Améliorer l'extraction et la description d'expressions polylexicales grâce aux règles transformationnelles

Aurélie JOSEPH[1,2]
(1) LDI, 99 avenue Jean-Baptise Clément F-93430 Villetaneuse
(2) ITESOFT, Parc d'Andron, le Séquoia, 30470 Aimargues
joseph.aurelie@gmail.com

RÉSUMÉ

Cet article présente une méthodologie permettant d'extraire et de décrire des locutions verbales vis-à-vis de leur comportement transformationnel. Plusieurs objectifs sont ciblés : 1) extraire automatiquement les expressions phraséologiques et en particulier les expressions figées, 2) décrire linguistiquement le comportement des phraséologismes 3) comparer les méthodes statistiques et notre approche et enfin 4) montrer l'importance de ces expressions dans un outil de classification de textes.

ABSTRACT

Enhance Multiword Expressions Extraction and Description with Transformational Rules

This paper presents a methodology to extract and describe verbal multiword expressions using their transformational behavior. Several objectives are targeted: 1) automatically extracting MWE and especially frozen expression, 2) describing linguistically their MWE behavior, 3) comparing statistical methods and our approach, and finally 4) showing the importance of MWE in a text classification tool.

MOTS-CLÉS: expressions polylexicales, expressions figées, locution verbale, extraction, transformation, classification de textes
KEYWORDS : multiword expression, verbal phrase, extraction, transformation, text classification

1 Introduction

Depuis quelques années l'extraction d'expressions phraséologiques (EP) est devenue un problème majeur dans le traitement du langage naturel. Du fait de l'universalité du phénomène à travers les langues, de l'importance des EP dans les corpus et de leur impact dans la compréhension, il devient primordial de décrire et de traiter cet objet linguistique

Notre principal but est de proposer une méthodologie permettant d'extraire ces mots polylexicaux englobant ce qui est connu sous le nom de noms composés (*pomme de terre, lune de miel*), locutions verbales (*prendre en compte, mettre fin*), idiomes (*casser sa pipe, peigner la girafe*), collocations (porter plainte)...

La plupart des méthodologies extraient généralement des expressions qui sont en réalité des éléments terminologiques ou des collocations. Peu d'études utilisent les contraintes transformationnelles qui sont des critères incontournables des EP. Nous voulons démontrer qu'une approche transformationnelle permet de retrouver ces expressions et même de donner des critères permettant de les décrire et ainsi de les catégoriser dans leur degré de figement. De plus, la majorité des approches traitent les expressions nominales (Daille, 1996 ; Watrin, 2007 ; François 2011). Dans cet article, nous nous focalisons sur les EP verbales (Kilgarriff, 2002 ; Smadja, 1993) qui sont plus difficiles à traiter.

2 Etat de l'art

Abordons brièvement les typologies liées aux phraséologismes et aux terminologies associées. Pour plus de détails, nous renvoyons à Mejri (2011), Gross (1996 ; 2012) ou Mel'čuk (2011) pour le français ou encore Abu Ssaydeh (2005) pour l'anglais. Dans notre étude nous nous concentrons sur :

- Les expressions totalement figées. Elles n'acceptent aucune variation, leur sens est souvent opaque et elles sont lexicalisées : *au fur et à mesure.*
- Les expressions semi-figées qui acceptent quelques variations. C'est ici que la plupart des séquences verbales se situent : *prendre une veste, casser sa pipe.*
- Les collocations. Ce sont des expressions qui « aiment » être ensemble (intimer l'ordre) mais dont le comportement syntaxique reste assez libre.
- Les routines langagières (*veuillez agréer mes sincères salutations*).

De nombreuses méthodes sont utilisées pour extraire ces éléments.

2.1 Les approches statistiques

Les mesures probabilistes telles que le rapport de vraisemblance ou la mesure de Dice, sont très souvent utilisées par les chercheurs pour déterminer les termes apparaissant fréquemment ensemble (Sinclair, 1991). Dias (2003) propose également une méthode sans ressource linguistique, utilisable indépendamment de la langue et sans contrainte dans le nombre de mots possibles dans la séquence. Les méthodes statistiques ont l'avantage d'être facilement implémentables, rapides et efficaces dans leur traitement mais laissent souvent de côté les expressions figées (Ramisch 2012) en faveur des

collocations. Elles nécessitent également des corpus volumineux.

2.2 Les approches hybrides

Même si certains chercheurs refusent les ressources linguistiques à cause de ses inconvénients (dépendantes de la langue étudiée, souvent longues à construire et à maintenir), les méthodologies les plus performantes et les plus utilisées, combinent statistique et des filtres linguistiques. Ces filtres peuvent être des nettoyages de mots grammaticaux (Manning et Schütze, 1999), des sélections de structures syntaxiques productives (Watrin, 2007). Quelquefois les chercheurs introduisent les transformations pour étendre leur extraction (Daille, 1996) ou pour vérifier la validité des candidats (Al Haj et Wintner, 2010 ; Abeillé et Schabes, 1989). L'outil mwetoolkit (Ramisch 2012) propose de nombreuses possibilités pour extraire ces expressions selon certains filtres.

2.3 Les approches multilingues

Très brièvement revenons sur les travaux de Villada Moirón et al. (2006) ou Archer (2006) qui proposent une extraction basée sur la comparaison de corpus parallèles. Ils partent du postulat que certaines EP ne sont pas traduisibles mot à mot. En d'autres termes, la traduction de chaque terme ne peut mener à la traduction de l'expression entière dans la langue cible.

2.4 Les approches basées sur des ressources linguistiques

Alors que des listes répertoriant les expressions figées peuvent être utilisées (Grezka et Poudat, 2012), pour reconnaître les expressions semi-figées qui sont les EP les plus fréquentes dans les corpus, la meilleure ressource est celle qui décrit les variations des composants et les transformations possibles. Le lexique-grammaire (initié par M. Gross) et les ressources du LDI (Ben-Henia Ayat, 2006 ; Cartier, 2010 ; Buvet, 2008) décrivent chaque expression de cette manière. Mais la description est très coûteuse en temps de réalisation. C'est pourquoi nous voulons améliorer cette approche en introduisant des descriptions automatiques.

Plus récemment, les méthodes basées sur le Web, ont émergé. Certaines utilisent les moteurs de recherche (Colson, 2010 ; Cartier et Joseph, 2011), d'autres utilisent Google ngrams (François, 2011) ou Wikipédia (Garcia-Fernandez et al., 2011).

3 Le corpus

Dans cette étude, nous utilisons un corpus existant dans notre entreprise composé de lettres écrites par des clients (1 533 documents, 273 669 formes). Ces courriers sont dactylographiés et concernent la relation entre un client et une entreprise télévisuelle. Chaque classe représente un sujet particulier (gestion d'abonnement, offres, annulation simple, annulation complexe, réclamation offre, réclamation financière). Le niveau d'orthographe et de grammaire du client entraîne inévitablement des erreurs qui ne sont pas corrigées. Cependant, la reconnaissance optique des caractères liée à la numérisation des courriers et les erreurs d'orthographes ne représentent que 2% des formes. Ce corpus spécialisé nous permet de nous confronter à des données réelles, problématiques et nous

permet d'appréhender la phraséologie de ce domaine.

4 Approche méthodologique

Notre approche utilise les contraintes transformationnelles liées aux EP (substitutions, insertion de modifieurs, passivation…) afin de les extraire, de les catégoriser automatiquement et de décrire leur comportement. L'extraction et la description sont réalisées à partir d'un corpus.

4.1 Architecture du système

Voici les différents processus pour extraire les phraséologismes :

- Un étiquetage morphosyntaxique permettant d'extraire une liste de candidats dans un corpus à partir de structures syntaxiques.
- La création d'un programme générant pour chaque candidat, les transformations possibles.
- La création d'un programme qui recherche les transformations dans les textes.

4.2 L'extraction par structures syntaxiques

Il a été prouvé depuis quelques années que les EP et en particulier les expressions verbales ne sont pas syntaxiquement déviantes de la langue comme avait pu le postuler Björkman (1978). Au contraire, la plupart des EP correspondent à des structures de la syntaxe libre. Les structures les plus productives (appelées moules syntaxiques) peuvent être listées. Par exemple, pour les noms : Mathieu-Colas (1988) ; pour les verbes : Gross (1982), Schmid (1991), Cartier et Joseph (2011) ; pour les adverbes : Grezka et Poudat (2012).

Dans cet article nous nous limitons aux EP verbales composées d'un seul nom car ce sont les structures les plus productives. Nous étudions donc les structures suivantes : VERBE NOM ; VERBE DETERMINANT NOM ; VERBE PREPOSITION NOM ; VERBE PREPOSITION DETERMINANT NOM. Après avoir taggué et lemmatisé notre corpus avec Treetagger, nous extrayons les EP candidats à partir de ces structures.

4.3 Transformations morphologiques, syntagmatiques et paradigmatiques

Les règles transformationnelles testées sont celles expliquées dans la littérature (Gross 1996, Lamiroy 2008…). Cependant, certaines sont supprimées car trop vagues pour être traitées automatiquement. Parmi elles, mentionnons la pronominalisation.

*Contacter votre service → le contacter ; Faire un geste → *en faire un*

Nous ne testons pas non plus l'insertion d'adverbes entre le verbe et le complément car la majorité des séquences verbales acceptent ce modifieur faisant d'elles des séquences semi-figées. Finalement trois grands types de transformations sont testés.

4.3.1 Les transformations morphologiques

Elles correspondent aux variations des composants, comme :

- La flexion nominale : le nombre du nom (*une étude, des études*)
- La nominalisation

G. Gross (2010) argue que ce critère se limite aux verbes prédicatifs (*résilier un contrat → une résiliation de contrat*). Mais selon M. Gross (1986), certaines expressions figées acceptent la nominalisation (*Mettre en scène → une mise en scène)*.

Ce critère peut toutefois permettre de catégoriser un grand nombre de prédicats. Notons également que la nominalisation peut entraîner des modifications syntagmatiques.

4.3.2 Les transformations syntagmatiques

Elles correspondent à des modifications sur l'axe syntagmatique, liées aux règles d'ordre des mots.

- Clivage : **C'est une fin que je mets à mon contrat*
- Passivation : *?Une fin est mise à mon contrat*
- Relativation : **La fin que je mets à mon contrat*

Les variations syntagmatiques peuvent également être dues à :

- La suppression ou l'insertion de déterminant : *Je mets une fin à mon contrat ; *Je fais point sur cette situation*
- L'insertion de modifieurs : *??se renvoyer la petite balle*
- L'insertion de syntagme entre le verbe et le complément : *?Faire sur la situation le point*

4.3.3 Les transformations paradigmatiques

Ces transformations concernent les substitutions des composants de la séquence avec un composant de même nature (un verbe avec un verbe, un nom avec un nom…). Contrairement aux méthodes statistiques, nous évaluons les substitutions possibles ayant la même structure syntaxique. Par exemple, pour la séquence *mettre fin,* 13 verbes de notre corpus sont substituables avec *mettre* (*avoir en fin, arrêter à la fin, attendre la fin*…). Cependant, en ajoutant une contrainte structurelle (*mettre fin* correspond à la structure VER NOM), seulement 4 sont conservées (*prendre fin*, *donner fin*, *arrêter fin* et *prévenir fin)*.

De plus, nous ne gardons que les substitutions possédant une catégorie grammaticale du contexte droit identique à celles possibles dans la séquence source. *Mettre fin* possède 3 contextes possibles : a) une préposition : mettre fin **à** mon contrat, b) une ponctuation (point, virgule…) : j'y mets fin**.**, c) un adverbe : je mets fin **immédiatement** à mon contrat. Parmi les substitutions précédemment sélectionnées, seules 2 séquences (*prendre fin* et *donner fin*) ont un contexte en commun avec *mettre fin* (ici une préposition : *prendre fin **en** septembre, donner fin **à** mon abonnement*). Ici, nous remarquons que la préposition à elle seule ne suffit pas pour montrer que nous avons deux contextes identiques (l'un complément d'objet indirect, l'autre complément circonstanciel de temps). Il faudrait améliorer l'analyse. Mais nous réduisons déjà un certain nombre de possibilités substitutionnelles.

Enfin, nous ne comptons pas le nombre d'occurrences de chaque substitution dans le corpus, mais le nombre de substitutions différentes. En d'autres termes peu importe le nombre de fois où *prendre fin* ou *donner fin* apparaissent dans le corpus seule compte le nombre de formes différentes substituables (ici 2).

4.3.4 Implémentation

Les transformations sont implémentées en utilisant de simples expressions régulières. Nous n'utilisons pas actuellement de ressources externes comme un parseur. Toutefois, nous sommes conscients de l'utilité de ces outils pour améliorer les traitements et résultats (Wehrli et al. 2010).

4.4 Seuil de fréquence, de la "règle de trois" à la "règle de deux"

Nous appelons règle de trois la méthode utilisée pour déterminer que les séquences sont utilisées assez fréquemment et possèdent une dispersion assez significatives pour être appelées collocations (Dubreil et Daille 2005). Pour cela, 3 règles doivent être vérifiées

- « - la cooccurrence de deux termes apparaissant au moins trois fois dans le corpus ;
- la cooccurrence de deux termes issus de trois articles différents;
- la cooccurrence de deux termes employés par trois auteurs différents. » (Dubreil et Daille, 2005)

Dans notre cas, ces trois critères peuvent être réduits à seulement deux. Chaque courrier est écrit par un client unique, donc « trois auteurs différents » et redondant avec « trois articles différents ». De plus, notre corpus n'étant pas très volumineux, nous réduisons l'apparition des termes au moins trois fois dans le corpus par seulement deux fois. De plus, n'oublions pas que nous voulons extraire également des séquences figées qui dans un corpus réduit ne sont pas très fréquentes. Ce seuil mis à « deux » est un bon compromis entre la non significativité d'une apparition unique et un seuil trop haut.

4.5 Score de figement

Le score de fixité permet de décider si une séquence candidate est une expression phraséologique. Pour comprendre ce score précisons son calcul. Tout d'abord, nous calculons un score pour chaque type de transformation (syntagmatique, morphologique et paradigmatique). Le calcul du score pour les transformations syntagmatiques et morphologiques est identique. C'est un simple ratio entre le nombre d'occurrences de la séquence candidate et la somme des différents tests et de la séquence candidate.

$$F_{synt}{}^{Si} = \frac{S_n}{(S_n + \sum T_n)}$$

Le calcul du score du figement paradigmatique est un ensemble d'heuristiques. Il dépend à la fois du nombre de déterminants substituables de la structure syntaxique et des substitutions verbales et nominales. Une différence entre les « upward collocations » et les « downward collocations » est également faite (Sinclair 1991). Les « upward collocations » sont dans notre cas des substitutions verbales, tandis que les « downward

collocations » sont des substitutions nominales. Disons brièvement que les « downward collocations » peuvent être très productives surtout avec des auxiliaires. Par exemple, dans notre corpus, *avoir* accepte 99 noms différents. Mais *fin*, nous l'avons vu, n'accepte que 2 substitutions dans une structure VER NOM. Donc, nous mettons en valeur les « upward collocations » c'est-à-dire les substitutions verbales possibles à partir d'un nom. C'est toutefois un choix risqué lorsque l'on sait que les collocations se réalisent majoritairement à partir du prédicat.

Enfin, nous fixons plusieurs seuils permettant de déduire qu'un candidat est une expression phraséologique : le score de figement morphologique (F-M) doit être supérieur à 0.8 ; le score de figement syntagmatique (F-S) doit être supérieur à 0.7 ; le score de figement paradigmatique (F-P) doit être supérieur à 0.6. Ces choix de seuils sont pour le moment pris de manière subjective, respectant tout de même une certaine probabilité dans les différentes transformations.

En prenant trois scores différents nous pouvons contrôler les seuils. L'application d'une moyenne entre les scores ou du calcul d'un score plus général entraîne plus de bruit.

4.6 Constitution d'une base de référence

Afin de comparer notre approche avec une base de référence, nous devons constituer cette base. Actuellement elle est réalisée grâce à différentes sources répertoriant des expressions figées (Lexique-Grammaire[1], expressio[2], DEL[3], Wiktionary[4]). Nous faisons remarquer immédiatement que le terme expression figée est pris de manière assez large selon les ressources. Des verbes supports peuvent même apparaître (*mettre fin*). Toutefois ils sont considérés comme des combinaisons avec une forte attraction et devenir des locutions comme en témoignent certains dictionnaires (notamment le TLFI). Nous les considérons de manière assez naïve sans remettre en considération leur présence. Finalement, 110 séquences évaluées comme étant figées par ces ressources ont été trouvées parmi nos candidats.

5 Résultats

L'extraction à partir de structures syntaxiques abouti à 5 148 candidats représentant 15 794 formes différentes (incluant les transformations morphologiques et syntagmatiques). 1 133 séquences peuvent prétendre au titre d'expressions phraséologiques après l'application de la règle de 2. Parmi elles 302 séquences sont assez figées pour être considérées comme des EP par notre approche, c'est-à-dire en appliquant les scores de figement.

[1] Revu et corrigé par Tolone 2011

[2] Georges Planelles 2011

[3] *Dictionnaire des expressions et des locutions* Rey et Chatreau 2006

[4] http://fr.wiktionary.org/wiki/Catégorie:Locutions verbales en français

5.1 Distribution des transformations

Les transformations représentent 44% des formes. Certaines transformations sont beaucoup plus utilisées que d'autres. Certaines sont même pratiquement inutiles. Par exemple, le clivage n'est utilisé que 3 fois dans notre corpus. Selon Riegel et al. (1994), le clivage est plus utilisé dans le langage parlé qu'écrit. Nous le vérifions ici également.

Règles transformationnelles	**Occurrences**
Insertion syntagme	8,09%
Insertion déterminant	2,42%
Inversion	4,66%
Clivage	0,06%
Relativation	3,68%
Passivation	9,05%
Nominalisation	5,75%
Insertion Modifieurs	3,72%
Flexion	4,33%
Suppression déterminant	0,95%

TABLE 1 – Distribution des transformations

La nominalisation est une transformation productive et intéressante. Les verbes impliqués sont des prédicats de premier ou de second-ordre.

Résilier = résiliation ; restituer = restitution ; rembourser = remboursement
Demander (résiliation) = demande ; attendre (confirmation) = attente

La passivation quant à elle, est une des transformations les plus utilisés. Autant les prédicats verbaux que les prédicats nominaux sont touchés par cette transformation.

Prélever la somme = la somme prélevée
Effectuer un prélèvement = un prélèvement est effectué

5.2 Extraction de phraséologismes

Avant de comparer les résultats avec des méthodes statistiques, regardons combien nous retrouvons de phraséologismes répertoriés par les ressources. 110 EP répertoriées par nos ressources externes sont présentes dans notre corpus. En appliquant la règle de 2, 57 sont conservées. Notre approche en retrouve 47 (soit 80%).

Séquences	F-S	F-P	F-M
mettre fin	1	0,97	1
prendre acte	1	0,96	1
mettre un terme	1	0,97	1
prendre fin	0,87	0,92	1
tenir compte	1	0,88	1
faire l'objet	1	1	1
tomber en panne	1	0,88	1
rentrer dans l'ordre	1	0,81	1
faire le point	1	1	1
faire foi	1	1	1
renvoyer la balle	1	1	1
porter plainte	1	1	1
mener en bateau	1	1	1
couronner le tout	1	1	1

TABLE 2 – Phraséologismes extraits avec la méthode linguistique grâce aux différents scores de figement syntagmatique (F-S) paradigmatique (F-P) et morphologique (F-M).

Ces résultats montrent un échantillon des expressions classifiées comme étant figées à la fois par les ressources externes et l'approche linguistique. Pour chaque séquence, les scores de figement sont spécifiés. Nous pouvons voir que certaines ne sont pas totalement figées. Toutes les séquences extraites n'ont pas le même degré de figement. Nous avons : a) des séquences figées avec un sens opaque : *Mener en bateau, renvoyer la balle ;* b) des collocations ou des verbes supports: *mettre un terme, mettre fin, porter plainte.* Par conséquent, la plupart des séquences non catégorisées comme figées sont à juste titre, des collocations libres ou des prédicats. Par exemple, *souscrire un abonnement* est un prédicat approprié pourtant catégorisé comme une séquence figée. Notre approche confirme bien qu'au vu du comportement transformationnel, nous n'avons en aucun cas une séquence figée (F-S : 0.29, F-P : 0.36).

Séquences	F-S	F-P	F-M
souscrire un abonnement	0,29	0,36	1
avoir droit	0,55	1	1
faire appel	0,63	1	1
souscrire un contrat	0,14	0,5	1
faire un plaisir	0,67	1	1
donner l'ordre	0,33	1	1
avoir le minimum	0,5	0,5	1

TABLE 3 – Phraséologismes non extraits par la méthode linguistique

5.3 Comparaison avec les approches statistiques

Nous comparons ces résultats avec les catégorisations proposées par les méthodes statistiques. Afin de ne pas valoriser une mesure plus qu'une autre, nous prenons en considération plusieurs algorithmes de l'état de l'art : l'information mutuelle spécifique (PMI) ; le rapport de vraisemblance (LL), le chi carré (X^2) et la mesure Dice. Pour chaque mesure statistique, le rang du candidat est trié par son score. Son rang final est déterminé par la médiane de tous les rangs.

Séquences	Rang	PMI Rang	LL Rang	X^2 Rang	Dice Rang
mener en bateau	1	1	125	1	1
calculer au prorata	1	1	125	1	1
adhérer à la convention	3	2	135	3	3
remercier par avance	4	298	1	5	2
raccrocher au nez	7	11	40	2	1
couronner le tout	7	3	144	6	7

TABLE 4 – Rang des séquences candidates selon les différentes mesures statistiques

Toutes les mesures donnent plus ou moins les mêmes résultats, même si le LL dévie étrangement des autres mesures. Mais, mise à part l'attraction entre les termes, nous ne connaissons pas le comportement indiquant si nous avons tous les éléments propres aux phraséologismes. Nous remarquons également que la structure VER PRP (DET) NOM est la plus fréquente dans l'extraction. Afin de comparer ces résultats aux nôtres, nous décidons de montrer combien de candidats doivent être extraits pour trouver un maximum d'EP répertoriées.

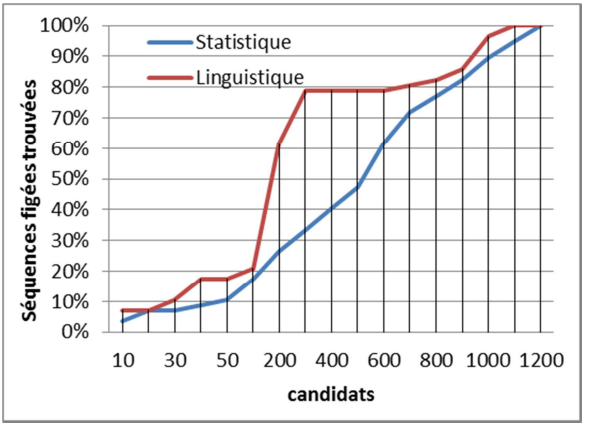

Figure 1 – comparaisons entre les approches linguistiques et statistiques pour extraire des phraséologismes parmi les candidats

Notre méthode extrait 80% des phraséologismes dès 300 séquences extraites tandis que

les statistiques ont besoin de 900 séquences candidates pour atteindre le même score. L'approche linguistique est donc plus précise et trouve les séquences figées plus rapidement que les statistiques dans un corpus non volumineux et spécialisé. Cependant plus de séquences ont besoin d'être annotées et cela de façon précise, afin de mieux comparer ces résultats et pouvoir faire un réel calcul rappel/précision.

6 Les expressions phraséologiques dans un outil de classification de textes

L'extraction des EP n'a pas d'intérêt si elle n'est pas intégrée dans une application fonctionnelle. Nous connaissons son importance dans le domaine de la traduction. Nous le supposons alors dans d'autres outils mais souvent sans preuve. C'est pourquoi, nous voulons connaître l'impact des EP dans la classification de textes. Nous avons alors procédé à une étude expérimentale. Pour effectuer cette tâche nous utilisons un classifieur propriétaire. Il utilise le corpus présenté précédemment. Celui-ci contient 6 classes. 2/3 des données sont utilisées par classe pour l'apprentissage et 1/3 pour le test. Après une phase d'apprentissage où les mots sont pondérés selon leur fréquence d'apparition, le classifieur utilise les K-plus proches voisins et une distance cosinus pour déterminer la classe la plus appropriée. Notre outil utilise également des conditions de rejets permettant de minimiser les confusions. Néanmoins cela implique un rappel plus faible. Nous voulons prouver que les phraséologismes aident à la classification. Nous voulons démontrer également que les séquences les plus figées ne sont pas les plus pertinentes pour cette tâche car elles correspondraient à des expressions linguistiques générales non spécifiques à une classe alors que les moins figées seraient plus proches de la terminologie et donc du sens de la classe.

Test	**Rappel**	**Précision**
Baseline	36%	82%
EP figée	44,9%	85.3%
EP moins figée	47,4%	87.7%
Toutes les EP	50,8%	87.6%

Table 5 – Rappel et précision dans la classification

Dans le tableau ci-dessus, l'hypothèse de départ semble être validée. Nous améliorons les résultats, que cela soit au niveau du rappel qu'au niveau de la précision. Cependant, les séquences les moins figées améliorent de 4% le rappel donné par les EP figées en augmentant la précision de 2,5% supplémentaires. Nous pouvons interpréter l'amélioration de la classification même avec des séquences complètement figées par le fait qu'elles ne sont pas seulement des séquences linguistiques générales (*accuser réception, faire part…*). Elles peuvent être des phraséologismes liées au sens de la classe (*mettre fin, tomber en panne*). Finalement les EP moins figées sont plus appropriées dans la classification car elles ressemblent à des collocations terminologiques (*renouveler un abonnement, résilier une option*). Notons toutefois que combinaison des deux listes améliore les résultats pour atteindre les 50% de classification. Par conséquent les premiers résultats laissent penser que les phraséologismes jouent un rôle dans

l'appréhension du sens et donc dans la classification de textes. Mais ceci nécessite une analyse plus approfondie.

7 Conclusions et perspectives

En résumé, nous proposons une méthode, appliquée aux expressions phraséologiques verbales, basée sur des critères linguistiques et en particulier sur leur comportement transformationnel. Ceci est effectué dans le but de les repérer et de les décrire automatiquement. Nous ne remettons pas en question l'intérêt des méthodes statistiques mais nous prouvons qu'elles ne sont pas assez précises et oublient souvent les expressions les plus figées notamment dans un corpus peu volumineux. Nous avons implémenté un système pour décrire semi automatiquement les variations des EP dans le but d'enrichir automatiquement une ressource composée d'expressions semi-figées. De cette manière, nous pouvons trouver de nouvelles entrées et ne sommes pas limités à une ressource finie.

Il serait prétentieux, à l'heure actuelle, de prétendre pouvoir catégoriser les EP sans aucun doute. D'une part, notre corpus est spécifique à un domaine et notre système doit être éprouvé avec un corpus plus générique et plus volumineux. D'autre part, notre base de validation mériterait une attention particulière pour séparer les réels phraséologismes des séquences libres. De plus, nous nous focalisons sur certaines structures syntaxiques, certes les plus productives, mais qui doivent être étendues à d'autres. Enfin, un réel critère sémantique est absent de notre étude. Les séquences pouvant avoir selon leur emploi, un sens littéral ou opaque ne sont pas distinguées, (par exemple r*envoyer la balle → se jeter les responsabilités les uns sur les autres*[5] ou r*envoyer une balle à quelqu'un* (dans le sport)). Pour améliorer la richesse du corpus et par conséquent l'identification des EP et de leur description nous intègrerons un module permettant d'utiliser le Web comme un corpus. Nous voulons utiliser les moteurs de recherche pour savoir si une transformation liée à une séquence existe (Joseph, 2012). Avec ce corpus nous pouvons extraire et faire une première description qui sera améliorée par l'utilisation du Web.

Enfin, dans cet article nous voulons prouver que les EP sont utiles pour la classification de textes. Les résultats préliminaires qui nécessitent plus de tests sont toutefois encourageants. Ils montrent que les séquences les plus figées sont moins significatives que celles possédant plus de variations mais permettent toutefois d'améliorer les résultats.

Références

ABEILLE, A. et SCHABES, Y. (1989). Parsing idioms in lexicalized tags. *In Actes de EACL (European Chapter of the Association for Computational Linguistics*, Manchester.

ABU-SSAYDEH, A.-F. (2005). Variation in multi-word units : the absent dimension. *Studia Anglica Posnaniensia : international review of English Studies*.

[5] Définition trouvée dans linternaute.com

AL-HAJ, H. et WINTNER, S. (2010). Identifying multi-words expressions by leveraging morphological and syntactic idiosyncrasy. *In Actes de COLING 2010 (Computational Linguistics)*, Beijing.

ARCHER, V. (2006). Acquisition semi-automatique de collocations à partir de corpus monolingues et multilingues comparables. *In Actes de TALN 2006 (RECITAL)*, Leuven.

BEN-HENIA AYAT, I. (2006). *Degrés de figement et double structuration des séquences verbales figées*. Thèse de doctorat, Paris 13, Villetaneuse.

BJÖRKMAN, S. (1978). *Le type avoir besoin. Étude sur la coalescence verbo-nominale en français*. Thèse de doctorat, Uppsala : Acta Universitatis Upsaliensis.

BUVET, P.-A. (2008). Quelle description lexicographique du figement pour le TAL ? le cas des adjectifs prédicatifs à forme complexe. In L*es séquences figées : entre langue et discours*, pages 43–54.

CARTIER, E. (2008). Repérage automatique des expressions figées : état des lieux, perspectives. In *les séquences figées: entre langue et discours,* pages 55–70.

CARTIER, E. et JOSEPH, A. (2011). Repérage automatique des séquences figées pour la classification des documents. *In La notion d'unité en sciences du langage*, Villetaneuse.

COLSON, J.-P. (2010). Automatic extraction of collocations : a new Web-based method. *In JADT 2010 (journées internationales d'analyse statistique des données textuelles)*, Sapienza.

DAILLE, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. *In.(Klavans et Resnik 1996)*, pages 49-66

DIAS, G. (2003). Multiword unit hybrid extraction. *In MWE (Workshop on multiword expressions)*, Sapporo.

DUBREIL, E. et DAILLE, B. (2005). Analyse sémantico-discursive des collocations lexicales en corpus spécialisé : la base « connaissance-s ». *In Actes de LTT 2005 (Lexicologie, Terminologie, Traduction)*, Bruxelles.

FRANÇOIS, T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Thèse de doctorat, Université Catholique de Louvain, Belgique.

GARCI-FERNANDEZ, A., LIGOZAT, A.-L., DINARELLI, M. et BERNHARD, D. (2011). Méthode pour l'archéologie linguistique : datation par combinaison d'indices temporels. *In TALN 2011 (Traitement automatique des langues naturelles)*, Montpellier.

GREZKA, A. et POUDAT, C. (2012). Building a database of french frozen adverbial phrases. *In LREC* 2012 *(Conference on Language Resources and Evaluation)*.

GROSS, G. (1996). *Les expressions figées en français noms composés et autres locutions*. Ophrys édition.

GROSS, G. (2010). Les verbes supports et l'actualisation des prédicats nominaux. *In Supports et prédicats non verbaux dans les langues du monde*, Cellule de Recherche en linguistique. Paris.

GROSS, G. (2012). *Manuel d'analyse linguistique*. Sens et Structure. Presses Universitaires du Septentrion.

GROSS, M. (1982). Une classification des phrases « figées » du français. pages 151–185.

GROSS, M. (1986). Les nominalisations d'expressions figées. *Langue française*, 69, 64–84.

JOSEPH, A. (2012). Pour un étiquetage automatique des séquences verbales figées : état de l'art et approche transformationnelle. *In RECITAL 2012 (Rencontre des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*, Grenoble.

KILGARRIFF, A. (2002). Sketching words. *In (Corréard 2002)*, pages 125–137.

MANNING, C. et SCHÜTZE, H. (1999). Collocations. *In Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Massachusetts, pages 141–177.

MATHIEU-COLAS, M. (1988). *Typologie des noms composés*. Technique 7, Paris 13, Paris.

MEJRI, S. (2011). Les dictionnaires électroniques sémantico-syntaxiques. *In (*Cardoso, Mejri, Mota), pages 159-187.

MEL'CUK, I. (2011). Tout ce que nous voulions savoir sur les phrasèmes, mais… *In Cahiers de lexicologie, revue internationale de lexicologie et de lexicographie.*

PLANELLES, G. (2012). *Les 1001 expressions préférées des français*. Editions de l'Opportu.

RAMISCH, C, (2012). Une plate-forme générique et ouverte pour l'acquisition des expressions polylexicales. *In RECITAL 2012 (Rencontre des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*, Grenoble.

REY, A. et CHATREAU, S. (2006). *Dictionnaire d'expressions et locutions*. Le Robert. Paris.

RIEGEL, M., PELLAT, J.-C. et RIOUL, R. (1994). *Grammaire méthodique du français*. Quadrige Manuels. Paris, PUF édition.

SINCLAIR, J. (1991). *Corpus, concordance, collocation*. Oxford, oxford university press édition.

TOLONE, E. (2011). *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français*. Thèse de doctorat.

VILLADA MOIRON, B. et TIEDEMANN, J. (2006). Identifying idiomatic expressions using automatic word-alignment. *In MWE 2006 (Workshop on Multiword Expressions)*, Italy.

WATRIN, P. (2007). Collocations et traitement automatique des langues. *In Actes de Lexis and Grammar*, pages 1530–1536, Bonifacio (France).

WEHRLI, E., SERETAN, V. et NERIMA, L. (2010). Sentence analysis and collocation identication. *In MWE 2010 (Workshop on Multiword Expressions),* Pékin.