Outils et méthode de constitution de dictionnaire de formes figées

Fabrice Issac

Lexiques, Dictionnaires, Informatique, UMR 7187, Université Paris 13

Résumé

L'objectif de cet article est de présenter un outil et une méthode d'extraction des « mots en plusieurs mots » en corpus à partir de classes sémantiques. Nous ne présentons pas une méthode autonome, mais plutôt un moyen permettant d'organiser les informations destiné à venir en aide aux spécialistes souhaitant créer des ressources linguistiques. Après avoir précisé le sens que nous attribuons à la notion de mot en plusieurs mots, nous présenterons Corpindex, la plate-forme logiciel, et nous décrirons les différentes étapes de la méthode. Nous conclurons en présentant des perspectives liées à l'expérimentation et à l'évaluation.

Abstract

The aim of this paper is to present a tool and a method, based on semantic classes, to extract « multiwords unit » from a corpus. We do not present an autonomous method, but rather a way to assist specialists wishing to create language resources. After clarifying the meaning we ascribe to the notion of « multiwords unit », we will present Corpindex, a software platform, and describe the various steps of the method. We conclude by presenting prospects.

Keywords : collocation, multiwords unit, concordance, software platform

1. Contexte

Les dictionnaires de mots simples ou de mots composés sont les ressources non seulement les plus faciles à constituer mais aussi les plus fondamentales puisqu'elles sont utilisées lors de la première étape de tout traitement linguistique informatique : l'analyse lexicale. Cependant, la langue n'est pas constituée uniquement de mots simples et on y trouve en grand nombre des « mots en plusieurs mots ».

L'objectif de cet article est de présenter un outil et une méthode d'extraction de ces « mots en plusieurs mots » en corpus à partir de classes sémantiques. Nous ne présentons pas une méthode autonome, mais plutôt un moyen permettant d'organiser les informations destiné à venir en aide aux spécialistes souhaitant créer des ressources linguistiques.

Après avoir précisé le sens que nous attribuons à la notion de mot en plusieurs mots, nous présenterons Corpindex, la plate-forme logicielle, et nous décrirons les différentes étapes de la méthode. Nous conclurons en présentant des perspectives liées à l'expérimentation et à l'évaluation.

1.1. Définition

Le figement est un phénomène linguistique à la fois très complexe et très répandu. Sa prise en compte par les systèmes automatiques est, en comparaison, bien faible. Cette faiblesse

provient de l'ambiguïté propre à la notion. D'un point de vue terminologique on constate la présence de nombreux termes se rapportant à un objet que tout le monde « voit » mais que l'on a du mal à discerner. Gaston Gross (Gross, 1996) présente un ensemble de termes dans son ouvrage sur le figement : expression idiomatique, mot composé, mot polylexical, collocation, locution ..., chacun présentant des caractéristiques spécifiques. Nous n'ajouterons pas notre propre pierre à cet édifice et nous ne considérerons ces entités uniquement d'un point de vue opératoire, les questions étant : comment les représenter et surtout comment les identifier ? La seule caractéristique que nous allons donc considérer pour désigner un figement est qu'il est composé de plusieurs mots. Le terme, volontairement flou, souvent employé dans la littérature anglophone pour désigner ces objets est « multiword unit » ou « multiword expression » que nous traduirons par « expression en plusieurs mots ». Le phénomène de figement se constate dans un continuum et répond à certains critères exprimés sous forme de contraintes, celles-ci étant d'ordre lexicale, transformationnelle, sémantique, …

1.2. Pourquoi s'intéresser aux expressions en plusieurs mots

Il existe plusieurs raisons pour lesquelles il est important de s'intéresser à ce type de structure linguistique, tout d'abord des raisons linguistiques-informatiques. Le traitement automatique des langues, en ce qui concerne l'écrit, procède à une analyse à partir de ses constituants les plus atomiques, les caractères, et effectue des regroupements pour tout d'abord former des éléments à partir d'un lexique puis des syntagmes pour finalement arriver au sens. Celui-ci peut à son tour être utilisé pour effectuer des inférences. Dans ce mécanisme, chaque étape dépend de l'étape qui la précède, et la notion de mot, qui précède toutes les autres, tient ici une grande importance.

Le problème est donc d'identifier ce qu'est un mot. D'un point de vue linguistique, la littérature abonde pour décrire le mot, sans pour autant qu'une définition précise en soit donnée. De multiples problèmes interviennent suivant la langue et le système d'écriture utilisé. Cependant il est un phénomène récurrent à toutes ces langues : le fait que certains mots prennent une indépendance par rapport au mode de fonctionnement « normal ». Ces unités qui dérogent à une règle idéale, qui en fait n'existe pas, sont non seulement très nombreuses mais aussi porteuses d'informations, lexicales, syntaxiques ou sémantiques et ne peuvent être ignorées lors de la construction du sens. Leur utilisation dans les systèmes de traitement automatique est donc fondamentale, quel que soit le type d'application envisagé. Les outils de traduction automatique et les plateformes d'apprentissage des langues, du fait de la confrontation de plusieurs langues, sont cependant plus emblématiques.

En ce qui concerne les outils de traduction, il suffit d'utiliser les produits à large couverture, comme ceux proposés en libre accès sur le web, pour se rendre compte rapidement que si les résultats sont compréhensibles, ils ne sont en rien acceptables en tant que tels et, lorsque les résultats sont bons, on se rend compte que c'est justement parce que les expressions en plusieurs mots, non seulement les expressions mais aussi les collocations, sont pris en compte.

Nous avons proposé la phrase suivante : *il exerce la profession de marchand de pomme de terre* à plusieurs traducteurs automatiques (Google, Reverso, Systran). Les résultats obtenus sont les suivants :

it works as a merchant potato (Google)

He(It) exercises the profession of walking(working) of apple of roam (Reverso)

he follows the occupation of potato merchant (Yahoo)

La plus mauvaise traduction est sans conteste celle de Reverso, par contre le traducteur de google prend en compte correctement les collocations dans la phrase *il exerce la profession de marchand de pomme de terre* puisqu'il propose *it works as a merchant potato* qui, sans présumer de la qualité de la traduction, est capable de proposer un synonyme correct pour *exercer une profession* et *pomme de terre*. Ce résultat est obligatoirement obtenu par l'utilisation d'un dictionnaire d'expressions en plusieurs mots.

2. Méthodes d'identification

L'identification ou le repérage des expressions en plusieurs mots sur corpus peut être envisagé selon un point de vue statistique ou selon un point de vue plus syntaxique. La première famille de méthodes nécessite un corpus relativement étendu puisque c'est par l'observation de phénomènes récurrents qu'il est possible d'inférer des interprétations linguistiques. La seconde famille s'appuie sur des langages de description de phénomènes linguistiques, citons par exemple l'utilisation de grammaires locales pour extraire des syntagmes nominaux à fonctionnement dénominatif (Bourigault, 1993).

Beaucoup de travaux sur l'identification en corpus d'expressions en plusieurs mots choisissent de focaliser sur les collocations. Celles-ci ont les caractéristiques suivantes :

1. Ce sont des associations conventionnelles, c'est-à-dire que leur construction respecte les règles de la grammaire traditionnelle ;
2. elles sont récurrentes, c'est-à-dire qu'elles reviennent dans le discours avec une fréquence supérieure à la normale ;
3. les éléments qu'elles contiennent ne sont pas, contrairement aux mots composés, obligatoirement contiguës ;
4. le sens est compositionnel, c'est-à-dire généralement transparent ;
5. l'échange d'un des constituants par un synonyme ne change pas obligatoirement le sens mais est généralement ressenti moins adéquat par un locuteur.

L'identification d'expressions plus figées même si elles apparaissent moins fréquemment, reste pour le traitement automatique nécessaire. Il n'est cependant pas possible de s'appuyer sur le nombre d'occurrences observées puisque même si les expressions en plusieurs mots sont fréquentes, une expression en particulier n'apparaîtra qu'à une faible fréquence (Colson, 2000).

Pour les collocations, le principe des différentes méthodes (p. ex., Evert, 2003) est d'utiliser le fait qu'elles sont des combinaisons apparaissant avec une configuration déviante par rapport à la normale. Le processus commence par la construction de paires de mots dans une fenêtre d'une taille restreinte, l'augmentation de la taille de la fenêtre induit une croissance importante de la complexité et présente de toute manière peu d'intérêt. Chaque paire de mots va se voir attribuer un score, celui-là permettra de classer l'ensemble des paires et d'établir, à partir d'un seuil, les collocations potentielles.

Il existe plusieurs méthodes pour attribuer un score d'association entre deux mots, score à partir duquel la décision de la nature plus ou moins figée est décidée. Les scores sont calculés à partir d'un principe de départ qui va conditionner l'ensemble des traitements et les résultats obtenus. Ainsi Smadja (1993) part du principe que les mots composant une collocation doivent apparaître ensemble de manière plus significative que ne le veut le hasard. Ce principe de départ est mis en œuvre à l'aide d'outils statistiques ou probabilistes – p. ex. le Z-test, le khi-carré ou l'information mutuelle – et affiné à l'aide d'heuristiques. Il est à noter que les différentes

mesures obtenues à partir de ces différentes méthodes ne peuvent pas être utilisées en dehors de leur contexte de calcul, elles sont utilisées afin d'effectuer des classements, c'est-à-dire que la mesure obtenue n'est qu'un indicateur au sein d'un corpus spécifique.

2.1. Le Z-test

Le principe du z-test est de proposer une mesure capable de rendre compte du fait que les mots qui composent une collocation apparaissent plus fréquemment ensemble que ne le voudrait le hasard. Une première étape consiste à construire, à partir d'un corpus, l'ensemble des fréquences entre un mot donné et ses co-occurrents. Parmi cet ensemble de triplets (mot, co-occurrent, fréquence), ceux dont la fréquence présente une déviation par rapport à la normale sont des candidats de collocations. Il est donc possible d'utiliser des tests statistiques tels que le Z-test, qui permet d'attribuer un score à une collection de fréquences.

La formule utilisée $Z_{m_p,m_i} = \frac{f_i - \overline{f}}{\sigma}$ calcule le score de la co-occurrence des mots m_p et m_i à partir de la fréquence d'apparition du mot *i*, de la moyenne des fréquences et de l'écart-type. Le résultat se présente sous forme d'une liste de co-occurrents associés chacun à une mesure, celle-là étant plus ou moins importante suivant la nature collocative de la co-occurrence.

2.2. L'information mutuelle

La méthode probabiliste basée sur l'information mutuelle est, quant à elle, issue de la théorie de l'information. Elle est utilisée pour calculer la quantité d'information partagée par deux mots. Cette mesure permet de rendre compte de la quantité d'information qu'un mot contient sur un autre mot.

La formule $\Im(m_1\, m_2) = \log\left(\frac{P(m_1 \wedge m_2)}{P(m_1)P(m_2)}\right)$ calcule le score de la co-occurrence des mots mots m_1 et m_2 à partir de la probabilité d'apparition des deux mots au sein d'une même structure et de la probabilité d'apparition de chacun des deux mots. Plus des mots apparaissent ensemble plus la mesure sera élevée. La quantité d'information qui est utilisée et représentée par ces probabilités est donnée par le calcul de la fréquence d'apparition des deux mots dans une fenêtre de taille fixée.

2.3. Évaluation

Une évaluation stricte des mesures décrites supra est délicate à plus d'un titre. En effet, les scores d'association obtenus à partir de différentes sources et dans des conditions différentes ne peuvent évidemment pas être directement comparés. Comment en effet comparer des résultats sur des langues différentes, sur des corpus différents, sur des postulats différents. Plus simplement la mise en place d'un processus d'évaluation sur un ensemble de méthodes ne permettra pas de les comparer puisqu'il n'y a pas consensus sur la définition même de l'objet linguistique à étudier. A propos des collocations Pearce (2002), décrivant un processus d'évaluation, indique la nécessité de la mise en place d'autres stratégies incluant notamment le jugement d'un locuteur natif.

3. Corpindex : une plate-forme de traitement linguistique

L'outil que nous avons développé appartient à la famille des concordanciers avec un langage d'interrogation riche et capable, à l'aide de dictionnaires, de traiter des textes « bruts ». De ce

point de vue Corpindex est très proche d'un outil comme Unitex [1]. Les différences se situent non seulement au niveau des fonctionnalités proposées mais aussi dans l'architecture même de l'application. En effet, le programme est écrit en python sous forme de bibliothèques autonomes utilisées par différents scripts en ligne de commande. Nous privilégions à ce stade non pas « l'utilisateur final non spécialiste en informatique » mais plutôt les possibilités offertes par l'outil et la facilité avec laquelle il est possible d'ajouter, de « brancher », différents types de traitements. La méthode d'extraction de séquences figées présentée ici est un exemple d'ajout de fonctionnalité.

3.1. Construction de l'index

Ce module construit à partir d'un texte, étiqueté ou non, une représentation du texte sous forme d'un index. Celle-ci permet un accès non plus séquentiel mais direct à partir de la forme, le lemme, la nature ou de toute autre étiquette associé à un mot. Une telle représentation permet par exemple de repérer instantanément tous les endroits d'un texte ou apparaît un lemme donné. Si le texte n'est pas déjà étiqueté alors il est possible d'utiliser des dictionnaires de mots simples et de mots composés qui réaliseront un étiquetage préalable sans levée d'ambigüité. Une prise en compte partielle des balises XML permet d'ajouter une dimension « structurelle » à l'index créé.

3.2. Requête

L'index construit, il est possible d'utiliser un langage capable de faire des requêtes non seulement sur la forme des mots mais aussi sur les informations attachées à ce mot. Le langage que nous avons développé fait explicitement référence à CQP (Corpus Query Language) développé à l'université de Stuttgart (Oli, 1994) dont il reprend une partie de la syntaxe. A titre d'exemple la requête suivante :

```
[l="un"][c~"^N"][*]?([l="vert"]|[l="jaune"])
```

permet d'extraire les suites de mots dont le premier mot a pour lemme « un » suivi d'un mot dont la catégorie commence par « N » (i.e. un nom dans le jeux d'étiquette que nous avons choisi) suivi, directement ou à une distance de 1, d'un mot dont le lemme est soit « jaune » soit « vert ». Le langage permet en outre de restreindre la recherche à l'aide de l'opérateur within sur une partie du document. C'est dans ce cas la valeur des identifiants que est prise en compte.

```
[c~"^V"][*]?[*]?[l="human"][l="right"] within ~"H$"
```

Dans l'exemple précédant la recherche porte exclusivement sur des parties de document encadrés par une balise avec un identifiant finissant par « H ».

3.3. Post traitements

A l'issue de la sélection des concordance il est possible d'appliquer des traitements. Soit pour « affiner » encore les résultats, soit pour les utiliser comme point de départ pour de nouveaux calculs. La travail présenté ici se place dans ce cadre là. En effet, l'ensemble des concordances est analysé puis regroupé au sein d'un tableau synthétique (cf. *infra*).

3.4. Modification de l'index

Outre les trois modules présentés ici, la plate-forme intègre la possibilité de modifier un index déjà construit. Le fonctionnement est identique à une requête comme décrit plus haut. Lorsqu'il y a identification le système effectue les remplacements pour chacun des tokens décrits.

[1] http://www-igm.univ-mlv.fr/~unitex/.

A partir de ce module il est par exemple possible, d'effectuer une levée d'ambigüité, de faire des regroupement de tokens (mots composés) ou encore d'ajouter des éléments de structure (p. ex. phrastiques).

4. Repérage d'expressions à l'aide de classes sémantiques

Les méthodes présentées portent essentiellement sur les collocations, phénomène très répandu et dont les contours sont relativement bien établis. Le principe central de toutes ces méthodes est d'identifier un comportement sinon impossible du moins inattendu par rapport à une norme. Notre objectif étant de prendre en compte à la fois des indices statistiques, syntaxiques et sémantiques, nous avons choisi de nous appuyer sur le modèle des classes d'objets (Le Pesant and Mathieu-Colas.,1998). Dans ce modèle l'association verbe support-argument est centrale. Cette mise en relation prend place dans un continuum allant de la syntaxe libre à la syntaxe figée en passant par les constructions à verbe support et les collocations (Mejri, 2008).

Nous utiliserons une stratégie basée sur l'identification d'un comportement incongru d'un argument au sein d'une classe sémantique. Les expressions à plusieurs mots présentent un certain nombre de caractéristiques que nous avons déjà énoncées. Pour construire notre système, nous considérons que (i) le critère de fréquence est à prendre en considération (ii) de nombreuses expressions ont un comportement sémantique faisant référence à la métonymie ou présentant des opacités.

4.1. Description de la méthode

Nous présentons les différentes étapes d'une méthode de classification de couples prédicats/arguments par rapport à la notion de figement. Notre objectif est, étant donné un ensemble de mots regroupés en une classe sémantique, de lister l'ensemble de leurs co-occurrents et de trier ceux-ci par rapport à leurs degrés de figement. Nous avons donc trois critères. Le premier est syntaxique et contraint la sélection des co-occurrents au sein d'un schéma. Celui-là doit prendre en compte la nature du prédicat, adjectival ou verbal dans notre étude, et de l'argument. De même la taille de la fenêtre dans laquelle la recherche s'effectue est indiquée. Le second critère, sémantique lui, identifie la co-occurrence du prédicat non pas avec un argument mais avec l'ensemble de la classe. C'est l'incongruence d'un des éléments de cette classe qui sera pris en compte pour le classement. Le dernier critère, d'ordre statistique, nous permettra d'éliminer les phénomènes marginaux.

Le corpus sur lequel à porter notre étude est constitué d'environ dix années du journal Le Monde. Après avoir été catégorisé à l'aide de Cordial l'ensemble du corpus a été indexé à l'aide de Corpindex.

Étape 1 : Constitution du corpus

La première étape consiste donc à constituer un corpus étiqueté de large couverture. L'homogénéité du corpus est évidemment corrélée avec l'objectif à atteindre. L'étiquetage pose un problème en raison même de l'existence de mots en plusieurs mots. Ainsi « *abaisse langue* » sera étiqueté verbe puis substantif au lieu de substantif pour l'ensemble. Cet aspect est donc à prendre en compte lors de l'étape suivante.

Étape 2 : Concordance

La seconde étape est l'extraction des concordances proprement dites. La requête ci-dessous permet de chercher un verbe (`[c~"^V"]`) suivi d'un ou deux mots n'étant ni un nom, ni un verbe, ni une ponctuation (`[c~"^[^NVY]"]?[c~"^[^NVY]"]`) suivi d'un mot dont le lemme est indiqué.

```
<item id="rue">[c~"^V"][c~"^[^NVY]"]?[c~"^[^NVY]"][l="rue"]</item>
<item id="route">[c~"^V"][c~"^[^NVY]"]?[c~"^[^NVY]"][l="route"]</item>
<item id="autoroute">[c~"^V"][c~"^[^NVY]"]?[c~"^[^NVY]"][l="autoroute"]</item>
<item id="avenue">[c~"^V"][c~"^[^NVY]"]?[c~"^[^NVY]"][l="avenue"]</item>
<item id="impasse">[c~"^V"][c~"^[^NVY]"]?[c~"^[^NVY]"][l="impasse"]</item>
<item id="allée">[c~"^V"][c~"^[^NVY]"]?[c~"^[^NVY]"][l="allée"]</item>
<item id="chemin">[c~"^V"][c~"^[^NVY]"]?[c~"^[^NVY]"][l="chemin"]</item>
<item id="sentier">[c~"^V"][c~"^[^NVY]"]?[c~"^[^NVY]"][l="sentier"]</item>
```

Figure 1 : Ensemble de requêtes d'extraction de co-occurrences

On le voit ici la fenêtre de recherche est égale à 4 avec obligatoirement un mot entre les extrémités. Ces contraintes sont évidement modifiables suivant la nature du type d'expressions à observer. Le résultat brut de cette étape est une suite de concordances en contexte, soit sous forme lemmatisée soit en conservant les formes fléchies :

```
Index/lm18 {rue} : de se trouver non (accompagner dans le rue) le nuit , avoir
Index/lm18 {rue} : sur le rambarde qui (border le rue) ) , de le barre de
Index/lm18 {rue} : pour lui donner l'occasion de (parader dans le rue) de Quimper . et
Index/lm18 {rue} : Alberto et Diego Giacometti , (remonter le rue) pour se installer au 46
Index/lm18 {rue} : avoir inventer un lieu , (occuper le rue) et tout le coin où
```

Étape 3 : Regroupement des concordances

Les concordances sont souvent difficiles à exploiter, c'est pourquoi il est important d'effectuer des tris sur leurs composants – contexte gauche, contexte droit, requête – afin de permettre une visualisation des regroupements. Dans notre cas, l'ensemble des concordances est regroupé de manière synthétique au sein d'un tableau. Nous reprenons en cela les travaux entrepris pour réaliser un concordancier adapté aux besoins linguistiques d'une sémantique distributionnelle (Pincemin et al., 2006). Le tableau a en colonne l'ensemble des arguments de la classe à observer, chaque ligne indique, pour un prédicat donné (i) la fréquence de co-occurrences avec chacun des éléments de la classe (ii) la fréquence totale (iii) le nombre d'arguments ayant déclenché le prédicat. Ces deux indicateurs ont pour la théorie des classes d'objets une importance puisqu'ils permettent d'exprimer le fait qu'un prédicat est approprié à tout ou partie de la classe d'arguments. Par rapport au précédent prototype développé, et parce que les objectifs étaient différents, nous ne procédons au cours de cette étape à aucun tri. Le résultat produit est un fichier au format csv, par conséquence importable dans un tableur, à partir duquel il est possible d'ajouter des calculs ou d'effectuer des tris. Ce parti pris permet de manipuler les résultats d'une manière très souple et d'expérimenter des calculs sans avoir à réitérer l'ensemble du processus de calcul.

Le résultat de l'étape précédente est formé d'un ensemble

$$P_i : (f_{A_1}, ..., f_{A_n}) \text{ pour } i \in 0, ... ,n$$

où
P_i est le i^{ème} prédicatif
f_{A_k} est la fréquence de la co-occurrence prédicat/argument $P_i\ A_k$
à partir de ces informations on calcule le nombre d'occurrences total de la classe (sPred) et le nombre d'arguments sélectionnés par le prédicat (sArg) :
$sPred = \sum f_{A_i}$
$sArg = \sum 1 \ si\ f_{A_i} > 0$

Étape 4 : Calcul du test

Au cours de cette étape nous effectuons les différents calculs destinés à attribuer une mesure à chaque couple prédicat/argument. Nous appliquons ici un test d'incongruence de classe basé

sur l'idée qu'un prédicat déclenché par une classe d'arguments présentera des caractéristiques différentes suivant la nature de la co-occurrence. Sans préjuger des résultats, en ce qui concerne la nature figée ou non du couple, nous appliquons donc un test. Le calcul de celui- se décompose en deux temps et s'effectue à partir d'un couple prédicat/liste d'arguments :

(i) identification des éléments de la liste d'arguments ayant un comportement incongru ;
(ii) pondération du précédent résultat par rapport à une fréquence d'apparition.

Ces tests sont de deux natures, le premier est un Z-test dont l'objectif est d'identifier pour chaque couple prédicat/liste d'arguments l'argument le plus atypique, le second test propose une mesure permettant de corréler la fréquence de la classe avec le nombre d'argument déclenchés par le prédicat.

$$t_j = \frac{f_{A_i} - \overline{f}}{\sigma_j}$$

où σ_i est l'écart-type des fréquences des arguments déclenchés par le prédicat P_j

L'objectif de ce test est d'identifier si au sein d'une classe d'arguments, un de ces composants apparaît avec une fréquence statistiquement significative. Pour chaque prédicat, on calcule le test donnant la valeur maximale et l'argument associé à cette valeur est proposé (colonne maxPred dans la table 1). La table 1 présente le résultat de ce premier calcul, le tri étant effectué sur le Z-test.

	maxPred	sPred	sArg	zTestMax	sentier	route	rue	impasse	chemin	allée	avenue	autoroute
parader	rue	12	1	2,65	-0,38	-0,38	2,65	-0,38	-0,38	-0,38	-0,38	-0,38
vendre	rue	12	1	2,65	-0,38	-0,38	2,65	-0,38	-0,38	-0,38	-0,38	-0,38
exhiber	rue	8	1	2,65	-0,38	-0,38	2,65	-0,38	-0,38	-0,38	-0,38	-0,38
débarquer	rue	8	1	2,65	-0,38	-0,38	2,65	-0,38	-0,38	-0,38	-0,38	-0,38
miner	route	7	1	2,65	-0,38	2,65	-0,38	-0,38	-0,38	-0,38	-0,38	-0,38
endommager	route	7	1	2,65	-0,38	2,65	-0,38	-0,38	-0,38	-0,38	-0,38	-0,38
démissionner	rue	6	1	2,65	-0,38	-0,38	2,65	-0,38	-0,38	-0,38	-0,38	-0,38
déneiger	route	6	1	2,65	-0,38	2,65	-0,38	-0,38	-0,38	-0,38	-0,38	-0,38
doter	autoroute	6	1	2,65	-0,38	-0,38	-0,38	-0,38	-0,38	-0,38	-0,38	2,65

Table 1 : Tableau de fréquence avec calcul du Z-test

On constate dans cet exemple que l'application de ce seul test ne permet pas de rendre compte de l'aspect répétitif du phénomène. Si le couple parader-rue semble intuitivement être une collocation (*parader dans la rue*) il n'en est pas de même pour le couple déneiger-route et encore moins pour le couple démissionner-rue. Parmi les résultats obtenus on remarque que la fréquence d'apparition (sPred) est peu élevée, sachant qu'elle peut atteindre des valeurs supérieures à 300 pour certains prédicats. Cette faiblesse explique les mauvais résultats du test. C'est pourquoi nous introduisons un second critère qui a pour objet de favoriser les prédicats dont la fréquence est d'autant plus élevée que le nombre d'arguments déclenchant est faible. Nous obtenons donc une mesure calculée comme suit :

$$M_i = t_i \frac{sPred}{sArg}$$

Une telle mesure va permettre d'identifier des associations dont un sens se superpose au sens obtenu par la combinatoire libre. Un tri par rapport à cette mesure permet d'obtenir le tableau suivant :

	maxPred	sPred	sArg	zTestMax	Mesure	sentier	route	rue	impasse	chemin	allée	avenue	autoroute
frayer	chemin	272	2	2,65	359,8	-0,38	-0,35	-0,38	-0,38	2,65	-0,38	-0,38	-0,38
prendre	chemin	939	8	2,23	261,2	-0,57	1,09	-0,46	-0,6	2,23	-0,6	-0,58	-0,51
sortir	impasse	695	7	2,56	254,66	0,24	-0,47	-0,27	2,56	-0,5	-0,53	-0,54	-0,5
faire	impasse	816	7	2,16	252,23	-0,69	-0,05	-0,54	2,16	1,11	-0,68	-0,7	-0,62
être	chemin	1022	8	1,46	186,43	-1,04	0,34	0,57	1,36	1,46	-1,02	-0,97	-0,71
reprendre	chemin	478	6	2,26	179,94	-0,56	1,04	-0,49	-0,57	2,26	-0,56	-0,57	-0,54
poursuivre	route	340	4	2,04	173,57	-0,58	2,04	-0,44	-0,6	1,37	-0,6	-0,6	-0,6
suivre	chemin	442	7	2,52	159,15	-0,35	0,41	-0,4	-0,56	2,52	-0,55	-0,54	-0,54
retrouver	chemin	477	8	2,61	155,43	-0,47	-0,26	0,02	-0,36	2,61	-0,5	-0,51	-0,52

Table 2 : Tableau de fréquences avec pondération du Z-test

Ainsi les premiers éléments du tableau peuvent se voir attribuer un sens combinatoire qui n'est pas le fait d'une collocation mais peut aussi être du type construction à verbe support ou encore un sens figuré. La liste ci-après montre les différents emplois possibles pour les couples :

se frayer un chemin : s'ouvrir un chemin/faciliter l'accès à quelque chose;

prendre le chemin : aller quelque part/s'engager dans une voie;

sortir de l'impasse : sortir réellement d'une impasse/sortir d'une situation semblant inextricable;

faire l'impasse : -/ne pas prendre en considération.

Le système offre aussi la possibilité d'effectuer l'ensemble des tâches présentées sur les formes fléchies des mots. Les résultats obtenus permettent de mettre en relief les formes les plus figées au détriment des expressions offrant plus de variabilité au niveau flexionnel.

Étape 5 : Traitement des résultats

Les résultats précédemment obtenus proposent des associations qu'il convient de valider. Pour cela nous procédons à une étude des variations observées en corpus pour chacune d'elles. Un calcul de fréquences est ensuite appliqué sur le résultat.

213	frayer-chemin	frayer un chemin
20	frayer-chemin	frayer son chemin
8	frayer-chemin	frayer le chemin
5	frayer-chemin	frayer leur chemin
4	frayer-chemin	frayer difficilement un chemin
3	frayer-chemin	frayer avec peine un chemin
2	frayer-chemin	frayer elle-même le chemin
2	frayer-chemin	frayer progressivement un chemin
2	frayer-chemin	frayer son propre chemin
1	frayer-chemin	frayer péniblement un chemin
1	frayer-chemin	frayer un autre chemin
1	frayer-chemin	frayer timidement un chemin
1	frayer-chemin	frayer tant bien que mal un chemin
1	frayer-chemin	frayer encore un chemin
1	frayer-chemin	frayer chemin
1	frayer-chemin	frayer un large chemin
1	frayer-chemin	frayer un bon chemin
1	frayer-chemin	frayer seul un chemin
1	frayer-chemin	frayer maintenant un chemin
1	frayer-chemin	frayer ainsi un chemin
1	frayer-chemin	frayer peu à peu son chemin
1	frayer-chemin	frayer à nouveau un chemin

Ensemble des fréquences d'une co-occurrence

404	sortir-impasse	sortir de le impasse
380	prendre-chemin	prendre le chemin
348	faire-impasse	faire le impasse
213	frayer-chemin	frayer un chemin
70	sortir-impasse	sortir de ce impasse
60	être-chemin	être le chemin
34	être-chemin	être sur le chemin
33	prendre-chemin	prendre un chemin
31	être-chemin	être un chemin
29	prendre-chemin	prendre le même chemin
24	prendre-chemin	prendre pas le chemin
20	frayer-chemin	frayer son chemin
20	être-chemin	être au bout du chemin
14	être-chemin	être en chemin
13	sortir-impasse	sortir de un impasse
13	faire-impasse	faire pas le impasse
11	sortir-impasse	sortir un impasse

Ensemble des contextes les plus fréquents

Table 3 : Calcul de fréquence sur les contextes internes

Les tableaux de la table 3 présentent l'ensemble des formes que peuvent prendre les différents contextes internes des co-occurrences.

5. Conclusion

La section précédente a présenté l'ensemble des outils mis en place pour extraire des expressions en plusieurs mots d'un corpus en se basant sur des indices syntaxiques, sémantiques et statistiques. Cette méthode va maintenant être appliquée sur la classe des « parties du corps » :

bedaine, bouche, bras, chevelure, cheveu, cheville, cou, coude, croupe, crâne, cuisse, dent, doigt, dos, front, genou, gorge, jambe, joue, langue, lèvre, mâchoire, main, menton, narine, nez, nuque, œil, ongle, oreille, paupières, peau, pied, poignet, poing, poitrine, pommette, sein, sourcil, tempe, torse, tympan, tête, ventre, visage, épaule

La première étape consiste à décrire le contexte de recherche des co-occurrences. Nous allons envisager trois types de co-occurrences, les couples *verbe-nom*, les couples *adjectif-nom* et les couples *nom-nom*. Les différentes requêtes seront donc de la forme :

```
[c~"^V"][c~"^[^NVY]"]?[c~"^[^NVY]"][l="<arguments>" &amp; c~"^N"]
[c~"^A"][l="<arguments>» &amp; c~»^N»]
[c~"^N"][c~"^[^NVY]"]?[c~"^[^NVY]"][l="<arguments>" &amp; c~"^N"]
```

La requête *nom-nom* ne rentre pas dans le cadre d'une structure prédicat argument, nous l'avons mis en place afin de vérifier l'apport de la partie sémantique du test. Les fenêtres de mots dans lesquelles s'effectue la recherche sont réduites, l'objectif ici est d'éviter de manipuler un volume de données trop important. Les résultats obtenus sont présentés dans le tableau ci-dessous. On recense le nombre de co-occurrences des lemmes, des formes en indiquant pour les deux cas les valeurs différentes de 1, considérées comme négligeables.

Type	*Nb lemme*	*Nb lemmes >1*	*Nb formes*	*Nb formes >1*
verbe-nom	3288	2231	11293	5925
adjectif-nom	460	226	684	346
nom-nom	9975	5160	16256	7599

Table 4 : Nombre de co-occurrences trouvées

En observant les résultats obtenus, on constate que si le classement permet effectivement de représenter un continuum allant des formes les plus figées vers les collocations puis les prédicats à support, il subsiste un certain nombre d'anomalies. Ainsi le premier élément de la liste est, pour les lemmes, la co-occurrence *parler-langue* qui n'est pas à proprement figé. Le même phénomène s'observe aussi en ce qui concerne les prédicats adjectivaux. Les résultats associés au schéma *nom-nom* ne présentent pas les caractéristiques de formes figées, le critère d'incongruence sémantique est donc pertinent pour l'identification de ceux-là.

Une démarche d'évaluation doit permettre de tester l'identification et le classement de l'ensemble des expressions ou collocations associées à une classe sémantique. Cette évaluation devra vérifier (i) si toutes les expressions ont été trouvées (rappel) et (ii) si les expressions trouvées sont figées ou non (précision). Une telle évaluation, pour être valide, doit porter non seulement sur un ensemble de classes relativement important mais aussi sur différents types de corpus. Il faut pour chaque classe construire un tableau associant pour chaque couple prédicat/argument sa position et son degré de figement (à titre d'illustration la figure 8 présente un classement

sur les cent premiers éléments). Il est à noter que ce dernier point nécessite l'intervention d'un opérateur humain et est par conséquent soumis à variations. Le tableau de la figure 8 montre la présence massive d'expressions figées en début de liste, ce taux tendant à diminuer par la suite ce qui tend à valider la méthode dans cette exemple.

Combinatoire\position	*0-19*	*20-39*	*40-59*	*60-79*	*80-99*
figée	17	14	11	7	7
coll/verbe support	3	4	3	5	6
libre	0	2	6	7	7

Table 5 : Classement des co-occurrences

Cette méthode n'a pas encore fait l'objet d'une évaluation à grande échelle, son objectif avant tout est d'être un outil parmi d'autres dans un processus de constitution de classes sémantiques. Ce type d'outil prend donc place au sein d'une problématique plus large qui est celle de la création de ressources linguistiques pour le TAL. L'élaboration de classes sémantiques est un processus qui n'est pas instantané et se nourrit de ses propres retours d'expériences. La construction de classes d'arguments va permettre de constituer des classes de prédicats appropriés et l'identification d'expressions pouvant être de nature plus ou moins figées. La suite à donner à notre travail concerne néanmoins une expérimentation à grande échelle, non pas pour quantifier les résultats mais pour affiner et adapter les tests suivant les nature des classes et des corpus.

Bibliographie

Bourigault D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *TAL*, 34/2 : 105-117.

Colson J.-P. (2000). Les locutions verbales françaises et allemandes dans le discours journalistique: pistes de recherche, fausses pistes, pistes brouillées. In Gréciano, G., editor, *Micro- et macrolexèmes et leur figement discursif. Actes du colloque international de Saverne*, décembre 1998. Bibliothèque de l'Information Grammaticale, Louvain-Paris : Peeters, pp. 173-189.

Evert S. and Kermes H. (2003). Experiments on candidate data for collocation extraction. In *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, pp. 83-86.

Gross G. (1996). *Les expressions figées en français : noms composés et autres locutions*. Paris : Ophrys.

Le Pesant D. and Mathieu-Colas M. (1999). *Introduction aux classes d'objets*, *Langages,* 131. Paris: Larousse.

Mejri S. (2008). Constructions à verbes supports, collocations et locutions verbales. In Mogorron Huerta, P. and Mejri, S., editors, *Las construcciones verbo-nominales libres y fijas. Aproximación contrastiva y traductológica*, Universidad de Alicante, pp. 191-202.

Oli C. (1994). A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*, Budapest.

Pearce D. (2002). A Comparative Evaluation of Collocation Extraction Techniques. In *Third International Conference on Language Resources and Evaluation*, Las Palmas, pp. 1530-1536.

Pincemin B., Issac F., Chanove M. and Mathieu-Colas M. (2006). Concordanciers : thème et variations. In *JADT2006*, Besançon : Presses Universitaires de Franche-Comté, vol. II, pp. 773-784.

Smadja F. (1993). *Retrieving Collocations from Text: Xtract.* Computational Linguistics, Volume 19, Number 1, March, Special Issue on Using Large Corpora: I.