Discours dictionnairique, moule phraséologique et corpus textuel

Lichao Zhu

Discours dictionnairique, moule phraséologique et corpus textuel[1]

Dictionary speech, phraseological mold and text corpus

Lichao Zhu

Résumé

Il s'agit dans cet article de partir d'un corpus lexicographique, en l'occurrence le dictionnaire *Le Trésor de la langue française informatisé*, pour y récupérer automatiquement des phraséologismes transparents comportant des indicateurs d'origines variables, afin d'en extraire des moules phraséologiques. Puis les données dictionnairiques sont utilisées à nouveau pour générer automatiquement des candidats de phraséologismes selon des critères sémantiques. Les phraséologismes générés seront projetés dans un grand corpus pour être validés.

L'objectif de notre expérience est de développer une méthode et un outil pour d'une part simuler le processus de mise en relation des données dictionnairiques par l'humain, et d'autre part modéliser les différents types de contenus du dictionnaire en reconnaissance et en génération automatiques à partir de phraséologismes. Cela permettrait de mettre en évidence le rôle que jouent les phraséologismes dans la création des liens sémantiques entres les unités lexicales.

Mots-clés

phraséologisme, dictionnaire, corpus, simulation, réseaux sémantiques

Abstract

Starting from the computerized dictionary *Le Trésor de la langue française informatisé* as lexicographic corpus, we shall automatically retrieve transparent phraseologisms which have indicators of varying origins, and to extract phraseological molds from them. Then, the dictionary data will be used again to automatically generate phraseologism candidates according to semantic criteria. Finally, the generated phraseologisms will be projected into a large corpus for validation.

The objective of our experience is to develop a method and a tool to, on the one hand, simulate the human process in linking dictionary data, and on the other hand, to model the different types of dictionary content in automatic recognition and generation of phraseologisms. This would highlight the role that phraseologisms play in linking semantically lexical units.

Keywords

phraseologism, dictionary, corpus, simulation, semantic networks

[1] J'adresse mes remerciements à Salah Mejri qui m'a apporté des éclaircissements théoriques importants et qui a nourri mes réflexions pour cet article. Je remercie également Luis Meneses Lerín pour nos discussions inspirantes.

Si le corpus est défini comme une collection de ressources langagières sélectionnées selon certains critères en tant qu'échantillon de la langue (Sinclair 1996)[2], le dictionnaire est un véritable corpus. C'est un échantillon lexicographique si riche que les utilisateurs y trouvent les connaissances linguistiques représentatives (générales ou spécialisées) d'une communauté linguistique.

Les dictionnaires tels que le *Trésor de la langue française informatisé* (désormais *TLFi*) et le *Grand Robert*[3] (désormais *GR*) se servent d'automates pour « simuler (…) le fonctionnement linguistique » (Martin, 2001 : 13). Il est évident que les automates, créés en langages informatiques pour la plupart, simplifient et accélèrent l'accès aux contenus du dictionnaire, qui sont balisés[4]. Mais l'utilisation de ces automates dans l'informatisation des dictionnaires (à l'instar du *TLFi*) nous donne la possibilité de simuler les processus de mise en relation et d'assimilation des données dictionnairiques de la part des utilisateurs.

Dans cet article, nous avons pour objectif d'extraire des moules à partir de phraséologismes transparents[5] existants (de type collocationnel), de générer des candidats de phraséologismes selon des classes sémantiques créées automatiquement, puis de projeter ces phraséologismes dans un corpus. Ce faisant, nous tentons de simuler le processus de mise en relation sémantique des unités lexicales réalisé par l'humain et de mettre en évidence le rôle que jouent les phraséologismes dans ce processus, à l'aide d'outils informatiques.

1. LES DEUX TYPES DE DISCOURS : DISCOURS DICTIONNAIRIQUE ET DISCOURS DANS LE CORPUS TEXTUEL DU WEB

Deux types de discours sont traités dans cet article, l'un sous forme dictionnairique et l'autre sous forme de corpus textuel du web. Nous nous focalisons d'abord sur les différences de ces deux types de discours :

– la nature : le discours dictionnairique est régi par des normes strictes qui sont constituées de structures (microstructure, mésostructure et macrostructure), de symboles, de terminologie et de formes de représentation (police et taille de caractères, interligne, etc.) ;

– la source : le discours dictionnairique puise ses sources dans la langue et le discours d'autorité (par exemple, des proverbes, des citations littéraires, des citations de célébrités, etc.) ; le discours de l'internet s'enracine dans le monde réel et ses sources sont plus variées : la presse écrite en ligne, les forums de discussion, les chats sur les réseaux sociaux, etc. (Mourlhon-Dallies *et al*. 2004) ;

– la validité : le discours dictionnairique est créé par des rédacteurs qui sont des experts en langue, en littérature et en lexicographie, ce qui fait son autorité, tandis que le discours de l'internet est très hétérogène et comporte souvent des erreurs de langue.

[2] « A corpus is a collection of pieces of language that are selected according to explicit linguistic criteria in order to be used as a sample of the language. » (Sinclair, 1996 : 4, cité par Neveu, 2011).

[3] *Le Grand Robert* dispose d'une version payante sous forme logicielle.

[4] Les automates du *TLFi* sont écrits en langage C++, tandis que le dictionnaire est balisé au format XML (Pierrel, 2003).

[5] Par « phraséologismes transparents », nous désignons les phraséologismes qui ne sont pas suivis d'une définition dans un dictionnaire et qui sont souvent compositionnels, à l'opposé des « phraséologismes opaques ».

Ces deux types de discours intéressent les linguistes pour des raisons différentes. Le discours dictionnairique s'articule autour de la langue, il est un condensé des productions linguistiques jugées prototypiques. Dans un dictionnaire électronique tel que le *TLFi*, l'article d'une entrée se construit autour du sens : les définitions qui sont les fils conducteurs de l'article sont formellement identifiables. Quant au discours de l'internet, il est le lieu de diffusion d'informations, de conversations et d'échanges, mais il s'agit également de textes, d'articles, de glossaires, etc.

Le discours dictionnairique consiste en un défi de normalisation et de modélisation des données (Martinez 2009 ; Kosem *et al*. 2015 ; Mangeot & Bellynck à paraître), car les dictionnaires électroniques ne sont pas *a priori* destinés au traitement automatique. Mais l'aspect sémiotique du discours est si saillant que la plupart des travaux en traitement automatique relèguent la langue à un second rôle. Par ailleurs, la taille très limitée d'un « corpus » dictionnairique impose aux chercheurs de modéliser les données, au lieu d'utiliser des méthodes de fouille de textes, comme le partitionnement des données en quelques blocs d'informations homogènes (le clustering).

Or, le dictionnaire est un excellent terrain d'expérimentation pour la linguistique fondamentale[6]. Il dispose de caractéristiques mécaniques de par sa structure : l'entrée est séparée de l'article ; les définitions sont numérotées ; les symboles et la nomenclature doivent être interprétés rigoureusement, etc. Mais l'interprétation des données dictionnairiques est également dépendante d'autres éléments, implicites, qui participent de la logique discursive du dictionnaire : p. ex., les phraséologismes, les exemples et les citations sont rattachés aux définitions ; les phraséologismes transparents et les phraséologismes opaques ne sont pas traités de la même manière dans le dictionnaire. L'interprétation de ces éléments explicites et implicites est exactement un processus de construction d'un discours dictionnairique de la part de l'utilisateur.

1.1. Traitement automatique du discours dans un corpus

Benveniste (1974) place la phrase au centre des niveaux d'analyse du discours. Il considère que le discours impose des contraintes syntaxiques aux syntagmes composés de morphèmes et que la phrase se trouve à la frontière entre l'univers sémiotique et l'univers sémantique. Avec une vision similaire, Harris (1969) conçoit le discours comme une succession de phrases qui intègre « la situation sociale ». Cette succession, « segmentable » jusqu'au morphème, s'appuie sur la « cohérence interne » des occurrences de la langue. Cependant, le cadre phrastique, la séparation nette entre le sémiotique et le sémantique ainsi que la conception du discours comme étant une multitude de phrases ont été critiqués dans plusieurs

[6] Martin (*op. cit*. : 14) explique les différents rôles des automates du point de vue TAL (« analyseurs morphologiques à fondement probabiliste ») et du point de vue lexicographique, ainsi que l'intérêt d'une approche simulative : « Les approximations probabilistes permettent des résultats rapides et impressionnants. Mais elles portent en elles des limites que seule l'approche simulative, assurément plus laborieuse, permet de franchir, parce qu'elle est indéfiniment perfectible. Plus important encore : la simulation conduit à invalider ou à confirmer, à rectifier ou à consolider tel ou tel aspect de la description ou de la théorie linguistique. Le linguiste y trouve un terrain d'expérimentation assurément utile. ».

travaux (Pêcheux 1969 ; Combettes 1987 ; Sarfati 1997 ; Rastier 1998 ; Charolles 2001 ; Adam 2015). En effet, dans le cadre du dictionnaire, l'organisation informationnelle efface le cadre phrastique au profit d'autres formes langagières et sémiotiques. Sous ce prisme, le discours dictionnairique, formé de bribes lexicographiques discontinues et séquencées, constitue un ensemble dont la cohérence découle des formalismes du moule lexicographique du dictionnaire (voir § 2.1.).

En linguistique, le corpus suscite des débats (Meneses-Lerín 2017) quant à son caractère holistique et à son exploitation ; en traitement automatique des langues, le corpus est souvent considéré comme un support langagier d'informations et de connaissances. Il est vrai que les opérations rationalistes exercées sur le corpus ont souvent pour objectif d'en tirer des bénéfices qui sont applicables au monde réel. Un corpus peut être construit selon des critères linguistiques préalablement établis, mais il peut être aussi une simple « collection de ressources textuelles » non ordonnées, si bien que le caractère « aléatoire » des ressources serait utilisé comme une justification de la représentativité des productions langagières d'une communauté linguistique. Néanmoins, d'un point de vue méthodologique, la représentativité d'un corpus ne peut être déterminée si l'ensemble des productions langagières à une période donnée n'est pas évalué, ce qui serait difficile à réaliser. En revanche, un corpus est un réceptacle de discours créés par des locuteurs qui maîtrisent la langue. Il nous semble par surcroît que le fait que les discours soient mis en ligne et extraits de sites internet de référence donnerait une forme de légitimité dans l'observation *in situ* de l'usage de la langue.

1.2. Discours dictionnairique et phraséologismes

Le discours dictionnairique impose un raisonnement déductif qui obéit à des normes. De nature hétérogène, il croise deux genres de discours, comme l'a démontré Rey-Debove (1985 : 22) :

> « Dans un article de dictionnaire, le mot dont on parle se trouve engagé dans deux discours inverses ; (…), tout article contenait (*sic*) un discours en usage sur un mot mentionné (l'entrée) et un discours mentionné (exemple) contenant le mot entré en usage. ».

En effet, le caractère autonymique du dictionnaire participe de la récursivité du discours dictionnairique. Cette récursivité relie les différents types de contenus du dictionnaire et forme des réseaux sémantiques et phraséologiques (Mejri & Zhu 2020).

Le dictionnaire est hautement structuré et ses différents types de contenus sont enchâssés ou classés (Kilgarriff 2005). Son aspect normatif peut être ramené à trois points, sa microstructure, sa macrostructure et le recours à des symboles et à des techniques de présentation :

– la macrostructure organise l'ensemble des données dictionnairiques, la nomenclature, les entrées, etc. assurant une verticalité dans la lecture du dictionnaire. La plupart des entrées sont des unités monolexicales. Dans les dictionnaires tels que le *TLFi* et le *GR*, les affixes et les suffixes sont mis au même plan que les unités monolexicales. En revanche, dans la plupart des dictionnaires de la langue française, les phraséologismes ne bénéficient pas du même traitement que celui des unités monolexicales, ils leur sont rattachés (voir § 2.2.) ;

– s'agissant de la microstructure, elle se caractérise par une binarité dont le premier élément est l'entrée lexicale, le second est l'ensemble des informations qui s'y rapportent. Les dictionnaires utilisent un canevas de représentation dans lequel ils présentent les informations lexicographiques.

1.3. Représentation du discours dictionnairique : les moules

Selon Mejri (à paraître), le moule est « une forme sémiotique correspondant à un type de traitement cognitif qui privilégie les traits globaux au détriment des détails, en vue de favoriser la construction et la reconnaissance des entités isolables (catégories générales) ». Le caractère global du moule peut néanmoins être divisé en une série d'éléments qui sont susceptibles d'être modélisés respectivement (Zhu 2016, 2019). Dans le cadre du discours dictionnairique, les inférences ne peuvent y être explicitées qu'à partir d'un jeu de données et d'opérations raisonnées faisant émerger les propriétés formelles et les relations corrélatives de ce discours. De plus, le discours se distingue par l'utilisation de symboles (les signes de ponctuation, les symboles alphanumériques, les tailles et les polices des caractères, des symboles géométriques comme les losanges, les carrés, etc.) et celle d'abréviations (grammaire, domaine de spécialité, etc.) comme dans[7] :

1. ◆Prendre son courage* à deux mains. (dans *Prendre*)
2. **Prescrire qqc. (à qqn); prescrire (à qqn) de** + inf. ; **prescrire (à qqn) que** + subj. (dans *Prescrire*)
3. *MÉD. Ventre en bateau.* „Forme particulière du ventre, qui est comme creusé dans le sens de sa longueur, dans la méningite des enfants, la colique de plomb, etc.`` (GUÉRIN 1892) (dans *Bateau*)

Dans 1, le losange est le symbole d'énumération de phraséologismes dans l'article de l'entrée et l'astérisque est le symbole de renvoi (en l'occurrence vers *courage*) ; dans 2, le schéma syntaxique est formalisé par des lettres en gras et des abréviations (*qqc, qqn, inf.*) entourées de parenthèses – signes d'optionalité – qui sont liées par le symbole de l'addition ; dans 3, l'abréviation en majuscules, suivie d'une définition, désigne le domaine de *Ventre en bateau* qui jouit d'un statut de sous-entrée. Le dictionnaire marque également les phraséologismes opaques qui s'apparentent à des entrées (dans l'entrée *Corne*) :

4. *Loc. fig. Prendre le taureau par les cornes.* Agir en s'attaquant de front à ce que l'on a à faire, à des difficultés que l'on doit surmonter, etc. *Attaquer l'ennemi là où il serait le plus dur, (...) prendre le taureau par les cornes* (DE GAULLE, *Mém. guerre,* 1956, p. 257)

L'étiquette « Loc. fig. » (en italique) est un mot clef[8] suivi de la séquence figée « prendre le taureau par les cornes » (en italique), qui est lui-même défini comme « Agir en s'attaquant de front à ce que l'on a à faire, à des difficultés que l'on doit surmonter, etc. ». Tout cela s'accompagne d'un exemple littéraire.

Ces types de structures sont récurrents. Leurs récurrences, traduites par des formalismes, sont observables dans chaque dictionnaire. Ce sont ces formalismes qui

[7] Ces exemples sont tirés du *TLFi* et cités tels quels.
[8] Dans la recherche assistée, les utilisateurs peuvent rechercher un phraséologisme par étiquette, en fonction d'autres critères.

construisent le cadre discursif du dictionnaire et jouent le rôle d'enchaînement, assuré par les unités de la langue dans un contexte discursif standard.

2. TROIS TYPES DE MOULE

Dans cet article, nous distinguons le moule dictionnairique, le moule phraséologique et le moule linguistique.

2.1. Moule dictionnairique

Les symboles alphanumériques, les polices, la ponctuation, les icônes et les techniques typographiques sont des supports sémiotiques pour distinguer les différents types de contenus. La structure globale traduit l'intention des rédacteurs et la hiérarchie des contenus du dictionnaire. Dans l'article, les contenus sont ordonnés à l'aide d'un formalisme strict qui varie d'un dictionnaire à un autre, et que l'on peut schématiser comme suit (l'exemple de « MAIN » dans trois dictionnaires, voir Figures 1-3 dans Annexe 2) :

- *DAF* (Dictionnaire de l'académie française 9e édition) : Entrée ([n° romain [n° en chiffre]])
- *GR* : Entrée ([n° romain [Lettre en majuscule [n° en chiffre]]])
- *TLFi* : Entrée ([n° de section [n° romain [n° en lettre majuscule latine [n° en chiffre [n° en lettre minuscule avec parenthèse [n° en lettre grecque minuscule]]]]]])

La microstructure de l'entrée « MAIN » est saillante dans chacun des trois dictionnaires. Entre les numérotations figurent les définitions qui sont représentées en police non oblique et rattachées à un symbole alphanumérique ; elles sont suivies d'exemples en italique : ce sont des citations, des informations supplémentaires ou des exemples fabriqués qui attestent l'emploi de l'unité lexicale dans un contexte discursif fermé[9]. De haut en bas, pour chaque acception, des nuances sémantiques sont apportées soit par la création d'exemples, soit par l'adjonction de définitions hiérarchiquement inférieures, p. ex., la première grande acception de « MAIN » dans le *TLFi* se présente comme suit, correspondant à Figure 3 dans Annexe 2 :

> **1re section** [Organe d'un être animé]
> **I.** — [Chez l'homme] Organe terminal du membre supérieur, formé d'une partie élargie articulée sur l'avant-bras et terminé par cinq appendices (les doigts), eux-mêmes articulés en plusieurs points et dont un (le pouce) est opposable aux quatre autres, organe qui constitue l'instrument naturel principal du toucher et de la préhension et, par là même, un moyen spécifique de connaissance et d'action.
> […]
> **A.** — *[Essentiellement comme objet, indépendamment de ses caractéristiques fonctionnelles]*
> **1.** [Considérée dans son apparence extérieure]
> **a)** [Parties constitutives]

[9] Les exemples et les citations sont souvent des paroles d'autorité. Il est rare qu'un exemple soit tiré d'un corpus *ad hoc* dont l'origine est mal connue.

[…]
b) [Caractérisations d'aspect]
α) *Main courte, frêle, grosse, noueuse, poilue; main d'homme.*
[…]
β) *Main brune, chaude, desséchée, molle.*

Non seulement la façon de numéroter le contenu dictionnairique est un trait formel saillant d'un dictionnaire, mais chaque bloc lexicographique a aussi ses formalismes spécifiques : le point après chaque numérotation, la lettre de numérotation en majuscule, la parenthèse après les minuscules, les tirets cadratins, les crochets qui entourent les textes, etc. Les lexicographes ont également recours à des marqueurs métalexicaux, seuls ou combinés, qui sont des indicateurs importants des formalismes du dictionnaire. Certains de ces traits sont communs à tous les dictionnaires et d'autres sont spécifiques à un ou plusieurs dictionnaires : les traits obéissent à un ordre de présentation stricte qui ressemble à un système indexé ; chaque trait formel est un indice de modélisation dans une opération rationaliste, qui est lié à un type de contenu lexicographique, p. ex., la structuration de la microstructure du *TLFi* peut être formalisée comme :

- Entrée ([n° de section **Sens 1** [n° romain **Sens 1.1.** [n° en lettre majuscule latine **Sens 1.1.1.** [n° en chiffre [n° en lettre minuscule avec parenthèse **Sens 1.1.1.1.** [n° en lettre grecque minuscule **Phraséologismes transparents**]]]]]]) ;

les niveaux dans cette structure sont interdépendants et l'accès à un niveau peut se faire à partir d'autres niveaux, p. ex., si l'on extrait le phraséologisme *Main brune*, on peut également extraire la relation d'affiliation entre ce phraséologisme et **Sens 1** qui est « [Organe d'un être animé] ».

2.2. Moule phraséologique

Les unités polylexicales ne sont pas traitées au même titre que les unités monolexicales (mots vedettes) dans les dictionnaires susmentionnés ; elles leur sont rattachées et n'ont pas d'entrée individuelle. Cependant, les rédacteurs réservent un traitement spécifique aux phraséologismes qui sont introduits dans le dictionnaire sous différentes formes :

- Marquage par étiquette. Il y a deux types d'étiquettes :
  1. Étiquettes domaniales (*ART MILIT.* (art militaire), *PAPET.* la papeterie, etc.) ;
  2. Étiquettes métalexicales (*Expressions, Au fig.,* (Au figuré), *P. méton.* (par métonymie)). Certaines étiquettes sont réservées aux phraséologismes, p. ex., l'étiquette SYNT. énumère des phraséologismes transparents qui peuvent être interprétés de manière compositionnelle :
     - **SYNT.** *Main longue, grande, petite, énorme, allongée, carrée, fuselée, large, délicate, fluette, forte, épaisse, fine, gonflée, grasse, maigre, potelée ; main crochue, décharnée, déformée, musclée, tordue ; main velue ; main rhumatoïde ; main charmante ; belle, jolie, pauvre, vieille main.*

Dans cet exemple, deux types de structures phraséologiques sont classées sous l'étiquette SYNT. :

a) La première : la base (= l'entrée) suivie du collocatif ; la combinaison se décline en deux variations comme suit :
   1. Base + collocatif énuméré, les collocatifs étant séparés par une virgule, la suite d'énumération se terminant par un point-virgule : *main propre, sale, soignée, souillée, terreuse* ;
   2. Base répétée + collocatif énuméré, le tout séparé par un point-virgule : *main velue ; main rhumatoïde ; main charmante*[10] ;

b) La seconde : la base (≠l'entrée) suivie du collocatif (l'entrée) : *belle, jolie, pauvre, vieille main*[11].

- Absence du marquage par étiquette. Il existe trois types :
  1. Les renvois internes de l'article : *À main droite, à main gauche. V. infra I A 2 c.*
  2. Les renvois externes : *Proverbes. Mains froides, chaudes amours (ou mains froides et cœur chaud). V. chaud I B 2 a.*
  3. Les phraséologismes opaques : ils sont suivis d'une définition, car leur opacité sémantique ne permet pas à l'utilisateur de les interpréter littéralement. Ces phraséologismes ne sont pas des exemples fabriqués et jouissent en réalité du statut de sous-entrée, p. ex. :

     Nu comme la main. Complètement nu ou dénudé. *C'est la lande, nue comme la main* (GIONO, *Colline*, 1929, p. 90). *Modigliani, nu comme la main et beau comme saint Jean-Baptiste* (CENDRARS, *Bourlinguer*, 1948, p. 201). *Chauve comme la main.* Complètement chauve. *Le petit qui prenait tant de truites dans la Nère, chauve comme la main sur le dessus du crâne, mais poilu du menton comme un bouc* (GENEVOIX, *Raboliot*, 1925, p. 199).

     Dans cet exemple, *nu comme la main* et *chauve comme la main* sont suivis respectivement d'une définition et d'une citation, comme une entrée monolexicale.
- Les remarques lexicologiques et d'usages : Les phraséologismes marqués et non marqués sont tous liés aux définitions, p. ex., les deux phraséologismes susmentionnés sont rattachés aux acceptions suivantes :
  - [Comme élément de référence]
    a) [d'aspect ou de modalité d'existence, le plus souvent en compar.],

  qui sont elles-mêmes rangées sous la définition **I** (voir § 1.3.).

Qu'ils soient marqués ou non marqués, les phraséologismes sont des « occurrences » de moules linguistiques. En donnant comme exemples les

[10] Une expérience similaire a été menée sur le *DAF* (voir Mejri & Zhu, 2020). Les configurations des collocations sont hautement ressemblantes dans les deux dictionnaires. En revanche, le *DAF* utilise des séparateurs différents, p. ex., au lieu d'un point-virgule, il utilise une virgule pour séparer des collocations de ce type.

[11] Ce type de collocation clôt souvent la liste des collocations sous l'étiquette SYNT.

phraséologismes, le dictionnaire impose un raisonnement inductif chez les utilisateurs concernant :

- la combinatoire : le dictionnaire liste des combinaisons prototypiques autour du mot vedette pour donner un aperçu de ses combinaisons usuelles ;
- le lien sémantique explicite : dans l'entrée « main » (*GR*) :
    - ◆ Donner*, (1636) prêter la main à qqn pour faire qqch. ➞ Aider ; appui ; main-forte.,

  le lien entre le phraséologisme *donner/prêter la main à qqn pour faire qqch* est mis en évidence avec *aider*, *appui*, *main-forte* ;
- le lien sémantique implicite : celui qu'entretiennent les phraséologismes et les définitions du mot vedette[12] et celui entre la base et le collocatif[13] du même phraséologisme (voir § 2.3.) sont implicites.

Les phraséologismes transparents font émerger des moules phraséologiques ; ils sont « donnés » par le dictionnaire et servent d'exemples de combinatoires à caractère figé, car il existe en réalité un grand nombre de variations. Le dictionnaire donne une forme globale et lexicalisée qui récapitule les combinaisons syntaxiques et lexicales possibles. Cette forme globale s'accompagne souvent de variations qui, à leur tour, confirment la globalité de la forme, p. ex., les syntagmes qui suivent l'étiquette SYNT. stipulent que :

- syntaxiquement et lexicalement, la forme du moule phraséologique est soit « X + main » soit « main + X » ;
- grammaticalement, les syntagmes en question sont nominaux et le X est adjectival ;
- sémantiquement, les candidats à X sont sémantiquement proches.

Les deux premiers aspects sont plus ou moins mis en évidence par le dictionnaire, tandis que le troisième aspect échappe à la structuration du dictionnaire. Par conséquent, un moule phraséologique est nécessairement une forme hybride qui mélange grammaire, lexique, syntaxe et sens.

2.3. Moule linguistique

Le moule phraséologique mène au moule linguistique et le phraséologisme transparent en est un exemple prototypique. L'une des caractéristiques de la présentation des phraséologismes transparents dans une entrée est que les liens sémantiques entre les collocatifs énumérés ne sont pas expliqués dans le dictionnaire. Souvent ignorés par les utilisateurs, ces liens sémantiques doivent faire l'objet d'une induction de la part des utilisateurs. Cette induction prend en compte les paramètres sous forme de moule tels que la grammaire (G), le lexique (L) et le sens (S) :

$$M \rightarrow GLS^{14}$$

[12] Cet aspect est pris en compte par certains dictionnaires informatisés, p. ex., dans « Recherche complexe » du *TLFi*, les utilisateurs peuvent lancer des recherches en précisant la relation de dépendance entre différents objets à l'intérieur de l'article d'une entrée.

[13] La plupart des dictionnaires ne distinguent pas la base du collocatif : le mot vedette peut être la base ou le collocatif d'un phraséologisme. Mais certains dictionnaires s'adjoignent des renvois pour mettre en relation deux entrées.

[14] L'idée est empruntée à la notion d'« unité de la troisième articulation du langage ». Mejri (2018 : 25)

- pour la partie lexicale, le moule est constitué de morphèmes et d'unités lexicales dont la forme est relativement stable ;
- le moule est doté d'une catégorie grammaticale ;
- s'y ajoute le sens qui est modélisable.

Prenons comme exemple l'entrée *Prendre* (dans le *TLFi*), sous l'objet « Construction » :

- **Qqn prend qqn/qqc. + compl. indiquant la partie saisie.**

qui se décline en deux formes :

- **Prendre qqn/qqc. à + subst.**
- **Prendre qqn/qqc. par + subst**.

La première forme donne lieu à : *prendre qqn à la gorge, prendre qqn au mot, prendre qqn à la lettre*[15], etc. ; la seconde variation donne lieu à : *prendre par le cou, prendre par la taille,* etc. Formellement, ces deux constructions partagent la même construction syntaxique ; « Prendre qqn/qqc. » et « subst. » sont les deux éléments qui constituent les invariants de la construction ; les éléments susceptibles de varier sont :

- la forme de « qqn/qqc » qui désigne soit un humain soit un objet non animé ;
- la préposition qui est soit « à » soit « par » ;
- la forme de « subst. ».

On propose de schématiser cette forme comme suit :

$$\underbrace{\textbf{Prendre}_{\text{lexique}} + \textbf{qqn/qqc}^{16} + \textbf{à/par}_{\text{lexique}} + \textbf{subst.}}_{\text{Construction syntaxique}}$$

Les exemples donnés par le *TLFi* montrent que la position de « subst. » est saturée par *la gorge, le cou, la taille*[17], etc., qui désignent des parties du corps humain. Cette orientation sémantique étant relativement stable, elle ajoute du sens à l'équation précédente :

$$\underbrace{\textbf{Prendre}_{\text{lexique}} + \textbf{qqn/qqc} + \textbf{à/par}_{\text{lexique}} + \textbf{subst.}_{\text{sens}}}_{\text{Construction syntaxique}}$$

Si l'on raisonne en termes de classe d'objets, ces termes appartiennent à la classe *<parties du corps>* (Gross 2008). Cependant, les classes d'objets sont créées et nommées par les linguistes par besoin de description ; elles ne sont pas issues d'une donnée objective et tangible. Dans notre expérience, les classes que l'on nomme les « classes sémantiques »[18] sont repérées automatiquement par notre programme informatique en croisant des items définitoires (Zhu 2019 ; Mejri & Zhu 2020). Ce faisant, sont regroupés les mots vedettes du dictionnaire partageant les mêmes items lexicaux, qui figurent dans les définitions respectives de ces mots. De ce fait, notre

stipule : « (…) l'idée que 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 ».

[15] *Prendre qqn au mot* et *prendre qqn à la lettre* sont marqués par l'étiquette « Loc. » dans le dictionnaire.

[16] *Qqn/qqc* subit également des contraintes sémantiques dans cette configuration.

[17] Ce sont des exemples de phraséologismes donnés par le *TLFi* ; ils ne comprennent pas les expressions figées telles que *prendre qqn à la lettre, prendre le taureau par les cornes,* etc.

[18] La notion de *classe sémantique* s'inspire des « classes d'objets » (Gross, 1994, 2008). Mais nous ne créons pas de classe manuellement ; la création de classes se fait de manière automatique à partir des données dictionnairiques, notamment des définitions.

démarche réduit une part de subjectivité dans la conception des classes et automatise ce processus.

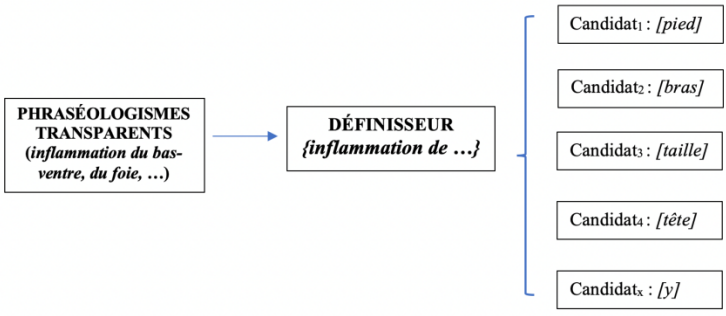
3. DEMONSTRATION

Notre expérience a deux objectifs. Le premier, empirique, consiste à modéliser les moules à partir des données lexicographiques ; le second, heuristique, a pour objectif de poser les bases d'une méthode pour repérer des moules, générer des phraséologismes et les projeter dans un grand corpus pour vérifier leurs attestations. En termes de technologie, nous utilisons le langage de programmation Python[19], qui est adapté à la modélisation des formalismes et aux traitements récursifs des chaînes de caractères.

3.1. Modélisation des moules

Pour illustrer notre démarche, nous prenons l'exemple du définisseur ***{inflammation de …}*** tiré du mot vedette *inflammation* du *TLFi*. Cette structure, extraite des phraséologismes transparents (*inflammation du bas-ventre, du foie, ...*), comporte une position ouverte (représentée par les points de suspension). En examinant les définitions des collocatifs (*bas-ventre, foie, gencive, intestin,* etc.) qui saturent cette position, nous avons repéré des items définitoires communs à ces unités (archisémèmes) tels que *région*, *partie*, *organe*, etc. Cela peut donner suite à deux traitements : le premier traitement consiste à isoler les items définitoires qui sont des archisémèmes des noms de partie du corps ou d'organe. Ils se trouvent vraisemblablement dans d'autres entrées (unités monolexicales) qui sont des « candidats » pour la génération, et leur définition respective contient l'un de ces items. Dans Schéma 1, ***{inflammation de …}*** reçoit une série de candidats *([pied], [bras], [taille]*, etc.) qui appartiennent à l'archisémème [partie du corps]. Ces candidats saturent ensuite un à un la position ouverte du définisseur pour créer des candidats de phraséologismes. Le second consiste à combiner à nouveau les candidats avec les items définitoires pour former des items définitoires complexes, c'est ce que nous démontrons dans cet article.

Schéma 1 Génération (récursive) de phraséologismes transparents

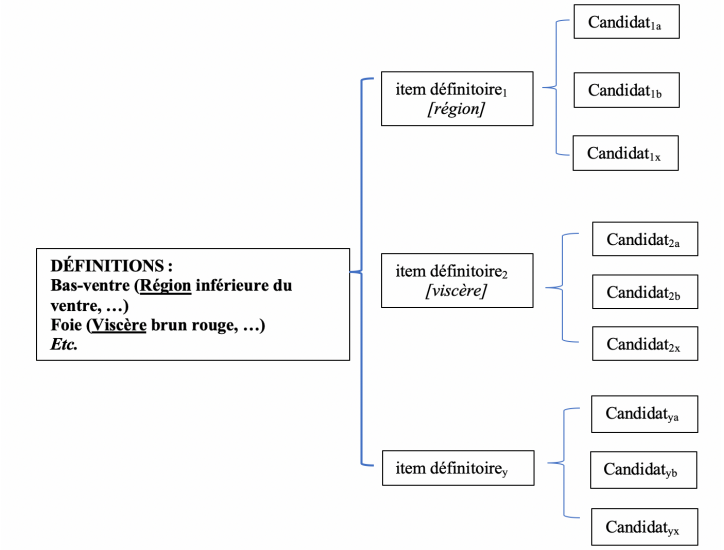


[19] Nous choisissons d'utiliser le langage de script, qui est *open source*, pour simuler les processus itératifs ou récursifs des opérations lexicographiques.

3.2. Construction automatique des classes sémantiques

Le processus de génération de phraséologismes est basé sur la cohérence sémantique des candidats par rapport aux phraséologismes transparents initiaux. Comme le démontre Schéma 2, les collocatifs *bas-ventre, foie, gencive, intestin, iris, matrice, muqueuse, rein* et *vessie* sont eux-mêmes définis par les items définitoires suivants : « Région », « Partie », « Membrane », « Viscère », « Organe » et « Poche »[20].

Schéma 2 Repérage automatique des candidats



[20] Les items définitoires dont les définitions sont tirées du *TLFi* sont les suivants :

- **Région** est l'item définitoire de *bas-ventre* dont la définition est « région inférieure du ventre, au-dessous du nombril » ainsi que de *reins* (Région lombaire) ;
- **Partie** est l'item définitoire de *gencive* (partie du parodonte qui recouvre les maxillaires et adhère fortement au collet des dents) et d'*intestin* (partie du tube digestif qui va de l'estomac à l'anus) ;
- **Membrane** est l'item définitoire de *iris* (« membrane arrondie, rétractile et diversement pigmentée, située au centre de la partie antérieure de l'œil, en arrière de la cornée et en avant du cristallin et qui est percée en son centre d'un orifice, la pupille ») et de *muqueuse* (membrane de revêtement des cavités naturelles de l'organisme (tégument interne) constituée d'un épithélium simple ou stratifié et de tissu conjonctif sous-jacent (chorion), qui contient des vaisseaux et nerfs) ;
- **Viscère** est l'item définitoire de *foie* (« viscère brun rouge, volumineux, de l'homme et des principaux vertébrés, situé dans l'hypocondre droit et une partie de la région épigastrique, sécrétant la bile et exerçant de multiples autres fonctions ») ;
- **Organe** est l'item définitoire de *matrice* (organe de l'appareil générateur de la femme et des mammifères femelles, situé dans la cavité pelvienne, destiné à contenir l'embryon et le fœtus jusqu'à son complet développement) ;
- **Poche** de *vessie* (poche membraneuse dans laquelle s'accumule l'urine sécrétée par les reins avant d'être expulsée par le canal de l'urètre).

Notre programme informatique retient l'entrée si l'une de ses définitions contient au moins l'un des items définitoires et qu'il s'agit de l'humain, à l'aide de l'algorithme suivant (*n* étant l'index de [définition] dans l'article d'une entrée, qui est lui-même indexé phrase par phrase) :

if lemme [item définitoire] **in** [définition] **and if** lemme[humain] **in** $\bigcup_{n+2}^{n-2}$ n
return [définition]

Le programme extrait la définition si le lemme de l'item définitoire s'y trouve et si le lemme de « humain » figure dans un cotexte allant de deux phrases à gauche de la définition à deux phrases à droite de la définition. Nous obtenons 61 entrées dont[21] :

AINE, subst. fém. : Anatomie Partie du corps humain situé entre le bas-ventre et le haut de la cuisse

ABDOMEN, subst. masc. : Partie postérieure du corps de certains invertébrés (arthropodes)

ANTÉRIEUR, EURE, adj. et subst. : ANAT. Partie du corps et en particulier muscle, occupant une place en avant

[…]

Puis les candidats (les mots vedettes) se combinent à nouveau avec les items définitoires de la première extraction. Ainsi, nous générons 366 candidats de phraséologismes (61 candidats d'unités monolexicales * 6 items définitoires) tels que : *région d'aine, partie d'aine, membrane d'aine, viscère d'aine, organe d'aine, poche d'aine, région d'abdomen, partie d'abdomen, membrane d'abdomen*, etc.

3.3. Projection dans un grand corpus

Nous nous référons notamment aux travaux de Colson (2017a, 2017b) sur l'exploitation d'un corpus en vue de l'extraction et de la modélisation des phraséologismes, selon une approche probabiliste. Dans notre expérience, le corpus nous sert de support de vérification afin de tester la validité des phraséologismes générés. Ainsi, les candidats générés sont projetés dans le Corpus French Web 2017 (frTenTen17) [22] . Il va de soi que tous les candidats ne sont pas valables linguistiquement, même s'ils sont générés selon la pertinence sémantique. Après vérification, nous constatons que certains phraséologismes recueillent *talis qualis* des occurrences dans le corpus :

Partie de peau : (68 occurrences)

il existe deux catégories de démangeaisons : - les démangeaisons localisées, c'est-à-dire n'intéressant qu'une **partie de peau** limitée ; - les démangeaisons généralisées, intéressant l'ensemble du revêtement cutané. (http://sante.lefigaro.fr/sante/symptome/demangeaisons/quest-ce-que-cest)

Poche de vessie : (1 occurrence)

le professeur vient de m'annoncer qu'il va subir une « dernière » grosse opération. On doit lui rajouter de la **poche de vessie**, obstruer le poche de la vessie pour la continence et créer un canal permettant (à la place du nombril) (http://www.pediatric-surgery.org/spip.php? article87) ;

[21] Les définitions sont tirées du *TLFi*.
[22] Le corpus contient 5 752 261 039 mots (consulté le 16 juillet 2020).

D'autres recueillent des occurrences avec des modifications morphologiques, grâce à la lemmatisation du moteur de recherche :

Région de corps[23] (337 occurrences) :

Ces derniers mesurent la différence d'intensité entre l'entrée et la sortie d'un faisceau de rayon X dans une **région du corps**. Le tube émetteur de rayons X et les capteurs tournent autour du corps au cours de l'examen. La mesure de l'atténuation (http://sante.lefigaro.fr/sante/examen/scanner-rhumatologie/quest-ce-que-cest)

Poche de ventre (1 occurrence) :

On peut faire un régime alimentaire pour changer le mode d'alimentation en fonction des objectifs qu'on veut atteindre : maigreur, avoir une taille normale, diminution des hanches, diminuer les **poches du ventre** ... ce régime doit être fait sous le regard vigilant d'un diététicien afin de ne pas subir des carences quelconques. (http://infos-regimes.fr/)

Parmi les phraséologismes générés, 79 séquences trouvent des occurrences dans le corpus dont 47 sont pertinentes (voir Annexe 1). Dans les 32 occurrences non pertinentes, les items définitoires homonymiques ou hautement polysémiques tels que *poche* et *partie* génèrent plus de résultats faux positifs que ceux des autres items, p. ex., la quasi-totalité des occurrences de « partie de jambe » sont imputées à *partie de jambes en l'air* qui est une séquence figée opaque.

CONCLUSION

Notre expérience simule les opérations rationalistes humaines, notamment les opérations de mise en relation des données lexicographiques. Nous avons rencontré les difficultés suivantes dans l'expérience :

- la modélisation des données dictionnairiques est coûteuse en termes d'opérations, les formalismes similaires de différents types de contenus dictionnairiques posent problème dans la récupération des différents types de données ;
- l'isolation des items définitoires est souvent confrontée à la polysémie et à l'homonymie de certaines unités comme *partie*, *région*, etc. ;
- l'imprécision du lemmatiseur nous induit souvent en erreur.

Cette expérience fait partie d'une série de réflexions que nous menons en matière de phraséologie et de sémantique. La méthode et l'outil informatique que nous avons développés pour l'expérience peuvent contribuer à la création des classes sémantiques et à la construction d'un dictionnaire de phraséologismes exploitable par la machine, à partir d'un dictionnaire électronique.

23 Ce syntagme est généré formellement par le script. Sketch Engine lemmatisant automatiquement les mots, il repère ainsi l'occurrence « région du corps » dans son corpus.

Références

ADAM J.-M. (2015), *La linguistique textuelle*, Paris, Armand Colin.

BECK D., GERDES K., MILICEVIC J. & POLGUÈRE A. (eds.) (2009), *Proceedings of the Fourth* International *Conference on Meaning-Text Theory*, Montreal, OLST.

BENVENISTE E. (1974), *Problèmes de linguistique générale*, tome II, Paris, Gallimard.

BERNET C. (2007), « Le TLFi ou les infortunes de la lexicographie électronique », *Mots* 84, 85-100.

CHAROLLES M. (2001), « De la phrase au discours : quelles relations », *in* Rousseau A. (éd.), *La sémantique des relations*, Université de Lille III, 237-260.

COLSON J.-P. (2017a), « A la croisée des corpus et de la phraséologie : une proposition d'outil informatique », *Studii de lingvistică* 7, 13-26.

COLSON J.-P. (2017b), « The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions », *in* Corpas Pastor G. et Mitkov R. (eds), *EUROPHRAS 2017: Computational and Corpus-Based Phraseology*, 16-28.

COMBETTES B. (1987), « Texte, discours, cohérence », *Repères pour la rénovation de l'enseignement du français* 71, 85-91.

GROSS G. (1994), « Classes d'objets et description des verbes », *Langages* 115, 15-30.

GROSS G. (2008), « Les classes d'objets », *Lalies*, Presses de l'École normale supérieure, 111-165.

HARRIS Z. S. & DUBOIS-CHARLIER F. (trad.) (1969), « Analyse du discours », *Langages* 13, 8-45.

KILGARRIFF A. (2005), « Informatique et dictionnairique », *Revue française de linguistique appliquée* X-2, 95-102.

KOSEM I., JAKUBÍČEK M., KALLAS J. & KREK S. (éds.) (2015), *Electronic lexicography in the 21st century: linking lexical data in the digital age*, *Proceedings of the eLex 2015 conference*, Ljubljana, Trojina, Institute for Applied Slovene Studies [consulté le 22 novembre 2020, https://elex.link/elex2015/wp-content/uploads/eLex2015-proceedings.pdf].

MANGEOT M. & BELLYNCK V. (2018), « Un devin de microstructures pour importer ou normaliser des ressources lexicales », *Actes de la conférence Lexicologie Terminologie Traduction LTT 2018*, [consulté le 22 novembre 2020, https://hal.archives-ouvertes.fr/hal-02063815/file/LTT2018_MMVB.pdf].

MARTIN R. (2001), *Sémantique et automate*. Coll. *Écritures électroniques*, Paris, PUF.

MARTINEZ C. (2009), « Une base de données des entrées et sorties dans la nomenclature d'un corpus de dictionnaires : présentation et exploitation », *Études de linguistique appliquée* 156, 499-509.

MEJRI S. (2018), « La phraséologie française : synthèse, acquis théorique et descriptifs », *Le Français Moderne*, Paris, CILF, 5-32.

MEJRI S. (à paraître), « Le mot démocratie dans le dictionnaire : structure et représentation réticulaire », *Les Cahiers du dictionnaire* 12, Paris, Classiques Garnier.

MEJRI S. & ZHU L. (2020), « Données dictionnairiques informatisées : phraséologie et inférence », *Le Français moderne*, Paris, CILF, 102-136.

MENESES-LERIN L. (éd.) (2017), *Corpus et ressources numériques : nouveaux paradigmes de recherche en linguistique, en didactique et en traduction*, *Studii de lingvistică* 7, Oradea, Editura Universităţii din Oradea.

MOURLHON-DALLIES F., RAKOTONOELINA, F. & REBOUL-TOURE, S. (2004), *Les* discours *de l'internet : nouveaux corpus, nouveaux modèles ?,* Paris, Presses Sorbonne Nouvelle.

NEVEU F. (2011), *Dictionnaire des sciences du langage*, Paris, Armand Colin.

PECHEUX M. (1969), *Analyse automatique du discours*, Paris, Dunod.

PIERREL, J.-M. (2003), « Un ensemble de ressources de référence pour l'étude du français : TLFI, FRANTEXT et le logiciel STELLA », *Revue québécoise de linguistique* 32, 155-176.

RASTIER F. (1998), « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage », *Langages* 129, 97-111.

REY-DEBOVE J. (1985), « Le métalangage en perspective », *Documentation et* recherche *en linguistique allemande contemporain* 32, 21-32.

SARFATI G.-É. (1997), *Éléments d'analyse du discours*, Paris, Nathan Université.

SINCLAIR J. (1996), *Preliminary Recommandations on corpus Typology*, EAGLES [consulté le 24 juin 2020, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.1988&rep=rep1&type=pdf].

ZHU L. (2016), « Pour une notion de moule dans le figement », *Les Cahiers du dictionnaire* 8, Paris, Classiques Garnier, 97-109.

ZHU L. (2019), « Moule locutionnel lexicographique et traitement des phraséologismes », *Les Cahiers du dictionnaire* 11, Paris, Classiques Garnier, 147-163.

Corpus

Sketch Engine – *Corpus French Web 2017*, https://www.sketchengine.eu