Pour un étiquetage automatique des séquences verbales figées : état de l'art et approche transformationnelle

Aurélie JOSEPH
LDI, 99 avenue Jean-Baptiste Clément, 93 Villetaneuse
ITESFOT, Parc d'Andron, Le Séquoia, 30470 Aimargues
joseph.aurelie@gmail.com

RESUME

Cet article présente une approche permettant de reconnaitre automatiquement dans un texte des séquences verbales figées (*casser sa pipe, briser la glace, prendre en compte*) à partir d'une ressource. Cette ressource décrit chaque séquence en termes de possibilités et de restrictions transformationnelles. En effet, les séquences figées ne le sont pas complètement et nécessitent une description exhaustive afin de ne pas extraire seulement les formes canoniques. Dans un premier temps nous aborderons les approches traditionnelles permettant d'extraire des séquences phraséologiques. Par la suite, nous expliquerons comment est constituée notre ressource et comment celle-ci est utilisée pour un traitement automatique.

ABSTRACT

For an Automatic Fixed Verbal Sequence Tagging: State of the Art and Transformational Approach

This article presents a resource-based method aiming at automatically recognizing fixed verbal sequences in French (i.e *casser sa pipe, briser la glace, prendre en compte*) inside a text. This resource describes each sequence from the view-point of transformational possibilities and restrictions. Fixed sequences are not totally fixed and an exhaustive description is necessary to not only extract canonical forms. We will first describe some transformational approaches that are able to extract phraseological sequences. The building of the resource will be then addressed followed by our approach to automatically recognize fixed sequences in corpora.

MOTS-CLES : séquences verbales figées, reconnaissance automatique, étiquetage, transformations linguistiques, ressources électroniques.

KEYWORDS: fixed verbal sequences, automatic recognition, tagging, linguistical transformations, electronic resources.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 255–266,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Le découpage en mots est la première opération effectuée dans un traitement automatique de la langue. Mais le terme mot est linguistiquement inapproprié car il correspond en informatique à une entité, appelée token, délimitée par des séparateurs graphiques (blancs, retour à la ligne…). Il n'est pas nécessaire de rappeler ici que la notion de mot est beaucoup plus complexe. Et lorsque nous disons complexe nous pensons tout autant à la difficulté de les cerner qu'à leur potentielle polylexicalité. En effet, alors que les traitements informatiques se sont concentrés sur le mot simple, les chercheurs ont prouvé que les mots complexes sont tout aussi importants dans le traitement des langues (Gross, 1982 ; Gross, 1996 ; Mejri, 1997 ; Lamiroy, 2006…). Le traitement automatique de ces séquences devient un enjeu incontournable et se doit d'être traité correctement car la bonne identification et donc l'étiquetage correct de ces séquences dites figées est utile pour de nombreuses applications telles que la traduction, l'extraction d'information, la classification... Notre approche consiste à créer une ressource électronique décrivant les séquences verbales figées (SVF) en termes de possibilités transformationnelles. Nous pensons, que comme les termes simples, elles ont besoin d'être, dans un premier temps, reconnaissables sous toutes leurs formes. Comme un verbe peut être reconnu dans un texte même s'il est conjugué, les SVF doivent être également reconnues malgré leurs modifications. Inversement, pour les séquences avec un dédoublement de sens (Mejri, 2003), les contraintes liées à certaines transformations peuvent nous diriger sur une séquence littérale et non globale (*prendre la veste verte* sera reconnu comme non figé).

2 Séquence figée et extraction automatique : état de l'art

Notre objet d'étude concerne la séquence figée (SF) connue aussi sous le nom d'expression figée, locution, expression idiomatique... Une séquence figée est un groupe de mots non nécessairement contigus, possédant une unité sémantique (sens global), et un figement à la fois morphologique (blocage du nombre), lexical (blocage du paradigme commutationnel) et syntaxique (blocage de la passivation, de la relativation… pour des séquences verbales) (Lamiroy et al., 2008).

A l'instar de Mejri (2011) nous distinguons les séquences figées de deux autres concepts :

- les séquences totalement figées (*au fur et à mesure*) qui n'acceptent aucune modification. L'ensemble est un bloc immuable et dont le traitement nécessite un simple référencement dans un dictionnaire.
- Les collocations : séquences répétées apparaissant fréquemment ensembles (Firth, 1957). Elles peuvent être propres à un domaine (collocation terminologique selon Smadja, 1993) ou typique d'une langue (comme les verbes supports ou les verbes appropriés)

Nous étudions ici, plus précisément la séquence verbale figée (*casser sa pipe, prendre le taureau par les cornes, faire faux bond*…). La problématique des SF et de façon plus importante des SVF, vient du fait qu'elles ne sont pas totalement figées (Gross, 1982 ; Gross, 1996 ; Lamiroy, 2005 ; Abeillé, 1989). En effet, elles autorisent certaines modifications d'ordre syntagmatique et/ou paradigmatique créant ainsi des degrés de

figement (Gross, 1996). Cependant, il n'est apparemment pas possible de définir a priori les transformations réalisables d'une séquence. Villada-Moiròn (2005) remarque que "*there is no uniform presence or absence of syntactic constraints in all fixed expressions since not all fixed expressions exhibit the same syntactic versatility*".[1] (Villada Moiròn, 2005:46). Balibar-Mrabti (2005) va plus loin en postulant que des séquences de même structure syntaxique n'acceptent pas les mêmes libertés transformationnelles (*bruler ses vaisseaux* vs *casser sa pipe*). Comment peut-on alors extraire de telles séquences malgré des modifications évidentes ? Il existe plusieurs approches assez répandues.

La première est purement syntaxique. Laporte et al. (2008) utilisent les patrons syntaxiques productifs en noms composés et vont les proposer à un transducteur (avec l'outil Unitex). En permettant certaines transformations (insertion, coordination…) ils récupèrent ainsi des séquences nominales correspondant syntaxiquement à des noms composés.

La deuxième approche est purement statistique. Ces méthodes utilisent une mesure pour déterminer au mieux la cohésion entre les éléments. Dias (2003) propose ainsi le GenLocalMax qui permet de calculer le degré de figement d'une séquence plus grande que deux mots non obligatoirement contiguë. Cependant l'approche est largement dépendante du corpus et de sa taille.

L'approche la plus utilisée conjugue à la fois syntaxe et statistique. Certains chercheurs (Manning & Schütze, 1999 ; Daille, 1996 ; Watrin, 2008…) commencent par un filtrage linguistique (sélection de lexèmes, patrons syntaxiques) pour ensuite prendre une décision basée sur un calcul probabiliste (*logarithme de vraisemblance, information mutuelle*…). D'autres à l'inverse, génèrent un premier filtrage par critères statistiques pour ensuite effectuer leur choix sur critères linguistiques (Smadja, 1993). Ces méthodes hybrides sont les plus prisées. Toutefois, elles permettent l'extraction de données terminologiques (souvent nominales) plus que l'extraction de séquences figées que nous pourrions appeler 'langagières', c'est-à-dire, qui peuvent se retrouver dans n'importe quel texte quel que soit le domaine. Les possibles modifications intégrées sont de l'ordre de l'expansion de séquence. Peu d'études (Al Haj et Wintner, 2010) testent les transformations morphologiques, lexicales ou même syntaxiques que peut subir une SF afin d'en calculer le degré de figement.

Une autre approche, permettant d'extraire des unités phraséologiques, est basée sur des dictionnaires électroniques : les travaux du LDI (notamment ceux de Ben-Henia Ayat, 2006, 2009 ; Buvet, 2008 ; Cartier, 2008) ou du Lexique-Grammaire (notamment Tolone, 2011). Les dictionnaires électroniques du LDI décrivent les emplois des termes de manière syntactico-sémantique. Les éléments sont catégorisés en prédicat, argument, actualisateur et leur comportement syntaxique lié au sens est déterminé par ces mêmes notions : *prendre(HUM,taureau,corne)*. Les séquences polylexicales sont également

[1] « *observationnel il n'y aucune uniformité dans la présence ou dans l'absence de contraintes syntaxiques dans toutes les expressions figées puisque toutes les expressions figées ne présentent pas la même polyvalence syntaxique* » (Villada Moiròn 2005:46).

traitées de la sorte : *prendre le taureau par les cornes(HUM)*. Cependant leur description est souvent liée à la syntaxe externe (les arguments qu'elles acceptent) et leur traitement interne n'est pas exhaustif. Une description des emplois en termes de prédicat et d'argument peut se révéler essentielle pour extraire, par exemple, des SF analytiquement fausses c'est-à-dire dont le rapprochement syntactico-sémantique n'est pas logiquement correct (*avoir un chat dans la gorge :* un humain ne peut pas avoir littéralement un chat dans la gorge) (Ben-Henia Ayat, 2009) ou pour les désambiguïser d'une séquence homonyme dont le sens est littéral (*prendre une veste*). Cependant, cette description s'avère coûteuse en description sous-jacente car le traitement d'un texte doit être très fin pour pouvoir être analysé.

Abeillé et Schabes (1989) proposent, grâce aux grammaires d'arbres adjoints une méthode pouvant extraire les SF malgré leur discontinuité (insertion, modifieur) et leurs potentiels changements syntaxiques (passivation, clivage…). Cela implique toutefois que la description transformationnelle soit complète.

Finalement, les approches hybrides sont assez rapides et nécessitent peu de ressources et de prétraitement. Toutefois elles incluent dans leur extraction toutes sortes d'unités phraséologiques (souvent des collocations terminologiques). De plus, elles ne prennent pas en compte toutes les possibilités transformationnelles d'une séquence en se limitant souvent à de simples expansions. Les dictionnaires électroniques, plus exhaustifs et précis sont néanmoins coûteux en réalisation et en prétraitement. De plus, même si les chercheurs décrivent la séquence figée comme ayant une double structuration, la description systématique de la structuration interne n'est pas détaillée.

Nous voulons constituer une ressource électronique répertoriant chaque SVF associée à toutes les transformations qu'elle accepte. Cette ressource sera utilisée dans un outil et permettra de reconnaitre toutes les SVF malgré leurs variations possibles.

3 Création de la ressource

3.1 Les transformations

Les transformations 'bloquées' des SVF ont été décrites dans la littérature (notamment Gross, 1982 ; Gross, 1996 ; Mejri, 1997 ; Lamiroy et al., 2006…). Nous l'avons dit précédemment, certaines de ces transformations peuvent être réalisées dans certaines SVF mais elles ne sont pas déterminables a priori. Nous reprenons donc chacune d'elles afin de transformer automatiquement chaque séquence[2] (Cartier et Joseph, 2011). Cette méthode s'apparente à l'utilisation de grammaire d'arbres adjoints (Abeillé, 1989) sans pour autant aspirer à des phrases toujours grammaticales.

- Conjugaison : le verbe est mis à toutes les personnes au présent, imparfait et futur.
- Flexion : chaque séquence est modifiée en changeant le nombre des noms et de leurs actualisateurs associés. Toutes les possibilités sont prises en compte. S'il y a

[2] Les séquences étudiées ont été récupérées dans diverses ressources et correspondent à des séquences particulières (cf Cartier et Joseph 2011)

deux noms dans la séquence nous aurons donc 4 possibilités (*prendre le taureau par les cornes, prendre les taureaux par les cornes, prendre le taureau par la corne, prendre les taureaux par la corne*).

- Substitutions : les noms, les adjectifs, les verbes, les adverbes sont substitués par leurs synonymes et antonymes (pour les adjectifs et verbes) les plus récurrents. Les déterminants sont substitués par d'autres déterminants (*indéfini, défini, possessif, démonstratif*).
- Modifieurs nominaux : seuls les compléments du nom et les relatives peuvent être modélisées pour être requêtés sur des données non étiquetées. (l'adjectif sera un simple mot joker).
- Suppression d'éléments : les adjectifs (*prêter main forte* ➔ *prêter main*), les adverbes, les déterminants, le verbe 'introducteur' sont tour à tour supprimés. Lorsqu'une séquence possède plusieurs syntagmes alors on supprime tour à tour les syntagmes (*prendre le taureau par les cornes* ➔ *prendre le taureau, prendre par les cornes*).
- Insertion : on teste la possibilité d'insérer un déterminant entre le verbe et le nom ou la préposition et le nom s'il n'existe pas (*prendre peur* ➔ *prendre la peur*).
- Négation / affirmation : les séquences affirmatives sont mises au négatif et inversement.
- Inversion : les syntagmes d'une séquence sont inversés s'ils sont au moins deux (*prendre le taureau par les cornes* ➔ *prendre par les cornes le taureau*)
- Passivation : la séquence est modifiée en une phrase au passif (*le taureau est pris par les cornes, le taureau par les cornes est pris*)
- Relativation : la séquence est modifiée en une relative (*le taureau que je prends par les cornes*).
- Clivage : chaque syntagme est clivé (*c'est le taureau que je prends par les cornes, c'est par les cornes que je prends le taureau*).
- Pronominalisation/Détachement : test non effectué actuellement automatiquement.
- Les questions : test non effectué actuellement automatiquement.

3.2 Aide à la création de la ressource

Afin que la décision sur la possibilité transformationnelle ne soit pas uniquement liée à l'intuition du linguiste, un programme (Cartier et Joseph, 2011) va permettre de soumettre chaque séquence transformée en tant que requête à un moteur de recherche (Google, Google Books, Google News). Ce moteur de recherche va nous retourner le nombre de liens trouvés dans le web. Le linguiste pourra finalement valider les possibles transformations en s'appuyant à la fois sur les résultats et les attestations retournées par les moteurs de recherche en vérifiant que la transformation effectuée est bien relative au sens de la séquence figée. Le linguiste reste malgré tout le garant de la validité des transformations en ayant une aide sur l'usage réel de la séquence.

Nous obtenons ainsi une ressource décrivant pour chaque SF les transformations qu'elle peut subir et par conséquent celles qui sont bloquées. Actuellement notre base se restreint à environ 500 séquences transformées et validées. Chaque séquence génère en moyenne 40 transformations. Le nombre de transformations réalisées dépend du nombre

de constituants. Précisons également que les trois quart des SVF de notre base correspondent à la structure 'verbe déterminant nom préposition déterminant nom'. Ce choix totalement arbitraire, avait pour but de trouver des règles transformationnelles potentielles malgré les postulats de départ.

Le peu de données actuelles est essentiellement dû au fait que le programme de transformation a nécessité une attention toute particulière afin de garantir sa robustesse pour qu'il puisse traiter toutes sortes de structures syntaxiques.

Voici comment se présentent les transformations d'une SVF, dans un premier temps celles qui affectent les constituants (Table 1), dans un deuxième temps celles qui affectent la séquence entière (Table 2).

WORD	MODIFIEUR	CONJUGAISON/FLEXION	SUBSTITUTION
prendre	False	True	saisir
taureau	True	True	boeuf
le	un Adj0	False	False
cornes	False	False	False
par	False	False	False
les	False	False	False

TABLE 1 – Transformations des constituants de la séquence *prendre le taureau par les cornes*

	LOCUTION	passivation_1	negation_neg	affirmation_aff	inversion_syntagme_021
▶	prendre le taureau par le cornes	TRUE	TRUE	TRUE	N1+modif

TABLE 2 – Transformations possibles de la séquence *prendre le taureau par les cornes*

Cette ressource nous permet de connaître sous formes de traits définitoires ce qui caractérise les séquences en termes de transformations possibles. De plus, il peut exister des liens entre les transformations. En effet, un déterminant peut se voir substituer si et seulement si le nom qu'il actualise est modifié. C'est le cas de *le* qui devient *un* si *taureau* est modifié par un adjectif placé avant lui (noté Adj0). Il est donc primordial de le répertorier.

4 Outil de reconnaissance des SVF

Nous proposons une méthode, qui devra être améliorée par la suite, permettant de reconnaitre automatiquement des SVF même si celles-ci ont été transformées et de les distinguer partiellement de séquences littérales homographes. Finalement nous pourrons les relier à leur forme canonique créant ainsi une lemmatisation de la SVF.

4.1 Corpus

Nous testons actuellement notre outil sur une base constituée des premières pages retournées par le moteur de recherche lors de nos requêtes sur les différentes transformations de *prendre le taureau par les cornes.* Dans l'état actuel des choses, la base

n'est utile que pour tester visuellement notre outil. Un corpus servant de référence, permettant de tester nos résultats doit être réalisé au plus tôt. Notre corpus actuel, de plus de 32000 mots, est étiqueté morpho-syntaxiquement et lemmatisé par Treetagger à l'exception des noms. En effet, une séquence figée se caractérise très souvent par un blocage flexionnel des noms (*singulier/pluriel*). C'est pour cette raison que nous ne lemmatisons pas les noms. Toutefois, les séquences ayant des noms acceptant les versions singulier et pluriel seront indiquées dans la ressource et ne seront bien évidemment pas écartées de l'analyse.

4.2 Étapes de reconnaissance

Notre programme procède en plusieurs étapes. La première consiste à constituer un dictionnaire d'entrées des composants de la SVF (noms, verbes, adjectifs, adverbes), qui n'acceptent ni substitution ni suppression. Nous les appelons 'invariants'. Ce terme utilisé de manière un peu abusive représente les termes obligatoirement présents dans la séquence mais peuvent toutefois varier en nombre. Ils peuvent s'apparenter à la « zone fixe » de Laporte (1988) correspondant à un ou plusieurs termes obligatoirement présents dans la séquence. Les 'invariants' se rapprochent de la définition de « tête » d'Abeillé (1989) qui correspond à un terme simple déclenchant la zone de recherche d'une potentielle SF.

Locution	Invariant		
mettre la tete a le carre	tete	carre	
mettre le doigt sur la bouche	doigt	bouche	
mettre le doigt sur la plaie	doigt	plaie	
mettre le feu a les poudres	feu	poudres	
mettre le nez a la fenetre	nez	fenetre	
mettre les pieds dans le plat	pieds	plat	
montrer le bout de le oreille	montrer	bout	oreille
occuper le devant de la scene	devant	scene	
prendre la balle a le bond	balle	bond	
prendre la cle de les champs	cle	champs	
prendre la main dans le sac	main	sac	
prendre le air de le bureau	air	bureau	
prendre le taureau par les cornes	cornes		
rater une vache dans un couloir	rater	vache	couloir

TABLE 3 – Échantillon de SVF associées à leurs invariants

Les 'invariants' sont utilisés comme des déclencheurs d'une potentielle SVF. En effet, le texte étudié est découpé en tokens et chaque forme différente (nom, adjectif, adverbe) est recherchée comme un possible invariant. Les SVF associées à cet invariant sont donc récupérées et deviennent les séquences candidates à évaluer de manière plus approfondie. Autour de cet invariant une fenêtre de recherche (que nous appellerons

'capture') est récupérée. Elle correspond à 10 mots de part et d'autre de l'invariant. Ce nombre a été choisi arbitrairement. Il sera par la suite réévalué, pouvant même dépendre du nombre de constituants de la séquence. Il s'en suit à partir de cette capture, une succession de tests, susceptibles d'éliminer la capture et par la même occasion la potentielle SVF associée. Ces tests concernent dans un premier temps, les constituants de la séquence :

- les autres 'invariants' : tous les éléments obligatoires dans la séquence doivent être retrouvés dans la capture.
- les éléments 'variants' : substituables ou supprimables.
  Si l'élément est supprimable sa présence n'est pas requise. Un élément supprimable peut être également le fait d'un syntagme supprimable (*avoir des fourmis dans les jambes, avoir des fourmis*).
  Un élément de la séquence peut être substitué. Dans ce cas les substitutions possibles sont listées sous formes de lemmes ou de classes d'objets (*avoir des fourmis dans <PARTIE DU CORPS>*)
- les modifieurs possibles ou impossibles : les compléments du nom, les adjectifs, les subordonnées relatives sont actuellement les trois modifieurs que nous recherchons dans la capture à partir d'un nom. Si ce nom ne doit pas être modifié et que des éléments représentants les modifieurs sont trouvés (*de, une étiquette <adjectif>, un relatif*) alors la séquence est éliminée.
- les modifieurs possibles selon une contrainte particulière (substitution du déterminant : *prendre une veste, prendre la veste de sa vie*).

Dans un deuxième temps, nous testons les transformations liées à la séquence entière (inversion, passivation, clivage, relativation). Pour ce faire, nous testons tout d'abord l'ordre dans lequel apparaissent les constituants. En effet, ces types de transformations impliquent un changement syntagmatique des éléments. Ainsi, en disant que *prendre le taureau par les cornes* possède 3 composants (*prendre, taureau, cornes*), ces composants constituent l'ordre suivant : 1 2 3. Le clivage de *taureau* modifie alors l'ordre en 2 1 3 (*c'est le taureau qu'il prend par les cornes*). En procédant de cette manière nous validons l'ordre des éléments par rapport à la transformation associée. Cependant, ceci n'est pas suffisant pour savoir si nous traitons bien la transformation cible. Il nous faut ainsi des éléments définitoires de chaque transformation devant être présents dans la capture pour que la transformation soit validée. Ainsi le passif est défini par la présence d'un verbe au participe passé avec l'auxiliaire *être*, la relativation par un pronom relatif et le clivage par la présence de *c'est* et un pronom relatif (Riegel et al., 1994).

Prenons un exemple : la ressource nous renseigne sur le fait qu'une passivation est acceptée (l'indice 1 indique que c'est une passivation de type : *le taureau est pris par les cornes*. Pour affirmer que nous avons ce passif, nous devons donc trouver l'ordre 2 1 3 mais également avoir un participe passé avec l'auxiliaire *être* (ou sans l'auxiliaire selon certaines conditions).

Précisons également que des interdépendances peuvent survenir et qu'il faut les prendre en compte. C'est le cas dans *prendre le taureau par les cornes* où lors d'une inversion *le taureau* doit être modifié par un complément du nom pour que l'inversion soit acceptée (*prendre par les cornes le taureau de la fantaisie à la française* dans *actualites.leparisien.fr*).

4.3 Prenons le taureau par les cornes : exemple

L'outil a été testé sur notre corpus constitué, comme nous l'avons dit précédemment, de différentes phrases incluant *prendre le taureau par les cornes* sous différentes formes (les différentes transformations). Les séquences libres et les séquences figées cohabitent donc dans ce corpus. La figure 1 illustre une partie des résultats retournés par l'outil.

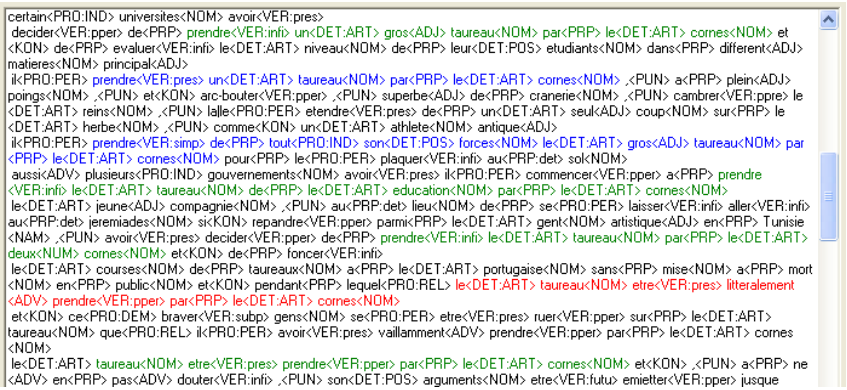

FIGURE 1 – Reconnaissance de *prendre le taureau par les cornes*

Nous pouvons remarquer que les SVF même transformées ont été retrouvées (en vert). Les séquences ne correspondant pas aux possibilités transformationnelles décrites ne sont pas extraites (en bleu). Elles correspondent en effet à la version littérale de la séquence. Cependant, nous ne réglons pas – et nous l'avions prévu – les conflits entre séquence littérale et séquence figée ayant les mêmes transformations ou la forme canonique (en rouge). L'outil aura donc tendance à privilégier la séquence figée. Pour régler de tels conflits, nous ne pourrons pas nous dédouaner d'un traitement de la syntaxe externe par l'analyse des arguments ou encore par un traitement sémantique plus étendu.

Nous présentons Table 4 les résultats de l'étiquetage de la séquence *prendre le taureau par les cornes*. La mesure prend en compte des séquences libres correctement ou incorrectement étiquetées.

	Prendre le taureau par les cornes
Rappel	99.29%
Précision	99.53%
F-score	99.41%

TABLE 4 – Reconnaissance de *prendre le taureau par les cornes*

Les principaux faux négatifs sont dus a des problèmes d'étiquetage morpho-syntaxique.

Précisons que ces données chiffrées sont à titre indicatives car elles ne représentent qu'une petite partie du problème, d'autant plus que même s'il peut avoir un sens littéral,

le sens opaque pour *prendre le taureau par les cornes* est plus fréquent. Le nombre d'attestations de séquences libres en témoignent avec seulement 1% des occurrences dans notre corpus. Ces résultats doivent également être comparés à d'autres méthodes (probabilistes, hybrides…). Toutefois, nous pensons que le début est prometteur et sera compétitif avec d'autres méthodes, notamment par le fait que la reconnaissance est indépendante de la taille du document et que l'importance est donnée aux séquences transformées qui représentent plus de 10% des occurrences, dans notre corpus.

5 Conclusion et Perspectives

Nous venons de présenter une méthode permettant de reconnaitre automatiquement des séquences verbales semi-figées. Cette approche est basée sur une ressource électronique constituée de SVF associées à leurs transformations possibles ou impossibles. Celle-ci répertorie à la fois les transformations liées aux composants de la séquence (modification flexionnelle, substitution, modifieurs…) mais également les changements liés à la séquence entière (passivation, relativation, inversion…) et les dépendances possibles entre les transformations (changement de déterminant en présence d'un modifieur). La ressource permet dans un premier temps, de trouver les mots du texte qui apparaissent comme des éléments obligatoirement présents dans la séquence, permettant de sélectionner des SVF potentielles qui seront testées sur une fenêtre autour de cet élément. Par la suite les composants sont vérifiés aussi bien par leur présence, leur substitution ou leur possible modification. Enfin l'ordre des composants est analysé s'il ne correspond pas à l'ordre canonique. Si aucune transformation syntagmatique n'est réalisable alors la séquence est rejetée. A l'inverse, si l'ordre trouvé correspond à une transformation possible celle-ci doit être validée par rapport à ses éléments définitoires (participe passé pour le passif). C'est ainsi, par l'utilisation d'une ressource électronique et une méthode qui se veut être la plus rapide possible (pour une éventuelle utilisation industrielle), que nous arrivons à extraire des SVF et à désambiguïser certaines de leur double littéral, indépendamment de la taille du corpus. Néanmoins, le travail est loin d'être terminé. La ressource établissant les différentes possibilités transformationnelles doit être complétée. Les SVF peuvent être décomposées en plusieurs types non pas selon leur degré de figement mais selon leur littéralité, leur dédoublement de sens, leur opacité, ou selon le domaine dans lequel on se trouve. En effet, les étapes de reconnaissance peuvent être allégées selon certaines conditions. De plus, la notion d'invariant doit être revue et élargie peut-être même jusqu'à une prise en compte de classes d'objets. L'ajout de la syntaxe externe doit également compléter la description pour une désambiguïsation totale. Enfin, des tests de robustesse doivent être effectués sur un corpus de référence, et les résultats doivent être comparés aussi bien à d'autres mesures qu'à d'autres ressources.

Références

ABEILLE, A. et SCHABES, Y. (1989). Parsing idioms in lexicalized TAGs. *In Proceedings of the the European Chapter of the Association for Computational Linguistics (EACL'89)*. Manchester, Angleterre.

AL-HAJ, A. et WINTNER, S. (2010). Identifying Multi-words Expressions by Leveraging

Morphological and Syntactic Idiosyncrasy. *In Proceedings of COLING 2010 (Conference on Computational Linguistics)*, Beijing, Chine.

BALIBAR-MRABTI, A. (2005). Semi-figement et limites de la phrase figée. *In LINX (53),* pages 35–54.

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

BEN-HENIA AYAT, I. (2009). Les séquences verbales figées métaphoriques. *In Synergie (1)*, pages 159–171.

BUVET, P.-A., (2008). Quelle description lexicographique du figement pour le TAL? Le cas des adjectifs prédicatifs à forme complexe. *In* (*Blumenthal et Mejri 2008*), pages 43–54.

CARTIER, E. (2008). Repérage automatique des expressions figées : état des lieux, perspectives. *In (Blumenthal et Mejri 2008)*, pages 55-70.

CARTIER, E. et JOSEPH A. (2011). Repérage automatique des séquences figées pour la classification des documents. *In LTT 2011 (Lexicologie, Terminologie, Traduction)*.

DAILLE, B. (2001). Extraction de collocation à partir de textes. *In TALN 2001 (Traitement automatique des langues naturelles*). Tours.

DAILLE, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *In (Klavans et Resnik 1996)*, pages 29–36.

DIAS, G. (2003). Multiword Unit Hybrid Extraction. *In Workshop on multiword expressions of ACL meeting (Association for Computational Linguistics)*. Sapporo, Japon.

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

GROSS, M. (1982). Une classification des « phrases figées » du français. *In Revue québécoise de linguistique*.

LAMIROY, B. et al. (2010). Les expressions verbales figées de la francophonie Belgique, France, Québec et Suisse. Ophrys, Paris.

LAMIROY, B. (2008). Les expressions figées : à la recherche d'une définition. *In (Blumenthal et Mejri 2008)*, pages 85–88.

LAMIROY, B. (2005). Le problème central du figement est le semi figement. *In LINX (53)*, pages 135–153.

LAPORTE, E. (1988). Reconnaissance des expressions figées lors de l'analyse automatique. *In Langages 23(90)*, pages 117–126.

LAPORTE, E., NAKAMURA, T. et VOYATZI, S. (2008). A French Corpus Annotated for Multiword Nouns. *In LREC 2008 (International Conference on Language Resources and Evaluation)*. Maroc.

LEPESANT, D. et MATHIEU-COLAS, M. (1998). Introduction aux classes d'objets. *In Langages 32(131)*, pages 6–33.

MANNING, C. et SCHÜTZE, H. (1999). Collocation. *In Draft*, pages 141–177.

MEJRI, S. (2011). Les Dictionnaires électroniques sémantico-syntaxiques. *In (CARDOSO et al. 2011)*, pages 159–188.

MEJRI, S. (2008). La place du figement dans la description des langues. *In (Blumenthal et Mejri 2008), p*ages 117–129.

MEJRI, S. (2003). Polysémie et polylexicalité. *In Syntaxe et sémantique (5)*.

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

SMADJA, F. (1993). Retrieving Collocations from Text: Xtract. *In Computational linguistics 19(1),* pages 144–177.

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

VILLADA MOIRON, M.B. (2005). Data-driven identification of fixed expressions and their modifiability. Thèse de doctorat. Université de Groningen, Pays-Bas.

WATRIN, P. (2007). Collocations et traitement automatique des langues. *In Lexis and Grammar*, Bonifacio.