Moule locutionnel lexicographique et traitement des phraséologismes

Lichao Zhu

Moule locutionnel lexicographique et traitement des phraséologismes

Introduction

Nous proposons dans cet article une méthode de modélisation afin de rendre compte à la fois de la forme et du contenu des phraséologismes. Pour ce faire, nous menons d'abord une réflexion méthodologique avant de présenter notre méthode qui repose sur un corpus lexicographique et des techniques informatiques.

1. Éléments méthodologiques

1.1. Linguistique de corpus et traitement statistique

L'avènement de la linguistique de corpus donne lieu à de nouveaux traitements des phraséologismes (Sinclair, 1991, Hoey 1991, Kilgarriff 1992). Le principe idiomatique (« idiom principle ») que pose Sinclair (1991) considère que les phraséologismes sont des « préfabriqués » de la langue ayant des contraintes spécifiques, mis à disposition du locuteur. Il prône une vision combinatoire et fréquentielle des phraséologismes. Sous cet angle, un phraséologisme est la représentation d'une série d'unités ayant une basse fréquence. La fréquence étant mesurée à l'aune de la représentativité des unités dans un corpus textuel, les phraséologismes ayant de fortes contraintes sont logiquement beaucoup moins représentés que ceux qui subissent de faibles contraintes.

Cependant, il est probable que des néologismes ayant, eux aussi, une fréquence basse dans un corpus soient pris pour des phraséologismes (Grezka et Zhu, 2017). De plus, le principe idiomatique ne prend pas en compte les variations du lexique et l'évolution de la langue. L'autre principe, dit « choix ouverts », est créé pour étendre la réflexions aux unités non phraséologiques (Sinclair, 1991 : 109) : « This is a way of seeing language text as the result of a very large number of complex choices. At each point where a unit is completed (a word, a phrase or clause), a large range of choice opens up and the only constraint is grammaticalness. ». Ce principe, en complément du principe idiomatique, considère les créations langagières comme des contraintes qui ne sont pas dans la courbe standard des fréquences.

Il va de soi que cette méthode ne traite pas de la langue comme un ensemble de signes ayant un sens et qu'elle ignore totalement la dimension sémantique des phraséologismes le traitement des données, au profit de la sous-séquence (n-gramme[1]) et de la statistique qui sont censées rendre compte des phraséologismes. La contribution de l'approche de la linguistique de corpus à

[1] Legallois (2009 : 46) déclare : « La méthode d'analyse adoptée est celle de l'identification des n-grammes : (…) est un n-gramme, toute suite de n formes graphiques non séparées par un délimiteur (signes de ponctuation usuels) qui se répète au moins deux fois. »

l'apprentissage de la langue est limitée, car la compréhension des phraséologismes requérant l'appréhension de leur polylexicalité et de leur dualité sémantique ne peut être obtenue simplement à travers une étude de corpus.

1.2. Traitement linguistique à des fins d'automatisation

Contrairement à la méthode fréquentielle, l'approche du lexique-grammaire aborde la problématique phraséologique d'une manière directe. Elle considère les phraséologismes comme des unités lexicales à part entière et leur applique le traitement auparavant réservé à la combinatoire libre. Basée sur la grammaire et appliquée à un cadre phrastique, cette approche démontre que les phraséologismes n'échappent pas à la règle générale de la grammaire, alors qu'ils étaient exclus de toute description systémique. La classification et la description systématiques de la syntaxe et du lexique constituent ainsi une grande avancée dans le traitement des phraséologismes.

En ce qui concerne la syntaxe, le lexique-grammaire utilise une série de tests de transformations pour examiner la fixité des phraséologismes. Des typologies structurelles sont réalisées (M. Gross 1968, 1990) afin de démontrer d'une manière exhaustive les constructions grammaticales des phraséologismes.

Pour ce qui est du lexique, il s'agit d'étudier les contraintes de combinaison des phraséologismes lors de leur emploi, qui sont révélatrices de leur congruence lexicale, co-textuelle et contextuelle dans l'énoncé (G. Gross 1996). Le sémantisme spécifique de certaines configurations d'expression figée comme les constructions à verbe support, les collocations, etc. donne lieu à des « classes sémantiques » qui possèdent des caractéristiques syntactico-sémantiques similaires. Considérons le schéma ci-dessous :

$$\left.\begin{array}{l} < Ondulation > : boucler, crêper, gaufrer \\ < Ajout > : pommader, laquer, poudrer \\ < Association > : attacher, enlacer, tresser \end{array}\right\} cheveux$$

(Mathieu-Colas, 2007)

Les classes sémantiques sont créées par les humains pour catégoriser sémantiquement les constructions verbales. La façon de nommer les classes n'est cependant pas déterminée d'une façon uniforme. Par conséquent, cette démarche qui est de nature inductive et subjective n'est pas automatisable dans le sens où, sans intervention humaine, la langue ne peut créer des outils de description sémantiques par elle-même.

2. Modélisation des phraséologismes à partir d'un dictionnaire

Les ouvrages lexicographiques tels que les dictionnaires et les manuels nous prouvent le contraire : l'humain est parfaitement capable de faire comprendre la langue en utilisant le métalangage et sa capacité consistant à établir les relations entre les unités lexicales est hautement automatisable (Martin 2001).

L'informatique peut « outiller » la linguistique grâce à la grande quantité de données automatisables, à la capacité de calcul des machines et à la puissance des algorithmes créés pour simuler l'activité langagière humaine. Cependant, la plupart des méthodes utilisées dans le traitement automatique des langues visent beaucoup plus le résultat que la valeur herméneutique de ce traitement. Dans cette optique, il n'est pas étonnant que les méthodes stochastiques ne donnent pas d'informations linguistiques. Abney (cité par Cori 2008) explique (1996b : 338-339) le rôle de la grammaire dans le traitement automatique des langues : « The grammar is not viewed as a linguistic description but as a programming language for recognizers. The goal is to write patterns that are reliable indicators of bits of syntactic structure, even if those bits of structure are "boundaries" or "kernels" rather than traditional phrases. ». Les « patterns », ainsi créés, ont pour objectif de rendre compte de la réalité des données dont la nature importe peu.

Nous proposons une méthode de modélisation qui se base sur la simulation des activités neuronales dans la compréhension humaine du langage naturel (Cardon et al., 2018). Pour ce faire, nous nous appuyons sur la théorie de la troisième articulation du langage qui est susceptible d'être formalisée par les algorithmes pour modéliser les phraséologismes.

2.1. Définition du moule locutionnel

Nous utilisons ici la notion de « moule » pour conceptualiser les modèles linguistiques des phraséologismes (Zhu 2016, Zhu 2018). Le modèle locutionnel dont la « locutionnalité » génère des combinatoires plus ou moins figées est perçu par R. Martin (1997 : 303) comme suit : « (…) les facteurs de locutionnalité peuvent générer, non pas telle ou telle locution particulière, mais un modèle locutionnel, plus ou moins productif. ». Examinons la définition de « moule » dans un dictionnaire :

1. Objet façonné en creux dans lequel on introduit une matière en fusion, liquéfiée, molle ou détrempée, qui, en se solidifiant, conserve la forme de ce modèle.
2. Objet plein servant de modèle, sur lequel on applique une matière malléable qui en prend l'empreinte ou les contours.
3. *Fig*. Modèle déterminé, type. (*Dictionnaire de l'Académie Française*[2], 9e édition, désormais DAF)

Le moule se caractérise par deux aspects principaux : une « matière malléable » et les « contours ». Le tout constitue un « modèle déterminé », un « type ». Considérons les phraséologismes suivants :

1. Que dieu vous __. (*Que dieu vous bénisse ! , Que dieu vous garde et vous protège !* , etc.)
2. Tel __, tel__. (*Tel père, tel fils*., etc.)
3. En avril ___, en mai ___. (*En avril, ne te découvre pas d'un fil ; en mai, fais ce qu'il te plaît*.)

[2] Consulté le 25 juin 2019 sur http://www.academie-francaise.fr/le-dictionnaire/la-9e-edition.

4. Noël au balcon, ___. (*Noël au balcon, Pâques au tison*.)
5. Quand on parle ___, ___. (*Quand on parle du loup, on en voit la queue.*)

Dans les exemples de 1 à 5, les espaces non saturés sont inférés à partir des configurations spécifiques de ces énoncés, qui sont figées formellement. Mais cette constatation ne sera valide que lorsque l'on délimite et définit ce qui est figé. La proposition *Quand on parle…* n'est pas figée en tant que partie de phrase[3], elle est figée seulement si l'énoncé entier acquiert le statut de phraséologisme.

Tout syntagme (ou phrase) figé(e) doit correspondre à un moule qui croise lexique, syntaxe et sémantique, qui prend la forme de caractères, de signes diacritiques et de ponctuation.

2.2. Unité de la troisième articulation du langage et ses formalismes

Différentes configurations des moules locutionnels sont envisageables. Par exemple, l'expression figée *Battre à plate couture* donne lieu aux strates suivantes :

- strate syntaxique : X0 <battre> (X1) à plate couture (N_0 V N_1 Prép C_2) (modèle du lexique-grammaire)
- strate lexicale : battre à (mort / la main / coups de poings, de marteau, de crosse / plate couture)

Cependant, les recherches linguistiques sur la modélisation des phraséologismes au moyen d'outils informatiques sont rares. Nous notons les travaux de Colson (2014, 2018) qui s'appuient sur la théorie de la grammaire de construction et de la troisième articulation du langage (Mejri 2006) pour modéliser les phraséologismes, en utilisant des outils et des techniques d'analyse informatiques tels que le séquençage en n-gramme et l'évaluation par le rappel et la précision. Mais aucune tentative de modélisation n'est proposée dans ces travaux.

La théorie de la troisième articulation du langage apporte un socle théorique sur lequel la modélisation peut s'opérer. S. Mejri considère que les langues sont triplement articulées du point de vue de l'énonciateur. En plus de la première articulation qui est phonématique, la deuxième qui est morphologique, il y a une troisième articulation qui est constituée des unités lexicales monolexicales et polylexicales. Il précise que (2018 : 25) : « (…) ce qui constitue l'ossature des séquences figées, c'est une sorte de moule qui sert de modèles pour former des séries idiomatiques, sans être nécessairement phraséologiques. (…) La portée théorique de la notion de moule est double : elle privilégie la globalité par rapport aux constituants et met en saillance la forme par rapport à la substance. ». Cette constatation sur la structuration du « moule » correspond à la définition donnée par le DAF. La « forme » et la « substance » constituent un moule régi par un encapsuleur qui est la catégorie grammaticale. Nous l'exprimons sous forme d'une équation :

[3] Par exemple, « Quand on parle de cette affaire, il est embarrassé. » n'est pas une phrase figée.

$$M \rightarrow C(\sum_{i=2}^{n} L)$$

Cette équation signifie que le moule locutionnel (M) implique l'ensemble (Σ) de *n* éléments de différents niveaux linguistiques (L), le nombre des éléments étant nécessairement supérieur à 1 (par conséquent, la somme Σ se situe entre 2 (le seuil minimal *i = 2*) et *n*), à cause du caractère hybride du moule. La catégorie grammaticale de l'unité (C) est, quant à elle, la fonction définie sur cet ensemble. De ce fait, le moule est le résultat d'une équation dont la fonction (mathématique) est la catégorie grammaticale et dont la variable est l'ensemble des éléments de différents niveaux linguistiques. Par exemple, le moule *de ... en ...* dispose d'une structure syntaxique qui est un élément linguistique syntaxique que l'on dénote en tant que L_1 ; cette structure est pourvue d'un sens lorsque l'on y injecte du lexique : *de mal en pis, de mieux en mieux, de pire en pire*, etc., que l'on dénote comme L_2 ; la catégorie adverbiale de ces séquences susmentionnées est leur fonction C. Ainsi, la séquence *de mal en pis* correspond à un moule qui peut être formalisé comme suit :

$$M\ (de\ mal\ en\ pis) \rightarrow C_{adverbiale}\ (L1+L2+\ldots+L_i)$$

Ce qui implique donc que l'ensemble des configurations morphologiques, syntaxiques, lexicales, etc. de l'unité est conditionné par la catégorie adverbiale.

2.3. Variant et invariant

La fixité qui est régie par les contraintes linguistiques, que ce soit lexicales, syntaxiques ou sémantiques, peut être extrapolée à travers « l'invariant » (Mejri 1998, Balibar-Mrabti 2005) d'une séquence figée. Considérons les trois séquences suivantes :

6. Jacques file à l'anglaise.
7. Jacques observe un rouge-gorge.
8. Jacques chante à tue-tête.

Dans 6, la configuration du syntagme verbal *filer à l'anglaise* constitue le prédicat (P) de la phrase, l'argument en première position (ARG0) étant un nom propre de la classe <humain>. Dans 7, *rouge-gorge*, dont le signifié est non-compositionnel et qui est versé dans un signifiant polylexical, est le nom vernaculaire d'une espèce d'oiseau. Dans 8, *à tue-tête* ne varie pas formellement. Cependant, ces trois phrases peuvent subir des modifications morphologiques ou

syntaxiques. Par exemple, il est à noter que le verbe *filer* se conjugue à cause de sa nature verbale qui est marquée par *-er* [4]:

6a Lucie et Mélina filent à l'anglaise.

l'insertion de l'adverbe *vraiment* est également acceptable dans 6, ce qui crée une rupture dans la continuité de la forme :

6b. Jacques file *vraiment* à l'anglaise.

Néanmoins, la fixité de la séquence figée *filer à l'anglaise* ne repose sur celle du syntagme *à l'anglaise*, qui est certes invariable, mais qui est associé également au premier segment verbal. Cette partie verbale comporte le morphème *fil-*, qui demeure invariant, et la désinence *-er* qui varie selon les désinences exigées par la conjugaison. Ce qui peut être formalisé comme suit :

- *fil- + -er + esp1+ à + esp2 + l +'+anglaise*

La première espace (esp1) est une espace « paradigmatisant(e) » (Blanche-Benveniste, 2002) à cause de la nature verbale de la séquence, tandis que la seconde espace (esp2) est insécable. Si le changement du morphème désinentiel *-er* ainsi que l'ouverture d'un nouveau paradigme sont liés au caractère verbal de *filer*, le reste de la séquence est la partie qui constitue réellement le socle rigide du phraséologisme. On peut alors proposer la formule suivante pour synthétiser la séquence :

- V (v (*fil-*, *-er*), *à l'anglaise*)

La catégorie verbale de la séquence (V) est conditionnée par la nature verbale du verbe *filer* qui peut donner lieu aux opérations suivantes :

- La conjugaison à travers la désinence *-er*,
- L'ouverture éventuelle d'un nouveau paradigme après le verbe

Nous constatons les mêmes résultats dans 7 et 8. Dans 7, la nature nominale (*N*) donne lieu à des modifications formelles avec le morphème du pluriel *-s* dans *rouges-gorges*, elle crée également un paradigme déterminatif qui précède au nom composé :

- *N(rouge, -, gorge)* = un/le/une espèce de… *rouge-gorge*, des/les/… *rouges-gorges*

Dans 3, la nature adverbiale (*ADV*) garantit l'invariabilité formelle de la séquence *à tue-tête* comme suit :

- V(*à tue-tête*) = *à, esp1, tue, -, tête*

De ces trois exemples découlent les constats qu'une séquence figée est régie par sa nature grammaticale qui est un invariant qui encapsule l'invariant formel de la séquence. La variabilité liée à la grammaticalité ne doit pas, selon nous, être considérée comme un élément déstabilisateur qui nuit à la fixité de la séquence,

[4] Le marquage morphologique n'est pas un critère x et suffisant de la modélisation. Il y a des verbes qui ne se conjuguent pas à toutes les personnes ni à tous les temps.

car elle est inhérente à toutes les formes lexicales[5]. C'est la globalité de l'invariant qui l'emporte.

La congruence, quant à elle, peut se manifester sous forme de réseaux d'inférences. Un phraséologisme est congruent lorsque son sens s'intègre parfaitement dans un énoncé, ce qui nécessite une stabilité sémantique. La dualité sémantique, qui est inhérente dans chaque phraséologisme, ne peut se révéler que dans certains types d'énoncés, car la plupart des phraséologismes s'intègrent avec leur sens global. Il suffit de circonscrire les périmètres sémantiques d'un réseau d'une unité lexicale phraséologique pour connaître ce qui est sémantiquement « acceptable » (Lyons, 1968).

Par conséquent, le degré de fixité d'une séquence serait corrélé à sa partie invariante qui forme la charpente liée la nature grammaticale de la séquence ; la partie variable s'inscrit dans les espaces permis par cette partie invariable. Ce n'est qu'à partir de ce constat que l'on peut modéliser les moules locutionnels.

3. Démonstration

Nous utilisons le *Dictionnaire de l'Académie Française* (DAF) en tant que corpus dont la pertinence lexicographique est reconnue. Notre script[6] télécharge automatiquement les données lexicographiques du dictionnaire, sous forme de texte brut, via le site portail du dictionnaire. Chaque entrée est ensuite reconstruite avec ses articles.

3.1. Phraséologismes dans une entrée

3.1.1. Microstructure et phraséologismes

Afin de montrer l'abondance des phraséologismes et le traitement spécifique que le dictionnaire leur réserve, nous examinons l'entrée *ARBRE* :

ARBRE nom masculin

xie siècle. Du latin arbor, -oris.

1. Végétal ligneux de grande taille dont la tige ne se ramifie qu'à partir d'une certaine hauteur. Arbre branchu, touffu. Les racines d'un arbre. Le fût, le tronc, la tige d'un arbre. Les branches d'un arbre. Le houppier, la cime d'un arbre. Arbre mort. Arbre à feuilles caduques. Arbre vert ou à feuilles persistantes, qui garde ses feuilles toute l'année. Planter, transplanter des arbres. Cet arbre a bien repris racine. Tailler des arbres. Élaguer, étêter des arbres. Arbre en espalier. Arbre en buisson. Un plant d'arbres. Arbre fruitier, forestier, ornemental. Arbre de plein vent, arbre qui croît sans abri. Arbre nain, arbre que l'on cultive de manière à réduire considérablement sa taille tout en conservant ses autres caractères. Les arbres nains des jardins japonais. Arbre à caoutchouc. Arbre à pain, voir Artocarpe. Arbre à vessies, voir Baguenaudier. Arbre du voyageur, palmier dont les feuilles en éventail conservent l'eau de pluie dans leurs gaines foliaires. Arbre de mai,

[5] Le degré de fixité d'une séquence figée est souvent corrélé à sa nature grammaticale. En français, les séquences verbales subissent des changements morphologiques contrairement aux séquences adverbiales. Ce postulat est également vrai dans la plupart des langues flexionnelles, mais il pose des problèmes dans une langue isolante comme le chinois. L'absence des déclinaisons annule la corrélation entre le degré de fixité et la nature grammaticale.

[6] La récupération s'effectue via un script en langage de programmation Python.

arbre que l'on plantait le 1er mai devant la porte d'une jeune fille à marier ou d'une personne que l'on voulait honorer. (…) (*DAF*)

Nous y isolons cinq types de phraséologismes :

1. Constructions et collocations - ce sont des mots ou des syntagmes avec lesquels l'entrée se combine : *Le fût, le tronc, la tige d'un arbre. Le houppier, la cime d'un arbre. Planter, transplanter des arbres. Élaguer, étêter des arbres. Arbre fruitier, forestier, ornemental*.
2. Combinaisons lexicales - ce sont des phraséologismes sémantiquement transparents mais figés dans leur emploi : *Arbre mort. Arbre à feuilles caduques. Tailler des arbres. Un plant d'arbres. Arbre en espalier. Arbre en buisson. Arbre à caoutchouc*.
3. Séquences opaques - ce sont les phraséologismes, avec une certaine opacité sémantique, qui nécessitent une explication : *Arbre de plein vent,* arbre qui croît sans abri. *Arbre nain,* arbre que l'on cultive de manière à réduire considérablement sa taille tout en conservant ses autres caractères. *Les arbres nains des jardins japonais. Arbre du voyageur,* palmier dont les feuilles en éventail conservent l'eau de pluie dans leurs gaines foliaires. *Arbre de mai,* arbre que l'on plantait le 1er mai devant la porte d'une jeune fille à marier ou d'une personne que l'on voulait honorer.
4. Phraséologismes phrastiques - Ce sont des phraséologismes donnés sous forme de phrase : *Cet arbre a bien repris racine.*
5. Renvois - Un phraséologisme qui figure sous une entrée est renvoyé vers une autre entrée dans le dictionnaire : *Arbre à pain,* voir Artocarpe. *Arbre à vessies,* voir Baguenaudier.

Des formalismes spécifiques des phraséologismes révèlent que le dictionnaire leur accorde une place importante. Nous avons modélisé la microstructure de l'entrée et les formalismes propres aux phraséologismes pour extraire automatiquement les phraséologismes (Mejri et Zhu, à paraître 2020).

3.1.2. Modélisation du moule locutionnel en tant qu'unité de la troisième articulation du langage

S'agissant de la forme et du sens des unités de la troisième articulation du langage, Mejri (2018 : 15) affirme : « Ainsi une telle définition couvrirait-elle tout le spectre des unités de la troisième articulation en y intégrant au moins trois caractéristiques indépendantes de la morphologie :

— la présence d'un constituant Y qui apporte « la forme du sens », c'est-à-dire un contenu catégoriel véhiculant tout le potentiel combinatoire et garantissant un fonctionnement global ;

— la fonction sémiotique de dénomination (on ne dénomme pas avec des morphèmes ou des phonèmes) ;

— la synthèse sémantique qui garantit l'unité et fait qu'une séquence polylexicale peut avoir comme équivalent une séquence monolexicale et vice-versa (casser sa pipe = mourir). » Cette tripartition donne corps au

moule locutionnel qui dispose d'une catégorie grammaticale, une forme polylexicale et un noyau de sens relativement monosémique. »

Dans le dictionnaire, des formalismes sont créés afin d'annoter les phraséologismes. Pour les illustrer, nous prenons comme exemple le premier article de l'entrée *Papier* :

1. Matière à base de cellulose obtenue à partir de fibres végétales diverses que l'on réduit en pâte, puis que l'on étend, fait sécher et débite par feuilles. Pâte à papier. Machine à papier, qui transforme la pâte en feuilles. Blanchiment, séchage du papier. Laver, laminer, calandrer le papier. Rouleau, bobine de papier. Rame, main, feuille de papier. Papier alfa. Papier de chiffons, de bois, de riz. Papier paille, papier maïs. Papier recyclé. Un bout de papier. Des serviettes en papier. Cocotte en papier.

▪ Cette matière, considérée selon sa consistance, et l'apprêt qui lui est donné. Papier fort, mince. Un papier qui a du corps, de la main. Papier mat, glacé, satiné. Papier vergé, gaufré, filigrané, bouffant. Papier crépon. Papier pelure. Papier de soie. Papier marbré, maroquiné, jaspé. Papier vélin, papier parchemin. Un tirage sur beau papier, sur grand papier.

▪ Selon le traitement qu'elle a subi. Papier gris. Papier bulle, voir Bulle III. Papier pur chiffon, pur fil, couché, encollé. Papier chine, japon, hollande. Papier à la forme. Papier adhésif, collant. Papier huilé, paraffiné, sulfurisé. ▪

Notre programme extrait les collocations comme suit :

- *Papier de chiffons, de bois, de riz. Papier paille, papier maïs.*
- *Blanchiment, séchage du papier. Laver, laminer, calandrer le papier*
- *Rouleau, bobine de papier. Papier mat, glacé, satiné.*
- *Papier vergé, gaufré, filigrané, bouffant. Papier marbré, maroquiné, jaspé.*
- *Papier grand aigle, petit aigle, carré, cavalier, colombier, coquille, couronne, écu, jésus, ministre, pot, raisin, tellière.*

Nous distinguons les collocations dont la base ou le collocatif (qui n'est pas l'entrée) est antéposé (*blanchement du papier, séchage du papier*, etc.) et celles dont la base ou le collocatif est postposé (Papier de chiffons, papier de bois, etc.). Les moules des collocations en tant qu'unités de la troisième articulation du langage (Y) peuvent être formalisés soit en X (entrée) + Z, soit en Z + X (entrée) :

Y = X (papier) + Z (chiffon)
(position : postposé, catégorie grammaticale : N, construction : *de Z)*
Y = Z (blanchiment) + X (papier)
(position : postposé, catégorie grammaticale : N, construction : *de Z)*

Nous retenons ici trois paramètres constitutifs de la modélisation du phraséologime en tant qu'unité de la troisième articulation : la position des éléments (autres que l'entrée), la catégorie grammaticale et la construction précise d'un phraséologisme.

3.2. Génération automatique de phraséologismes à partir d'un moule

La chaîne de traitement est la suivante (voir Schéma 1) : à partir de l'entrée (Ent 1), nous repérons des phraséologismes (Phras 1) tels que *Papier de riz*. Il serait alors possible d'isoler le mot vedette « riz » et la construction du phraséologisme

(Constr 1). Nous accédons ensuite à son entrée dans le dictionnaire (Ent 2, en l'occurrence *riz*) dont la définition est « Céréale annuelle de la famille des Graminées, cultivée pour son grain dans les terres humides, marécageuses ou irriguées des régions chaudes, et qui constitue l'une des grandes sources de l'alimentation humaine. ». Nous extrayons ensuite l'item définisseur (Déf 1), « céréale ». Notre programme cherchera ensuite dans le dictionnaire toutes les entrées (Ents 3) dont l'item « céréale » est le définisseur. Finalement, nous générons de nouveaux phraséologismes (Phras 2) à partir de la Forme 1 et de l'Ent 3.

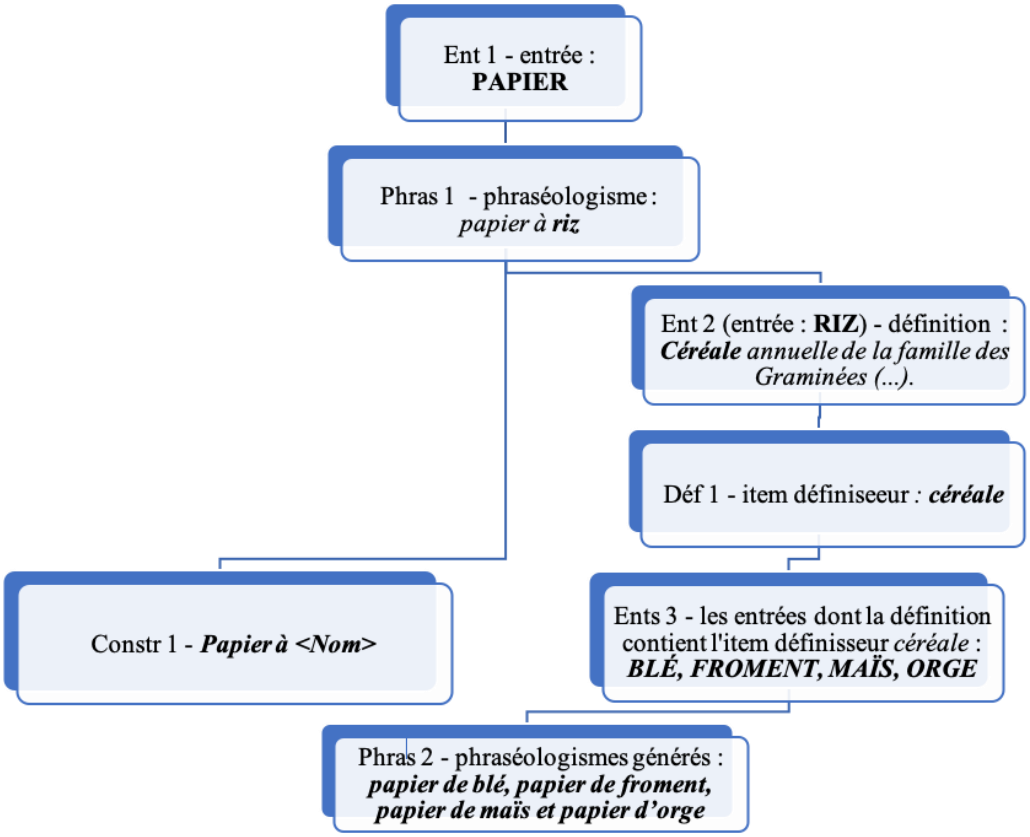

Schéma 1

Notre programme reconnaît les entrées suivantes :

*BLÉ : **Céréale** herbacée annuelle, de la famille des Graminées, cultivée pour ses grains dont on tire la farine servant à la fabrication du pain.*

*FROMENT : **céréale** (se dit aussi bien de la plante que du grain récolté).*
*MAÏS : **Céréale** de la famille des Graminées, originaire d'Amérique, à tige droite, à larges feuilles lancéolées, qu'on cultive pour ses graines ou comme plante fourragère.*
*ORGE **: Céréale** herbacée annuelle, de la famille des Graminées, au feuillage vert et aux épis barbus, cultivée pour son grain.*

Ce faisant, nous établissons un lien sémantique entre l'Ent 1 (*papier*) et les Ents 3 (*blé*, *froment*, etc.) à travers l'Ent 2 (*riz*). Ainsi, pourrions-nous générer *papier de blé*, *papier de froment*, *papier de maïs* et *papier d'orge*, qui ne figurent pas dans le dictionnaire et dont certains ont des attestations sur internet[7]. Ce traitement permettra *in fine* de mettre en relation les entrées grâce à deux types de lien : le premier, de nature lexicale, est sous forme de phraséologisme ; le second, de nature sémantique, a la forme de définisseur.

Conclusion

Notre étude a trois objectifs :

- montrer qu'une réflexion méthodologique est indispensable dans la linguistique outillée. L'apport technologique de l'informatique est indéniable mais il doit servir à étudier, grâce aux données, les problématiques linguistiques ;
- illustrer le pouvoir explicatif de la méthode de modélisation et la pertinence de la notion de « moule locutionnel » ;
- faire une démonstration à partir d'un corpus dictionnairique : d'abord, nous avons identifié et extrait différents types de phraséologismes, puis nous avons modélisé un type de collocation selon les positions et les catégories grammaticales de la base et du collocatif. Nous avons généré, à partir de locutionnels, des phraséologismes. En revanche, la précision de la génération des moules dépendant de l'univocité et de la normalité du modèle définitionnel, des typologies seront alors indispensables à la modélisation définitionnelle et à la création d'un moteur d'inférences. En plus de la relation hyperonymique, il serait également question de traiter la relation dérivationnelle avec des marqueurs métalinguistiques (*en parlant de*, *se dit de*, etc.) (Martin 2001, Mejri et Zhu, à paraître).

Deux points demeurent en suspens :

[7] *Papier de blé*, *papier de maïs* et *papier d'orge* ont chacun plus de 16 000 occurrences selon le moteur de recherche Google (consulté le 23 août 2019).

- la précision de la modélisation et de la génération dépend du modèle définitionnel qui diffère d'un dictionnaire à l'autre ;
- il nous reste à améliorer la méthode de modélisation. L'utilisation d'un annotateur automatique contribue, certes, à un traitement plus efficace, mais les annotateurs sont souvent confrontés à des problèmes de prise de décision et d'homonymie pendant la lemmatisation (par exemple, *porte* est interprété par TreeTagger tantôt comme un nom du lemme « porte » tantôt comme un verbe du lemme « porter ».

Références

Balibar-Mrabti Antoinette 2005, « Semi-figement et limites de la phrase figée », Linx, n° 53, p. 35-54.

Blanche-Benveniste Claire 2002, « Auxiliaires et degrés de « verbalité »», Syntaxe et sémantique, vol. 3, n° 1, p. 75-97.

Colson Jean-Pierre 2018, « Les traces du figement dans les corpus linguistiques : une étude de cas », Le français moderne, n° 1, p. 129-145.

Cori Marcel 2008, « Des méthodes de traitement automatique aux linguistiques fondées sur les corpus », Langages, n° 3, p. 95-110.

Garrigues Mylène 1992, « Dictionnaires hiérarchiques du français. Principes et méthode d'extraction ». Langue française, n° 96. *La productivité lexicale*, sous la direction de André Dugas et Christian Molinier, p. 88-100.

Gross Gaston 1996, *Les expressions figées en français. Les noms composés et autres locutions.* Paris, Ophrys.

Gross Maurice 1968, *Grammaire transformationnelle du français : syntaxe du verbe*, Paris, Larousse.

Gross Maurice 1990, *Grammaire transformationnelle du français : syntaxe de l'adverbe*, ASSTRIL.

Kilgarriff Adam 1992, *Polysemy*, thèse de doctorat, University of Sussex, Sussex.

Lamiroy Béatrice et Klein Jean-René 2016, « Le Figement. Unité Et Diversité. Collocations, Expressions Figées, Phrases Situationnelles, Proverbes », L'information Grammaticale, p. 15-20.

Legallois Dominique 2009, « À propos de quelques n-grammes significatifs d'un corpus poétique du XIXe siècle ». L'Information Grammaticale, n° 121, p. 46-52.

Martin Robert 2001, *Sémantique et automate*. Écritures électroniques, PUF.

Martin Robert 2016, *Linguistique de l'universel*. AIBL.

Mathieu-Colas Michel 2007, « Domaines et classes sémantiques ». Verbum, Presses Universitaires de Nancy, n° 29, p. 11-24.

Mejri Salah 2018, « La phraséologie française : synthèse, acquis théorique et descriptifs », Le Français Moderne, p. 5-32.

Mejri Salah et Gross Gaston (dir.) 2016, *Phraséologie et profils combinatoires : Lexique, syntaxe et sémantique Hommage à Peter Blumenthal*, Honoré Champion, Paris.
Mejri Salah 2006, « Polylexicalité, monolexicalité et double articulation : la problématique du mot », Cahiers de Lexicologie, n° 89, p. 209-221.
Mejri Salah 1998, « La conceptualisation dans les séquences figées ». *L'Information Grammaticale*, n° 2, Numéro spécial Tunisie. p. 41-48.
Mejri Salah et Zhu Lichao (à paraître), « Données dictionnairiques informatisées : phraséologie et inférence ». *Le Français moderne*.
Meneses-Lerín Luis 2017 (dir.), *Corpus et ressources numériques : nouveaux paradigmes de recherche en linguistique, en didactique et en traduction*, Studii de lingvistică, Vol. 7, Editura Universităţii din Oradea.
Miller Philip et Torris Thérèse 1990, *Formalismes syntaxiques pour le traitement automatique du langage naturel*. Paris, Hermès.
Sinclair John 1991, *Corpus, concordance, collocation*, Oxford University Press.
Stranak Pavel 1994, *Guide de la recherche phraséologique en langue de spécialité*, Bureau de la traduction, pagination discontinue.
Zhu Lichao 2016, « Pour une notion de moule dans le figement », *Les Cahiers du dictionnaire*, n° 8, p. 97-110.
Zhu Lichao 2017, « Modularité du figement », Revista de Letras, vol. 36, n° 1, Universidade Federal do Ceaŕap, Brésil, p. 70-79.

Résumé
Notre article entreprend une réflexion méthodologique et herméneutique de la linguistique outillée et démontre, en manipulant les données brutes du Dictionnaire de l'Académie Française (9e édition), la méthode de modélisation et la pertinence de la notion de « moule locutionnel » dans le traitement linguistique des phraséologismes.

Abstract
Our article undertakes a methodological and hermeneutical reflection of the tooled linguistics and demonstrates the method of modeling and the relevance of the concept of "locutional mold" in the linguistic processing of phraseologisms, by manipulating the raw data of the Dictionary of the French Academy (9th edition).

Lichao Zhu