaison de méthodes ou un sous-ensemble de méthodes qui fonctionne mieux pour un type d'EPL en particulier. La plupart des travaux récents se concentrent sur l'extraction d'EPL à partir de corpus prétraités (Seretan, 2008) et sur le filtrage automatique et le tri grâce à des mesures d'association (Evert, 2004; Pecina, 2010), mais peu d'auteurs fournissent une vue d'ensemble de la chaîne de traitement d'EPL.

Un des avantages de la plate-forme et de l'outil proposés dans cet article est qu'ils modélisent le processus d'acquisition par des tâches modulaires qui peuvent être enchaînés de plusieurs façons. Chaque tâche a de multiples techniques disponibles pour leur accomplissement. Par conséquent, il est hautement personnalisable et permet un paramétrage détaillé selon les types d'EPL cibles, contrairement à des outils similaires tels que NSP [6] et UCS. [7]

De plus, les techniques développées ne dépendent pas d'une longueur fixe d'expressions candidates (par exemple, les paires de mots) ni sur l'hypothèse de contiguïté. Grâce à cette souplesse, cette méthodologie peut être facilement appliquée à un grand nombre de langues, de types

6. http://search.cpan.org/dist/Text-NSP
7. http://www.collocations.de/software.html

d'EPL et de domaines, ne dépendant pas d'un formalisme donné ou d'un outil. [8] Pour une langue donnée, si certains outils de prétraitement tels que les étiqueteurs morpho-syntaxiques et/ou analyseurs sont disponibles, il n'y a pas de raison pour ne pas s'en servir. Dans ce cas, intuitivement, les résultats seront bien meilleurs que sur du texte brut non analysé. Mais comme toutes les langues ne disposent pas de ces outils, la méthodologie a été conçue pour être appliquée même dans le contexte où aucun outil de prétraitement n'est disponible. Dans l'avenir, nous voudrions valider cette hypothèse en réalisant l'extraction d'EPL sur un corpus dans une langue pauvrement dotée.

4 Résultats

Ici, nous présentons les résultats de l'évaluation de notre plate-forme dans le cadre de la lexicographie assistée par ordinateur. Tout d'abord, nous introduisons une nouvelle classification pour les différents axes d'évaluation de l'acquisition automatique d'EPL (§ 4.1). Ensuite, nous résumons l'évaluation quantitative et qualitative extrinsèque de la plate-forme d'acquisition d'EPL proposée précédemment (§ 4.2).

4.1 Évaluation de l'acquisition d'EPL

Comme l'a souligné Pecina (2005), « l'évaluation des méthodes d'extraction de collocations est une tâche complexe. D'une part, les différentes applications exigent différents [...] seuils. D'autre part, les méthodes donnent des résultats distincts selon les intervalles de leurs scores d'association ». Nous structurons l'évaluation de l'acquisition d'EPL selon les critères suivants :

- **Selon la nature des mesures :**
  - **Quantitative** : consiste à évaluer l'acquisition à travers des mesures objectives telles que la précision, le rappel, la F-mesure et la précision moyenne. Alors que de nombreux articles calculent uniquement la précision sur les premières *n* EPL retournées, il faut aussi évaluer le rappel, ce qui est rarement fait. Néanmoins, cela est d'une importance capitale dans l'attribution de l'utilité d'une méthode. Une méthode très précise qui n'extrait qu'une douzaine d'expressions quand il y a en réalité des milliers d'expressions à récupérer n'est pas plus efficace que la force brute ou la recherche manuelle. La quantité d'EPL découvertes est un facteur aussi important que leur qualité, et il est difficile d'évaluer combien d'EPL sont « suffisantes » pour que l'acquisition automatique soit utile (Villavicencio *et al.*, 2005; Church, 2011).
  - **Qualitative** : le but de l'évaluation qualitative est d'obtenir une compréhension approfondie des EPL obtenues et des erreurs commises par la méthode d'acquisition. Cela consiste à observer les motifs récurrents en analysant les listes résultantes en termes de leur adéquation à l'application cible. Cette évaluation est souvent itérative, c'est-à-dire que les améliorations possibles sont retro-appliquées sur la méthode d'acquisition, avec une nouvelle évaluation, et ainsi de suite.
- **Selon l'objectif de l'acquisition :**

8. Toutefois, la méthodologie est conçue pour traiter les langues qui utilisent des espaces pour séparer les mots. Ainsi, lorsque l'on travaille avec du chinois, du japonais, ou même avec des noms composés en allemand, un prétraitement supplémentaire est nécessaire.

Langue	Type	Corpus	Mots	Cand.	EPL	Publication
anglais	VàP	Europarl-frg	13M	5,3K	875	(Ramisch *et al.*, 2012)
grec	NC	Europarl	26M	5K	815	(Linardaki *et al.*, 2010)
portugais	EV	PLN-BR-FULL	29M	407K	773	(Duran *et al.*, 2011)

TABLE 2 – Acquisition d'EPL appliquée à la lexicographie assistée par ordinateur.

  - **Intrinsèque** : la plupart des résultats d'évaluation publiés dans les références bibliographiques citées dans cet article sont intrinsèques, c'est-à-dire, ils considèrent les EPL en elles-mêmes, directement, en tant que produit final d'un processus. Même si elle présente de nombreuses limitations, l'évaluation intrinsèque donne toujours une estimation de qualité qui permet une comparaison inter-méthodes fiable.
  - **Extrinsèque** : il est souvent plus facile d'estimer la qualité du résultat pour une tâche de TAL concrète que pour une liste d'EPL dont on ne connait pas l'application. Ainsi, l'évaluation extrinsèque, c'est-à-dire l'utilisation des EPL dans une application de TAL extérieure, peut être très concluante pour démontrer si les EPL acquises sont utiles.
- **Selon le type d'EPL :**
  - **Fondée sur les types** : certaines expressions ne sont pas ambiguës et peuvent être annotées hors contexte, comme des entrées dans un lexique. C'est souvent le cas quand il s'agit de noms composés, de termes techniques et de constructions à verbe support. La décision de savoir si une séquence de mots est une EPL, dans ce type d'annotation, est indépendante du contexte dans lequel elle apparaît.
  - **Fondée sur les occurrences** : cette annotation doit être effectuée quand les EPL cibles sont ambiguës, comme les verbes à particule et les expressions idiomatiques. Hors contexte, il est impossible de dire si les mots doivent être traités comme une unité ou indépendamment.

D'une part, les résultats de ces *évaluations intrinsèques* sont souvent vagues ou peu concluants. Bien qu'ils fassent la lumière sur les paramètres optimaux pour un scénario donné, ils sont difficiles à généraliser et ne peuvent pas être directement appliquées à d'autres configurations. La qualité des EPL acquises mesurée par des critères objectifs dépend de la langue, du domaine et du type de la construction cible, ainsi que de la taille et du genre du corpus, des ressources déjà disponibles [9], des filtres appliqués, des étapes de prétraitement, ...

D'autre part, l'*évaluation extrinsèque* consiste à insérer les EPL acquises dans des applications de TAL réelles et à évaluer l'impact de ces nouvelles données sur la performance globale du système. Ainsi, une contribution originale du présent travail est l'application de l'évaluation extrinsèque de l'acquisition d'EPL sur une étude de cas : la lexicographie assistée par ordinateur. Notre objectif à long terme est d'étudier (1) quel est l'impact de ces EPL sur les applications de TAL en général et (2) la (ou les) meilleure(s) méthodes(s) pour les intégrer dans le pipeline complexe de l'application cible.

9. Il est inutile d'acquérir des EPL déjà présentes dans le dictionnaire.

4.2 Lexicographie assistée par ordinateur

Nous avons travaillé pour cette évaluation en collaboration avec des collègues linguistes et lexicographes expérimentés, dans le but de créer de nouvelles ressources lexicales contenant des EPL. Les langues des ressources sont l'anglais, le grec et le portugais. Le tableau 2 résume les résultats de chaque évaluation.

Nous avons extrait les verbes à particule (VàP) à partir d'un fragment de la partie anglaise du corpus Europarl. [10] Nous avons considéré un VàP comme étant formé par un verbe (à l'exception de *be* et *have*) suivi d'une particule prépositionnelle [11] éloignée d'au plus 5 mots après le verbe. [12] Nous avons obtenu 5 302 candidates à VàP qui apparaissent plus d'une fois dans le corpus. L'évaluation de ces candidates a été effectuée de façon automatique, en les comparant avec un dictionnaire de référence. Parmi les candidates, 875 ont été classifiées comme de véritables VàP. Cette évaluation quantitative et fondée sur les types est une première étape d'une expérience en cours sur l'intégration de ces constructions dans un système de traduction automatique.

Pour le grec, il existe une vaste littérature portant sur les propriétés linguistiques des EPL, mais les approches informatiques sont encore limitées (Fotopoulou *et al.*, 2008). Dans nos expériences, nous avons extrait de la partie grecque du corpus Europarl, étiquetée morpho-syntaxiquement, des noms composés (NC) correspondant aux motifs suivants : adjectif-nom, nom-nom, nom-déterminant-nom, nom-préposition-nom, préposition-nom-nom, nom-adjectif-nom et nom-conjonction-nom. Les candidates ont été comptées dans deux corpus et classées par quatre mesures d'association. Les premières 150 candidates selon chaque mesure d'association ont été évaluées intrinsèquement par trois locuteurs natifs. Ainsi, chaque annotateur a jugé environ 1 200 candidates. Finalement, les annotations ont été combinées, entraînant la création d'un lexique avec 815 EPL nominales en grec.

L'objectif du travail avec les expressions verbales (EV) en portugais était de réaliser une analyse qualitative de ces constructions bien comme de la méthode d'acquisition. Nous avons étiqueté morpho-syntaxiquement le corpus PLN-BR-Full [13] et ensuite nous avons effectué quelques itérations d'une phase d'évaluation qualitative par un lexicographe expérimenté suivie d'une nouvelle phase d'extraction. Finalement, nous avons extrait des séquences de mots correspondant aux motifs verbe-[déterminant]-nom-préposition, verbe-préposition-nom, verbe-[préposition/déterminant]-adverbe et verbe-adjectif. Le processus d'extraction a ainsi conduit à une liste de 407 014 candidates qui ont ensuite été filtrées avec des mesures d'association.

Durant l'évaluation quantitative fondée sur les types, un annotateur humain expert a validé manuellement 12 545 candidates, parmi lesquelles 699 ont été annotées comme des EV compositionnelles et 74 comme des EV idiomatiques. Ensuite, une analyse fine de chaque motif d'extraction a été réalisée dans le but de trouver des corrélations entre la flexibilité syntaxique et les propriétés sémantiques telles que la compositionnalité. Les retours fournis par le lexicographe ont montré que l'outil est plus flexible et plus efficace que les concordanciers traditionnels, car il permet d'identifier des EPL candidates sans fixer une liste pré-définie de verbes support. Ainsi, des constructions non prototypiques ont pu être identifiées grâce à l'utilisation de notre plate-forme.

10. `http://statmt.org/europarl`
11. *up, off, down, back, away, in, on.*
12. Même si la particule pourrait apparaître plus loin que 5 mots après le verbe, de tels cas sont suffisamment rares pour être ignorées dans cette expérience.
13. `www.nilc.icmc.usp.br/plnbr`

5 Conclusions et perspectives

Dans cet article, nous avons décrit une plate-forme pour l'acquisition des EPL à partir de corpus monolingues. La contribution principale de ce travail réside dans le fait qu'il représente une étape vers l'intégration des EPL automatiquement extraites dans des applications réelles. Premièrement, nous avons proposé un *cadre méthodologique* unifié, ouvert et flexible pour l'acquisition automatique des EPL. Deuxièmement, nous avons effectué une vaste *évaluation de l'acquisition d'EPL*, afin de disséquer l'influence des différents types de ressources utilisées dans l'acquisition sur la qualité des EPL résultantes. De plus, nous avons proposé une nouvelle taxonomie qui classifie les résultats de l'évaluation d'acquisition d'EPL selon trois axes principaux.

Nous sommes actuellement en train de développer des méthodes pour l'intégration des EPL verbales acquises automatiquement dans un système de TA empirique. Les EPL verbales sont des constructions syntaxiquement flexibles. Par conséquent, elles constituent un défi pour les systèmes de TA empirique fondés sur les séquences de mots. Après quelques expériences préliminaires, nous avons constaté le besoin d'appliquer une méthode identification fondée sur la syntaxe. En même temps, nous analysons les alignements lexicaux et les entrées de la table de traduction du système sur un ensemble d'EPL prototypiques, afin de mieux comprendre l'impact des EPL sur les résultats du système. Enfin, nous voulons réaliser des expériences sur la simplification d'EPL, par exemple, le remplacement d'un verbe polylexical comme *come back* par sa forme simple *regress*. Ainsi, l'intuition qui veut que l'on fasse ressembler la langue source à la langue cible facilite la tâche d'apprentissage d'alignements inter-langue (Stymne, 2011). Comme ces améliorations dépendent du paradigme de TA choisi, nous voulons également évaluer les stratégies pour l'intégration des EPL verbales dans les systèmes de TA experte tels que ITS2 (Wehrli et Ramluckun, 1993) et Etap-3 (Apresian *et al.*, 2003).

En dépit d'un important effort de recherche dans ce domaine, le traitement d'EPL dans les applications de TAL est encore un problème ouvert. Bien sûr, ceci n'est pas vraiment une surprise dans la mesure où les linguistes ont démontré la complexité de ce problème depuis des décennies (Sag *et al.*, 2002). Au début des années 2000, Schone et Jurafsky (2001) ont posé la question si l'identification automatique d'EPL était un problème résolu, et la réponse que cet article apporta à l'époque fut négative. De même, les préfaces du numéro spécial sur les EPL de la revue *Language Resources and Evaluation* (Rayson *et al.*, 2010) et de l'atelier MWE 2011 (Kordoni *et al.*, 2011) indiquent que, d'un point de vue pratique, plusieurs défis sont à relever dans le but d'obtenir des résultats moins artificiels pour les EPL dans les systèmes de TAL. Nous croyons que, à long terme, le présent travail de recherche contribuera à la conception d'applications de TAL qui intègrent pleinement le traitement des EPL comme une étape très importante dans la constitution du lexique et de la grammaire. Néanmoins, étant donné la complexité du problème, ce traitement doit être continuellement amélioré, car ils nous semble peu probable que, dans un avenir proche, on puisse proposer une solution définitive et unifiée pour le traitement des EPL dans les applications de TAL.

Remerciements

Je remercie mes tuteurs Christian Boitet et Aline Villavicencio, ainsi que les collègues qui ont contribué activement à ce travail : Evita Linardaki, Magali Sanchez Duran et Vitor De Araujo. Merci aux réviseurs pour leurs suggestions et à Antoine Gay pour la relecture. Ce travail est financé par une allocation du Ministère de l'Enseignement Supérieur et de la Recherche et par le projet CAMELEON (CAPES-COFECUB 707-11).

Références

Acosta, O., Villavicencio, A. et Moreira, V. (2011). Identification and treatment of multiword expressions applied to information retrieval. *In* (Kordoni *et al.*, 2011), pages 101–109.

Apresian, J., Boguslavsky, I., Iomdin, L. et Tsinman, L. (2003). Lexical functions as a tool of ETAP-3. *In Proc. of the First MTT Conference (MTT 2003)*.

Araujo, V. D., Ramisch, C. et Villavicencio, A. (2011). Fast and flexible MWE candidate generation with the mwetoolkit. *In* (Kordoni *et al.*, 2011), pages 134–136.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., Macleod, C. et Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. *In Proc. of the Third LREC (LREC 2002)*, pages 1934–1940, Las Palmas, Canary Islands, Spain. ELRA.

Carpuat, M. et Diab, M. (2010). Task-based evaluation of multiword expressions : a pilot study in statistical machine translation. *In Proc. of HLT : The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 242–245, Los Angeles, California. ACL.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. *In RIAO'88*, pages 609–624.

Church, K. (2011). How many multiword expressions do people know ? *In* (Kordoni *et al.*, 2011), pages 137–144.

Church, K. W. et Hanks, P. (1990). Word association norms mutual information, and lexicography. *Comp. Ling.*, 16(1):22–29.

Constant, M. et Sigogne, A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. *In* (Kordoni *et al.*, 2011), pages 49–56.

de Cruys, T. V. et na Villada Moirón, B. (2007). Semantics-based multiword expression extraction. *In* Gregoire, N., Evert, S. et Kim, S. N., éditeurs : *Proc. of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007)*, pages 25–32, Prague, Czech Republic. ACL.

Dias, G. (2003). Multiword unit hybrid extraction. *In* Bond, F., Korhonen, A., McCarthy, D. et Villavicencio, A., éditeurs : *Proc. of the ACL Workshop on MWEs : Analysis, Acquisition and Treatment (MWE 2003)*, pages 41–48, Sapporo, Japan. ACL.

Duran, M. S., Ramisch, C., Aluísio, S. M. et Villavicencio, A. (2011). Identifying and analyzing brazilian portuguese complex predicates. *In* (Kordoni *et al.*, 2011), pages 74–82.

Evert, S. (2004). *The Statistics of Word Cooccurrences : Word Pairs and Collocations*. Thèse de doctorat, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany.

EVERT, S. et KRENN, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):450–466.

FINLAYSON, M. et KULKARNI, N. (2011). Detecting multi-word expressions improves word sense disambiguation. *In* (Kordoni *et al.*, 2011), pages 20–24.

FOTOPOULOU, A., GIANNOPOULOS, G., ZOURARI, M. et MINI, M. (2008). Automatic recognition and extraction of multiword nominal expressions from corpora (in greek). *In Proceedings of the 29th Annual Meeting, Department of Linguistics*, Aristotle University of Thessaloniki, Greece.

GRALIŃSKI, F., SAVARY, A., CZEREPOWICKA, M. et MAKOWIECKI, F. (2010). Computational lexicography of multi-word units : How efficient can it be ? *In* (Laporte *et al.*, 2010), pages 1–9.

GROSS, M. (1986). Lexicon - grammar the representation of compound words. *In Proc. of the 11th COLING (COLING 1986)*.

JACKENDOFF, R. (1997). Twistin' the night away. *Language*, 73:534–559.

KELLER, F. et LAPATA, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Comp. Ling. Special Issue on the Web as Corpus*, 29(3):459–484.

KORDONI, V., RAMISCH, C. et VILLAVICENCIO, A., éditeurs (2011). *Proc.of the ACL Workshop on MWEs : from Parsing and Generation to the Real World (MWE 2011)*, Portland, OR, USA. ACL.

LAPORTE, E., NAKOV, P., RAMISCH, C. et VILLAVICENCIO, A., éditeurs (2010). *Proc.of the COLING Workshop on MWEs : from Theory to Applications (MWE 2010)*, Beijing, China. ACL.

LINARDAKI, E., RAMISCH, C., VILLAVICENCIO, A. et FOTOPOULOU, A. (2010). Towards the construction of language resources for greek multiword expressions : Extraction and evaluation. *In* PIPERIDIS, S., SLAVCHEVA, M. et VERTAN, C., éditeurs : *Proc. of the LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*, pages 31–40, Valetta, Malta. May.

PEARCE, D. (2001). Synonymy in collocation extraction. *In WordNet and Other Lexical Resources : Applications, Extensions and Customizations (NAACL 2001 Workshop)*, pages 41–46.

PEARCE, D. (2002). A comparative evaluation of collocation extraction techniques. *In Proc. of the Third LREC (LREC 2002)*, Las Palmas, Canary Islands, Spain. ELRA.

PECINA, P. (2005). An extensive empirical study of collocation extraction methods. *In Proc. of the ACL 2005 SRW*, pages 13–18, Ann Arbor, MI, USA. ACL.

PECINA, P. (2010). Lexical association measures and collocation extraction. *Lang. Res. & Eval. Special Issue on Multiword expression : hard going or plain sailing*, 44(1-2):137–158.

RAMISCH, C., ARAUJO, V. D. et VILLAVICENCIO, A. (2012). A broad evaluation of techniques for automatic acquisition of multiword expressions. *In Proc. of the ACL 2012 SRW*, Jeju, Republic of Korea. ACL.

RAMISCH, C., VILLAVICENCIO, A. et BOITET, C. (2010a). Multiword expressions in the wild ? the mwetoolkit comes in handy. *In* LIU, Y. et LIU, T., éditeurs : *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China. The Coling 2010 Organizing Committee.

RAMISCH, C., VILLAVICENCIO, A. et BOITET, C. (2010b). mwetoolkit : a framework for multiword expression identification. *In Proc. of the Seventh LREC (LREC 2010)*, Malta. ELRA.

RAMISCH, C., VILLAVICENCIO, A., MOURA, L. et IDIART, M. (2008). Picking them up and figuring them out : Verb-particle constructions, noise and idiomaticity. *In* CLARK, A. et TOUTANOVA, K., éditeurs : *Proc. of the Twelfth CoNLL (CoNLL 2008)*, pages 49–56, Manchester, UK. The Coling 2008 Organizing Committee.

RAYSON, P., PIAO, S., SHAROFF, S., EVERT, S. et MOIRÓN, B. V. (2010). Multiword expressions : hard going or plain sailing ? *Lang. Res. & Eval. Special Issue on Multiword expression : hard going or plain sailing*, 44(1-2):1–5.

SAG, I., BALDWIN, T., BOND, F., COPESTAKE, A. et FLICKINGER, D. (2002). Multiword expressions : A pain in the neck for NLP. *In Proc. of the 3rd CICLing (CICLing-2002)*, volume 2276/2010 de *LNCS*, pages 1–15, Mexico City, Mexico. Springer.

SCHONE, P. et JURAFSKY, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem ? *In* LEE, L. et HARMAN, D., éditeurs : *Proc. of the 2001 EMNLP (EMNLP 2001)*, pages 100–108, Pittsburgh, PA USA. ACL.

SERETAN, V. (2008). *Collocation extraction based on syntactic parsing*. Thèse de doctorat, University of Geneva, Geneva, Switzerland.

SMADJA, F. A. (1993). Retrieving collocations from text : Xtract. *Comp. Ling.*, 19(1):143–177.

SPINA, S. (2010). The dictionary of italian collocations : Design and integration in an online learning environment. *In Proc. of the Seventh LREC (LREC 2010)*, Malta. ELRA.

STYMNE, S. (2011). Pre- and postprocessing for statistical machine translation into germanic languages. *In Proc. of the ACL 2011 SRW*, pages 12–17, Portland, OR, USA. ACL.

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. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):365–377.

WEHRLI, E. et RAMLUCKUN, M. (1993). ITS-2 : an interactive personal translation system. *In Proc. of the 6th Conf. of the EACL (EACL 1993)*, page 476, Utrecht, The Netherlands. ACL.