Données dictionnairiques informatisées. Réseaux inférentiels et phraséologiques

Salah Mejri, Lichao Zhu

Données dictionnairiques informatisées

Réseaux inférentiels et phraséologiques[1]

Salah MEJRI et Lichao ZHU

Avec l'avènement des dictionnaires informatisés[2], on assiste à une prise de conscience des perspectives que l'informatisation offre à la linguistique, la lexicographie et le traitement automatique. Avec le développement des outils informatiques, l'objet de la linguistique change de forme : la taille des corpus est de plus en plus grande, la manière dont on traite les questions de syntaxe, lexique, morphologie, sémantique se modifie (on n'est plus dans les exemples bien choisis, adaptés et taillés sur mesure ; les notions de paradigme et de combinatoire ne couvrent plus les mêmes réalités ; le sens n'est plus cette nébuleuse qu'on cherche à cerner à travers les relations sémantiques et lexicales limitées à un nombre réduit d'unités lexicales). Le dictionnaire lui-même échappe ainsi à ses élaborateurs. La navigation hypertextuelle et la recherche avancée impliquant plusieurs paramètres modifient totalement l'objet même du dictionnaire : on a accès à toutes les occurrences d'un mot dans le dictionnaire, qu'il s'agisse d'entrées, de définitions, d'exemples, de citations, d'informations étymologiques, etc. Ainsi l'utilisateur se trouve-t-il confronté à des données qui échappent à la conscience des concepteurs même du dictionnaire. Grâce à ces changements, dont nous décrivons d'autres aspects, les linguistes pourraient accéder à une mine d'informations relatives à la langue, objet de leurs analyses.

Le dictionnaire, la seule œuvre linguistique revendiquant l'exhaustivité dans sa description des langues, ainsi transformé, pourrait représenter un corpus unique, pour une meilleure connaissance du fonctionnement des langues. Il est, à notre connaissance, le seul ouvrage métalinguistique qui revendique la description d'une langue à travers son lexique, considérant cet aspect de la langue comme le point de départ pour rendre compte de toutes les autres dimensions. Le choisir comme corpus, c'est situer la recherche au niveau de la réflexion même sur la langue. Cette dimension autonymique n'exclut nullement l'usage « mondain » (non autonymique), puisque les dictionnaires comportent tout un sous-corpus d'exemples et de citations dont la fonction consiste à illustrer les descriptions, les compléter et les enrichir.

Partant de ces considérations, nous essaierons de voir l'impact de la numérisation sur les potentialités des dictionnaires initialement conçus pour être consultés dans des versions papiers, d'examiner les obstacles qui font de l'automatisation de ces données une tâche très compliquée, et pour finir, de proposer une méthode de génération automatique des réseaux inférentiels et phraséologiques.

1. Le dictionnaire, lieu d'une expertise linguistique

Le dictionnaire est un ouvrage particulier ; il est l'objet d'une appréciation ambivalente. D'un côté, il est considéré comme la référence en matière de description linguistique ; de l'autre, on lui reproche de renfermer toutes sortes d'étrangetés[3]. Il faut préciser que l'appréciation positive se situe du côté du grand public qui voit dans le dictionnaire l'institution reconnue socialement comme étant la référence en matière de lexique. L'appréciation beaucoup plus réservée est du côté des spécialistes, notamment les linguistes. Nous ne reprenons pas les arguments des uns et des autres, – tel n'est pas notre propos –, nous présentons les éléments du dictionnaire que nous considérons comme pertinents pour les objectifs que nous nous sommes fixés. Ils sont au nombre de cinq :

- Le dictionnaire est un ouvrage qui affiche l'ambition de décrire la langue dans sa complexité, de la manière la plus exhaustive et la plus précise possible, même si l'on sait par ailleurs que le lexique d'une langue est par définition ouvert et que l'absence de limites claires entre langue générale et langue des domaines spécialisés conduit à des approximations le plus souvent subjectives et nécessairement floues, les marges du lexique étant souvent ignorées[4] ;

[1] Nous remercions les deux relecteurs pour la pertinence de leurs remarques, qui ont aidé à améliorer la version initiale tant sur le plan de la forme que sur celui du contenu.

[2] À distinguer des dictionnaires électroniques dont l'ambition est l'élaboration d'un dictionnaire entièrement automatisé, permettant des exploitations dans le traitement automatique des langues. Le dictionnaire informatisé est un dictionnaire traditionnel, numérisé, permettant des recherches hypertextuelles.

[3] Voir par exemple D. Corbin (1982).

[4] Voir par exemple les lexiques très changeants et éphémères des jeunes des cités, ou ceux qui relèvent des variétés francophones, auxquels sont consacrés des ouvrages de plus en plus nombreux, qu'il s'agisse de dictionnaires ou non. Voir en particulier P. Goudaillier (1997), J.-M. Klinkenberg (1997).

- Le dictionnaire est un genre[5] ; son discours obéit à des normes : il se caractérise par une organisation générale qui repose sur deux structures. La microstructure comporte l'ensemble des fonctionnalités assurées par l'article consacré à chaque entrée (vedette + un ensemble d'informations [de prédicats] associé à cette entrée). Leur nature et leur nombre varient en fonction de la nature du dictionnaire, du public visé et des choix des lexicographes. On constate par ailleurs que, dans la pratique, il est rare que les règles que les lexicographes s'imposent soient scrupuleusement respectées (voir par exemple § 3.3, pour les marqueurs métalexicaux qui commencent parfois par une majuscule, dans d'autres cas par une minuscule : Fig. 2 et Fig. 3) ;

- Le dictionnaire est un ouvrage didactique : son premier objectif est de mettre à la disposition de ses lecteurs l'ensemble des informations jugées nécessaires à l'amélioration leurs compétences linguistiques. Même si la doxa privilégie la définition fournie pour chaque entrée, – vision communément admise par le public –, il est toujours utile de rappeler que le dictionnaire est extrêmement riche en données utiles pour la connaissance des différents aspects de la langue : catégorie grammaticale, étymologie, orthographe, prononciation, datation, morphologie, constructions syntaxiques, différentes significations, domaines d'emploi, registres, champs lexicaux et sémantiques, dérivés et composés, unités phraséologiques, homonymes, synonymes, antonymes, variantes, etc. Le tout est généralement illustré et complété par des exemples et des citations. Le recours aux exemples et aux citations, en plus de la fonction illustrative qui leur est assignée, est souvent à l'origine d'un complément d'informations ne figurant pas dans le reste de l'article ;

- La macrostructure du dictionnaire, qui n'est pas directement accessible, est le lieu d'informations portant sur l'organisation de cet ouvrage : l'ensemble des renvois entre articles en constitue l'ossature. Certains dictionnaires en font une spécificité, comme c'est le cas pour les dictionnaires analogiques, ou ceux qui structurent leur description en fonction des radicaux partagés par les groupes de mots (famille de mots) ;

- Une spécificité des dictionnaires, rarement retenue, concerne la dimension idiomatique de l'expression et, par conséquent, la dimension culturelle. Bien que très prégnante, la présence du culturel passe inaperçue aux yeux des usagers natifs. Mais cette dimension est bien réelle pour toutes les variantes de la même langue territorialement séparées (exemple de la variation dans la francophonie, l'hispanophonie, la lusophonie, l'arabophonie, etc.). Malgré les efforts que certains dictionnaires fournissent dans la description de cette dimension, celle-ci n'obéit pas pour autant à une systématisation. On n'y trouve pas par exemple une distinction claire entre la manière dont on dit les choses dans une langue, c'est-à-dire les mots ou la combinaison de plusieurs mots, et l'ensemble des inférences rattachées à l'usage des mots.

Toutes ces spécificités en font un ouvrage problématique : il affiche l'exhaustivité sans jamais l'atteindre ; il revendique la clarté et la cohérence tout en véhiculant des incohérences et tout en recelant des ambiguïtés et des approximations ; il revendique de fait une expertise dans la description des langues tout en faisant l'objet de critiques sévères de la part des spécialistes des langues. C'est autour de ce dernier point que se cristallisent les critiques, dont nous retenons les points suivants :

- La tradition lexicographique montre que les lexicographes ne sont pas nécessairement des spécialistes de la langue (grammairiens, rhétoriciens, linguistes, etc.), mais des gens à qui on reconnaît une bonne connaissance et une maîtrise de la langue (auteurs la plupart du temps). Cette reconnaissance, qu'on pourrait appliquer aux Académiciens, n'est pas mise en saillance dans des cas aussi prestigieux dans ce domaine que celui de *Larousse* et celui du *Robert*. Jean Pruvost situe le passage de la linguistique à la lexicographie et vice-versa à la deuxième moitié du 20e siècle : « [...] en plein développement de la linguistique, dans la mouvance structuraliste, une autre période commençait où les lexicographes devenaient linguistes et où les lexicologues entraient dans l'aventure lexicographique » (2002, p. 71). Se détache de ce tableau la grande entreprise du *TLF*, qui a été initiée par des linguistes comme Paul Imbs et qui a pris naissance dans la dynamique de la recherche linguistique. La direction en a été confiée, après Paul Imbs, à Bernard Quemada[6] et Robert Martin. Robert Martin, qui a associé dès le début de sa carrière, lexicographie et linguistique, veille toujours sur le *Dictionnaire du moyen français*[7]. Sa recherche est pétrie de données lexicographiques, dont on retient particulièrement *Sémantique et automate* (2001).

- Même si le dictionnaire n'est pas nécessairement l'œuvre de spécialistes des langues, il ne dispose pas moins d'un appareil terminologique qui trahit une vision de la langue et reflète, d'une manière ou d'une autre, le savoir partagé sur la langue décrite. Un tel savoir est présenté le plus souvent d'une manière diffuse dans le dictionnaire. Le mode éclaté de sa présentation ne lui donne pas la saillance suffisante pour saisir ce savoir en tant que tel, il n'est pas pour autant très important : certaines données grammaticales, comme la nature grammaticale des entrées, leur morphologie, l'étymologie, les relations sémantiques, etc. sont systématiquement mentionnées ; parallèlement, on découvre dans les articles une information foisonnante sur les constructions syntaxiques, les différentes contraintes d'emploi, les irrégularités combinatoires, les emplois des modes, temps,

[5] Le dictionnaire prototypique est le dictionnaire de langue. Le dictionnaire encyclopédique, qui ne nous intéresse pas ici, obéit à d'autres types de structuration.
[6] *Dictionnaire et lexicographie* 1, 1990.
[7] Accessible à l'adresse officielle suivante : http://www.atilf.fr/dmf. La version du 31 juillet 2019 comporte 65 720 entrées, 470 125 exemples, environ 200 200 000 caractères.

aspects, etc. Un dictionnaire est, sous cet angle, un vrai manuel de grammaire[8] dont le contenu reflète généralement la norme en la matière. Seules certaines informations sont formellement explicitées. On les retrouve dans la liste des abréviations : la nature grammaticale, les catégories de genre et de nombre, les constructions verbales, les fonctions, etc. Elles sont le plus souvent mélangées avec d'autres signes conventionnels de présentation et occupent dans un dictionnaire comme le *Grand Robert* en 6 volumes (2001) une dizaine de pages.

- Ce genre de codes sémiotiques obéit à une organisation qui ne manque pas de rigueur, nécessaire à la présentation de la matière abondante du dictionnaire ; il trahit néanmoins un certain nombre de catégories ; ce qui donne lieu à des différences parfois notables entre les dictionnaires. Un domaine, celui des registres de langue, n'est pas toujours traité de la même manière d'un dictionnaire à un autre. Des mots comme *flic* et *poulet* sont marqués tous les deux *fam.* (Familier) dans le *Grand Robert*, mais respectivement *pop. et fam.* et *Arg. et pop.* dans le *TLF*. Même si l'on reconnaît que les frontières dans ce domaine ne sont pas très claires, il y en a d'autres où les connaissances sont relativement assez stabilisées, comme c'est le cas de celui des catégories grammaticales (parties du discours) ; pourtant on relève à ce propos soit des silences, soit des partis pris, soit le recours à l'implicite. L'exemple des locutions est assez représentatif des modes de traitement. Une séquence comme *à la carte* est traitée par le *Grand Robert* dans l'article *carte* comme suit :

> - À LA CARTE, Manger à la carte, en choisissant librement sa carte (opposé à *au menu*, à prix fixe). *Repas à la carte.* Par ext. *À la carte* : au choix, Voyages individuels à la carte.

Bien que tous les éléments descriptifs fournis, notamment les syntagmes illustratifs, montrent clairement que cette séquence peut avoir, selon le contexte, la valeur d'une locution adjectivale ou adverbiale, rien n'est explicitement fourni au lecteur.

Dans le traitement de cette locution, le parti pris peut conduire à des erreurs d'interprétation de la part du lecteur. Ainsi en est-il dans cet exemple :

> - Mod. Loc. Adv. (fam.) *Va-comme-je-te-pousse, à la va-comme-je-te-pousse* : n'importe comment, de façon désordonnée. *Ce travail a été fait à la va-comme-je-te-pousse.*
>
> > [...] Ils [nos contemporains] portent au hasard leurs mains innocentes et brusques sur des sièges et des bémols, et vas-y comme je te pousse, Léon-Paul Fargue, *Lanterne magique*, p. 182.

Les questions qui se posent sont les suivantes : Quelle forme correspond à la locution adverbiale ? La première ? La seconde ? Quel statut donner à la forme illustrée par l'exemple ?

Le silence est on ne peut plus clair dans ce cas (l'article *on*) :

> Loc. Avec *pouvoir* et *savoir*, marquant soit un haut degré (*on ne peut plus, on ne peut mieux*, etc.), soit l'indétermination (*on ne sait qui, on ne sait quoi, on ne sait où, on ne sait comment*, etc.) *J'ai tout ça, on ne peut mieux présent à l'esprit*.

Le caractère locutionnel est certes bien souligné, mais rien n'est dit sur la catégorie grammaticale. Le silence est d'autant plus lourd de conséquence qu'il porte sur le fonctionnement de deux types de locutions pouvant fonctionner différemment : avec *on ne peut plus*, il s'agit d'un adverbe qui modifie un adjectif, comme l'illustre l'exemple fourni. Avec les locutions construites avec le verbe *savoir*, on peut avoir toutes sortes d'emplois, comme l'emploi nominal : *Il a ce je ne sais quoi d'imposant.*

On pourrait multiplier les exemples. L'objectif n'est pas d'accabler les lexicographes qui ont beaucoup de mérite dans l'élaboration de ce genre d'ouvrage ; nous voudrions tout simplement montrer la complexité de leurs tâches. Pour construire une charpente descriptive, il faut saisir le degré de complexité de l'objet à décrire. Or la langue est un objet dont la structure générale et l'extrême complexité de fonctionnement échappent encore aux théories linguistiques, encore plus aux approches lexicographiques. Pourtant, ce qui caractérise le dictionnaire, c'est l'ambition affichée de rendre compte de la totalité de la langue. Il est à notre connaissance, à côté des grammaires, la première tentative de simulation[9] jamais menée depuis plusieurs siècles[10]. Or la simulation présuppose une théorie globale par laquelle les lexicographes essaient de simuler la langue. La première difficulté consiste à s'entendre sur la définition même de la théorie. Si l'on opte pour l'approche des biosémioticiens comme Dario Martinelli[11], cité par Jacques François (2017)[12], l'on admet que :

> La *semiosis* est le résultat de l'interaction entre un sujet et un objet, entre une structure et une contre-structure, entre un récepteur et un véhicule de sens. Ces deux parties échangent des informations constamment et mutuellement. En fait l'échange lui-même est le véritable générateur de tout phénomène sémiotique, puisque le second (le véhicule de sens) n'existerait tout simplement pas si le sujet n'était pas affecté par lui et ne l'affectait pas en retour. Toute recherche zoosémiotique, des

[8] La grammaire est l'autre composante de la langue. Une bonne partie en est retenue par les dictionnaires, comme signalé plus haut.
[9] Il ne s'agit pas évidemment de simulation informatique. *Simuler* est pris ici dans le sens que lui donne le *TLF* : « Reproduire artificiellement une situation réelle à des fins de démonstration ou d'explication. »
[10] *Cf.* par exemple la tradition lexicographique arabe.
[11] Dario Martinelli 2010.
[12] *Cf.* Salah Mejri, compte rendu de cet ouvrage dans *Bulletin de la Société de linguistique de Paris*, à paraître, 2020.

phéromones aux chants des baleines, devrait prendre en compte une telle conception, sinon elle risque de pervertir l'essence même du phénomène de *semiosis*[13].

Selon cette vision très générale, la « règle sémantique » qui régit toute la vie de la tique serait l'odeur de l'acide butyrique qui se dégage de la sueur des mammifères, qui la réveille à l'approche des mammifères et lui permet de sucer à son contact le sang dont elle se nourrit. Donald Favareau[14], qui a développé cet exemple, le généralise à tous les mammifères : « […] les propriétés de tout mammifère – que ce soit un homme, un chien, un cerf ou une souris – activent par contrepoint la règle de vie de la tique » (cité par Jacques François, *op. cit*., p. 159).

Si l'homme porte en lui des représentations de son environnement naturel et social, qui ont conditionné son adaptation et son évolution, il a élaboré en plus un système symbolique lui permettant de les fixer dans des formes langagières (gestes, mimiques, vocalisations, images…) qui ont abouti aux langues humaines telles que nous les connaissons. Si notre cognition est en quelque sorte la théorie de tout ce qui nous est extérieur (l'univers dans toutes ses composantes), et conséquemment ce qui nous est intérieur comme les représentations que nous faisons de nous-mêmes, nos émotions, etc., son expression langagière fait de la langue une théorie fixée dans des symboles qui renferment non seulement le monde extérieur mais également la manière dont on l'appréhende et les visions qui s'y attachent. En d'autres termes, cette théorie symbolique n'est pas fondée sur un isomorphisme strict avec le monde « objectal », mais sur une relation diffractée, selon la dynamique de l'interaction entre le sujet parlant et son monde environnant.

Cette complexité fait que toute tentative de simulation de la langue doit intégrer une double précaution méthodologique : la première concerne le caractère métathéorique de la langue (dans le sens de la théorie fournie plus haut) ; la seconde porte sur le caractère autorégulé de la langue qui lui donne, en tant que système sémiotique autonome, une plasticité telle qu'elle intègre continuellement des variations de toutes sortes (du monde extérieur et de sa dynamique interne) tout en se réadaptant aux exigences des principales fonctionnalités régissant son usage au sein de la collectivité et aux aléas des évolutions. L'approche lexicographique se situe, selon cette perspective, dans une position « méta-méta-théorique », une théorie au second degré : toute simulation de la langue est nécessairement une approximation d'une complexité exponentielle. S'ajoute à cette complexité le symbolisme employé pour élaborer la simulation lexicographique, qui n'est rien d'autre que les signes linguistiques eux-mêmes ; ce qui crée toujours l'illusion qu'on traite directement de la langue, alors qu'on manipule en réalité des signes qui renvoient par autonymie à eux-mêmes. Pour résumer cette mise en abyme :

environnement global ↔ représentations cognitives ↔ système langagier ↔ simulation lexicographique

Pour l'élaboration de la simulation lexicographique[15], une théorie langagière gouverne la cohérence de tout discours que les dictionnaires comportent. Cette théorie n'est pas une simple spéculation ; elle a une dimension pragmatique : la description d'une langue. Elle doit satisfaire aux trois conditions générales suivantes :

- La globalité et la cohérence : la globalité du projet lexicographique est assurée par le choix du lexique comme entité pour la description de la langue. Un tel choix n'est pas théoriquement anodin : il a tout l'air de couler de source, mais il marque une prise de conscience de la centralité, dans la description des langues, des unités de la troisième articulation du langage que sont les unités lexicales (Mejri 2018)[16] : un tel choix n'était pas évident dans certaines entreprises lexicographiques : qu'on pense à la première édition du *Dictionnaire de l'Académie* où les entrées ont été disposées selon les racines, « c'est-à-dire, en rangeant tous les mots derivez ou composez après les mots dont ils descendent » (préface de la deuxième édition, 1758), ou à la tradition lexicographique arabe qui a longuement privilégié la matière consonantique, souvent trilitère, porteuse d'un contenu sémantique partagé par l'ensemble des unités à la formation desquelles elle participe. Le recours aux unités lexicales comme point d'entrée, en l'occurrence le mot[17], représente un choix, certes pertinent, mais très difficile à réaliser. Les Académiciens le soulignent bien dans la même préface où ils expliquent pourquoi ils ont choisi de substituer les mots aux racines :

  En effet rien n'est plus penible que d'avoir à déterminer sur un mesme mot les idées diverses & souvent tout opposées, qu'il doit exciter en nous, suivant les differentes manieres dont il peut estre lié avec tous les autres mots de la mesme Langue. (*Ibidem*).

[13] Jacques François (2017), p. 158.

[14] Donald Favareau 2010.

[15] La simulation lexicographique représente une étape vitale pour la conservation, la diffusion et l'apprentissage d'une langue. Evidemment, les conditions *sine qua non* de cette description sont l'écriture et ce qu'on dénomme couramment la grammatisation d'une langue (*i.e.* la doter des descriptions qui rendent compte de son fonctionnement) ; ce qui donne toute leur signification aux entreprises lexicographiques, quelle qu'en soit la forme.

[16] Depuis Martinet, il est d'usage de distinguer la première articulation, par laquelle la segmentation des énoncés donne lieu aux plus petites unités linguistiques ayant un sens (les morphèmes ou les monèmes dans la terminologie de Martinet) et la deuxième articulation dont les unités sont des phonèmes. Cette analyse mérite d'être reprise à la lumière d'une orientation contraire, plus ouvrante, celle du producteur de l'énoncé. Le locuteur est censé partir des unités de la deuxième articulation de Martinet (première pour la triple articulation) pour former des unités douées d'un sens non grammaticalisé (deuxième articulation), auxquelles s'ajoute la dimension grammaticale au moyen d'un encapsuleur morphologiquement marqué ou pas (l'appartenance à une partie du discours) qui assure au système linguistique toute sa puissance, qu'elle soit de nature syntaxique ou autre (pour les détails, voir Mejri 2018).

[17] Le mot étant la forme monolexicale de l'unité de la troisième articulation (*cf.* Mejri 2018).

Dit en d'autres termes, aborder la description de la langue à partir de son lexique, c'est prendre en compte l'ensemble des réseaux que constitue le système de la langue, tout en tenant compte de la combinatoire des unités qui la composent. Que seraient les différentes dimensions linguistiques (phonologie, morphologie, syntaxe, sémantique) sans le support lexical, ce lieu où tout s'entrecroise ?

À la globalité est associée une cohérence qui en découle. Se donner comme objectif de décrire par le menu l'ensemble des caractéristiques jugées pertinentes de chaque unité du lexique, impose une démarche qui repose sur la cohérence générale du discours du dictionnaire. Cette cohérence est assurée par la permanence d'une microstructure reproductible dans la structuration des articles et par l'économie générale qu'on cherche à construire à travers la microstructure qui garantit à l'ensemble de l'édifice lexicographique sa solidité et sa pertinence[18]. Entre la microstructure et la macrostructure, une mésostructure prend en charge, de fait, des éléments qui échappent à la monolexicalité, comme la phraséologie (voir § 3.3.) et les faits pragmatiques et culturels[19].

- La systématicité et le degré de granularité : la systématicité s'observe à travers le type de description dont l'objectif final est de rendre compte du système que la langue représente, un système fait de sous-systèmes avec des jeux de relations diverses et variées, se situant à plusieurs niveaux de leur agencement et concourant toutes à assurer au système une souplesse inégalée qui garantit toutes les fonctionnalités de la langue. Le grand défi d'un dictionnaire réside dans la simulation de la dynamique interne du système linguistique. C'est pourquoi la systématicité doit être corrélée au degré de granularité choisie par chaque dictionnaire. La granularité d'un dictionnaire se construit horizontalement et verticalement et en profondeur, donnant du système une vision tridimensionnelle. L'horizontalité prend en charge l'ensemble des paramètres descriptifs : phonologie, morphologie, syntaxe, pragmatique, etc. Plus ce maillage est riche en paramètres, plus l'information retenue est variée. La verticalité concerne l'ensemble des liens formels et sémantiques entre les unités lexicales : parmi les liens formels, il y a lieu de retenir les configurations morphologiques réalisées dans le cadre de la monolexicalité et dans celui de la polylexicalité, au-delà desquelles se profile l'hypothèse de moules formels générateurs de nouvelles unités[20]. La profondeur s'exprime à travers une sorte d'orientation vectorielle à partir de chaque unité lexicale établissant tous les rapprochements directs ou indirects à partir de n'importe quel ingrédient de l'unité prise comme point de départ[21]. Quand on parle d'ingrédient, on n'exclut aucun aspect de la forme, du contenu et des divers aspects kaléidoscopiques de chaque unité lexicale. Cela impliquerait entre autres les profils dénominatifs, les combinaisons phraséologiques idiomatiques, les inférences culturelles (mythes, religion, histoire, littérature, anthropologie, sociologie, etc.). Plus la granularité est fine et variée, plus l'épaisseur et la densité vectorielles sont importantes[22].

- L'explicitation et la visée d'apprentissage : malgré les différents reproches qu'on peut faire au dictionnaire en rapport avec les implicites, les non-dits, etc., celui-ci demeure néanmoins l'ouvrage le plus explicite dans sa description. Outre les aspects normatifs qu'il véhicule, il a recours à tout un système métalangagier qui mélange symboles, abréviations et signes typographiques[23] pour verser tous les contenus selon une typologie et une conception préétablies. Tout cet arsenal est souvent détaillé dans la préface ou dans un texte spécialement dédié, où l'on en explicite le rôle afin d'en faciliter la consultation. Le tout concourt à fournir le plus de moyens possibles pour rendre compte de l'extrême complexité du système linguistique. On oublie souvent qu'un dictionnaire est également une syntaxe, une morphologie, un système phonologique, une stratification des marques d'emploi, un système lexical, des réseaux référentiels, sémantiques et culturels. L'explication de la nature de ces données est le meilleur garant de l'efficacité descriptive, et par conséquent, de la dimension pédagogique et didactique. Le dictionnaire est l'ouvrage didactique par excellence : nous le consultons pour glaner les informations, précisions ou détails qui nous manquent. S'il ne répond pas à ces besoins et à ces exigences de structuration, il perd de son efficacité pédagogique. Certes il y a des différences de taille, de densité descriptive, d'illustration et de tant d'autres caractéristiques qui changent selon le public visé, les conditions du marché, les moyens techniques disponibles, etc., mais une constante est toujours là : le souci de simplicité et de clarté, indispensable à toute démarche didactique réussie. En plus des grammaires, le dictionnaire est l'ouvrage fondamental dans l'apprentissage et la diffusion de toute langue.

2. L'outil numérique

[18] L'exemple de la réussite dans la durée de quelques dictionnaires est impressionnant : le *Dictionnaire de l'Académie, Larousse, le Robert.*

[19] Pour la mésostructure, voir Xavier-Blanco Escoda et Salah Mejri (2018).

[20] Il s'agit là d'une hypothèse intéressante qui pose le problème des unités de la langue à un niveau strictement sémiotique. Elle soulève entre autres l'idée de l'existence de formes hybrides régulières reproductibles comme c'est le cas dans les systèmes autorégulés. Des travaux en sciences cognitives, en génétique, en informatique, etc. s'intéressent à une telle problématique. Certains la rattachent à la théorie de la Gestalt (voir pour cette notion Fiorenza Toccafondi 1999).

[21] C'est cette dimension qui a toujours fasciné les poètes et les écrivains qui voient dans le dictionnaire un espace de voyage où l'imprévu et la découverte le disputent aux aléas de la navigation entre les articles. Avec la navigation hypertextuelle, la tridimensionnalité n'est que plus prégnante.

[22] Ce qui donne du dictionnaire une représentation qui n'est pas plate (réduite à deux dimensions : l'horizontalité et la verticalité) mais sphérique, dont les limites demeurent ouvertes (Voir en Annexes, Figures 18 et 19).

[23] Certains dictionnaires, comme ceux qui s'adressent à des jeunes, ont recours à des jeux de couleurs ou autres formes de présentation (Voir par exemple *Le Robert illustré* 2018).

Avec l'avènement des formats numériques, le dictionnaire a changé de forme et s'est affranchi de plusieurs contraintes imposées par les versions papiers : la nature du support numérique présente un certain nombre d'avantages qui vont avoir par la suite un grand impact sur la taille du dictionnaire, notamment celle de la nomenclature, la diversification des bases de données à partir desquelles l'élaboration des dictionnaires s'effectue, notamment pour ce qui est des illustrations, des citations, des attestations, des données relatives à la datation, l'étude de la néologie qui connaît un bouleversement méthodologique, puisqu'il est actuellement possible de suivre les nouvelles créations lexicales dans leur évolution avec une très grande précision, faisant ainsi de l'étude des changements linguistiques, des évolutions du système et des modifications qui s'y fixent durablement, un champ de recherche complexe où il serait possible de montrer comment la diachronie se construit par couches successives de synchronies. On peut multiplier les exemples des divers avantages que l'informatisation apporte dans ce domaine. Contentons-nous de rappeler les nouveaux modes de consultation des versions informatisées des dictionnaires tels que *Le Grand Robert*, celui de l'*Académie française*, et évidemment le *Trésor de la langue française informatisé* : en plus de la recherche classique, qui se fait par entrée, d'autres possibilités sont offertes à l'utilisateur. *Le Grand Robert* comporte plusieurs fenêtres en rapport soit avec la nomenclature complète soit avec l'entrée recherchée. Dans le premier cas, on a accès à toutes les entrées, toutes les formes et toutes les locutions en rapport avec chaque entrée. La fenêtre réservée à l'entrée comporte toute une série de possibilités de consultations comme la forme abrégée de l'article, les synonymes et contraires, les citations (en mode simple et en mode étendu), les exemples et expressions, les homonymes, l'étymologie, etc. La recherche par article donne entre autres accès au texte intégral. Cette consultation hypertextuelle, qui se fait par le biais d'une navigation à l'intérieur du dictionnaire, donne accès à l'ensemble des occurrences qui existent dans le dictionnaire. En effectuant une recherche par critères sur le mot *femme* et en sélectionnant le champ des citations, on obtient les 4 628 occurrences avec les citations qui les renferment. Une telle donnée est riche de significations, parce qu'elle donne accès à des emplois disséminés dans un très grand nombre d'articles qui porteraient d'éventuelles nouvelles informations sur cet item ne figurant pas nécessairement dans l'article *femme*.[24] Une telle recherche peut être complétée par d'autres types de possibilités sur le même item, par exemple la recherche des locutions dont ce mot fait partie, les dérivés et composés, les exemples et expressions, etc. Pour le linguiste, le dictionnaire se transforme ainsi en un objet de recherche, notamment dans la perspective de son automatisation.

Robert Martin a consacré un ouvrage en 2001[25] à l'automatisation de la dimension sémantique du dictionnaire en vue d'élaborer un moteur d'inférences. Il a montré comment une partie au moins des données sémantiques d'un dictionnaire comme le *TLF* sont automatisables. Nous pensons qu'une telle exploitation ne concerne pas uniquement la dimension sémantique, qui est à notre avis la plus difficile, mais qu'elle la dépasse pour couvrir toutes les autres dimensions.

Rappelons que le caractère informatisé est une condition nécessaire à toute automatisation, qui est une opération par laquelle on peut accéder directement, sans intervention humaine, c'est-à-dire grâce à des algorithmes, au contenu du dictionnaire en fonction des besoins de l'utilisateur. Le dictionnaire électronique idéal serait un dictionnaire qui, une fois projeté sur des corpus, se charge de l'explication de ces textes, de leur modification (comme pour les résumés automatiques) et, dans tous les cas, de l'extraction des données recherchées. On n'en est pas encore là. Notre ambition actuelle est d'isoler ce qui est automatisable dans les dictionnaires disponibles pour en extraire une information pertinente favorisant une meilleure connaissance du système linguistique décrit (voir § 3 pour la génération automatique des réseaux inférentiels et phraséologiques).

Partant du constat que le format numérique des dictionnaires présente des avantages incontestables pour la recherche des données dans la totalité de ces ouvrages, deux possibilités s'offrent à nous :

- Se limiter à la structuration des données mises à la disposition des lecteurs et bénéficier de tous les avantages de ce type de format, qui ne sont pas négligeables ;
- Récupérer toutes les données et les traiter automatiquement pour en extraire les informations recherchées, indépendamment des contraintes imposées par les modes de consultation des formats accessibles au public[26].

De toute évidence, c'est la seconde option qui a un vrai intérêt pour la recherche linguistique. Elle présuppose au moins deux conditions :

- La première concerne une hypothèse de recherche à vérifier ;
- La seconde, la possibilité technique de l'élaboration d'un algorithme prenant en charge toute la chaîne de traitement des données.

Nous développons ces deux points.

2.1. Hypothèses

[24] Toutes ces informations échappent évidemment à un utilisateur lambda ou même aux élaborateurs du dictionnaire. Elles peuvent se révéler précieuses pour un dictionnaire électronique dont la qualité dépend de la richesse de sa description des unités lexicales. Rappelons qu'un dictionnaire électronique est un dictionnaire-machine.

[25] Robert Martin (2001).

[26] Voir la démonstration dans le § 3.

Nous partons de l'idée, déjà mentionnée, que le dictionnaire est l'unique simulation disponible du système linguistique décrit[27]. À ce titre, il comporte toutes sortes d'informations globales qui donneraient la configuration générale du système, tendances qu'on ne peut dégager qu'au moyen de données systématiques, données que seul ce genre d'ouvrage peut fournir. En plus de la taille et de l'exhaustivité des données que les formats numériques peuvent mettre à la disposition des linguistes[28], il y a lieu d'émettre l'hypothèse que les données dictionnairiques renferment les ingrédients essentiels sur la structuration sémantique et phraséologique de la langue, notamment les réseaux inférentiels et phraséologiques (voir § 3 et 4.).

S'agissant des réseaux sémantiques, nous partons de l'idée que le dictionnaire comporte, à travers l'ensemble des définitions qu'il fournit de chaque entrée, une structuration générale sémantique reflétant les différentes imbrications de sens entre les définitions. La pertinence de cette approche repose sur la conception suivante du sens des unités lexicales, en tant qu'unités de la troisième articulation[29] : contrairement aux phonèmes qui ont une pertinence phonologique, et aux morphèmes associés à une pertinence sémantique, l'unité lexicale se caractérise par une pertinence sémiotique et syntaxique[30]. La pertinence sémiotique fait que la fonction dénominative est assurée par les unités lexicales, qu'elles soient mono- ou polylexicales. Comment cette fonction sémiotique s'effectue-t-elle concrètement ? C'est dans la recherche néologique[31] qu'on pourrait trouver des réponses. Pour qu'il y ait une nouvelle dénomination, il faut qu'on dispose d'un ensemble de prédicats hiérarchisés donnant lieu à un besoin sémiotique qui favorise la création d'un signe dédié à ce faisceau de prédicats. L'exemple suivant, emprunté à la *Banque des mots*[32], permet de l'illustrer :

- X constitue un ensemble d'individus (prédicat 1)
- X constitue une même espèce (prédicat 2)
- X est un ensemble d'individus se reproduisant entre eux (prédicat 3)
- X constitue une population génétiquement isolée (prédicat 4)
- X est isolé le plus souvent pour des raisons géographiques ou écologiques (prédicat 5)

Tous ces prédicats sont encapsulés dans une catégorie grammaticale (substantif ou nom) ayant pour signifiant *dème* (n.m.). Présenté sous forme de définition, cela donne celle de la *Banque des mots* :

> *Dème* n. m. Ensemble des individus d'une même espèce se reproduisant entre eux (gamodème) et constituant une population génétiquement isolée, le plus souvent en raison de barrières géographiques ou écologiques.

La définition d'une unité lexicale n'est donc rien d'autre qu'un ensemble de prédicats encapsulés dans une catégorie grammaticale qui en détermine le fonctionnement combinatoire. Si les faisceaux prédicatifs régissent les contraintes sémantiques, l'encapsuleur grammatical prend en charge la partie combinatoire des énoncés dans lesquels s'insère l'unité lexicale. En d'autres termes, l'encapsuleur véhicule l'ensemble des virtualités combinatoires de l'unité en question. Cette vision, que nous avons eu déjà l'occasion de détailler[33], fait du dictionnaire le lieu idéal pour décrire le système de la langue. Ainsi, puisque tout énoncé n'est en fin de compte que le cadre dans lequel les unités linguistiques se déploient avec leurs faisceaux prédicatifs et leurs virtualités combinatoires, le dictionnaire, en partant de la description de ces unités, fournit en même temps tous les ingrédients nécessaires à la formation de ces énoncés, lesquels énoncés sont souvent représentés par des exemples ou des citations.

Avec cette vision de l'unité lexicale, on pourrait partir des entrées et de leurs définitions, formées d'unités lexicales également, pour dégager les différents liens sémantiques entre toutes les unités du lexique, le lexique étant par définition de nature ouverte. C'est cette ouverture qui lui assure une autorégulation : alimentée par le générateur sémiotique à l'œuvre dans la création symbolique (grâce à la fonction dénominative), le lexique s'enrichit de nouvelles unités lexicales qui en modifient perpétuellement la configuration.

Tenant compte de toutes ces considérations, notre hypothèse de travail consiste à pouvoir générer automatiquement les réseaux sémantiques que les unités lexicales entretiennent entre elles (voir § 4.1).

[27] Il s'agit évidemment des dictionnaires monolingues.

[28] Notamment pour ce qui concerne l'étymologie, la syntaxe, la morphologie et les dimensions pragmatiques.

[29] Voir Salah Mejri (2018).

[30] D'aucuns se demanderaient pourquoi les morphèmes ne seraient pas dotés d'une pertinence syntaxique et sémiotique. Nous rappelons très rapidement ce que nous entendons par morphème, unité lexicale et pertinence syntaxique et sémiotique. Le morphème, qu'il soit autonome ou pas, souffre d'une incomplétude qui le prive d'une forme du sens, c'est-à-dire une forme qui encapsule son contenu sémantique en vue de permettre son agencement avec les autres unités dans le cadre de l'énoncé. Cette forme est la partie du discours dans laquelle les morphèmes, indépendamment de leur configuration (autonome ou pas, monolexicale ou polylexicale), sont versés (encapsulés). On passe alors de l'unité de morphème à celle de l'unité lexicale, unité syntaxiquement combinable avec les autres unités de la même nature. Des morphèmes comme *–ons, il* et *sac* représentent des configurations différentes parmi celles que peuvent avoir ces unités : dans le premier cas, il s'agit d'un morphème grammatical, non autonome, dépourvu d'une partie du discours ; dans le deuxième cas, il s'agit également d'un morphème grammatical, pourvu d'un encapsuleur le rangeant dans la catégorie des pronoms, dont découlent toutes les règles de son emploi ; dans le dernier cas, le morphème *sac* est de nature lexicale. Bien que dépourvu d'une marque morphologique nominale (morphème zéro), il est rangé dans la classe des noms. Limiter l'analyse aux morphèmes conduirait à faire abstraction de la dimension grammaticale qu'apporte la troisième articulation du langage.
Pour ce qui est de la pertinence sémiotique, il faut rappeler que c'est par le biais des unités lexicales qu'on dénomme les objets du monde, non par les morphèmes : pour qu'une relation durable entre le signe linguistique et l'objet dénommé s'inscrive dans la langue, il faut une unité lexicale.

[31] Voir Salah Mejri (2018) ; voir également Louis Guilbert (1971).

[32] *La Banque des mots*, n°. 89, 2015, CILF, p. 15.

[33] Voir Salah Mejri (2018).

Pour ce qui est de la phraséologie, il est rare qu'on parle de réseaux phraséologiques, même si les travaux sur la dimension phraséologique sont des plus abondants. Pour en démontrer l'existence, nous partons de l'hypothèse émise depuis 1997 (Voir également 2003 et 2006) selon laquelle « la polylexicalité serait aux séquences figées ce que la polysémie est aux unités monolexicales »[34] (2003, p. 9). Certes, on s'est beaucoup occupé de la polysémie ; de la polylexicalité beaucoup moins, notamment en tant que pendant systémique du lexique. L'idée est que toute unité monolexicale (tout mot) participe à l'enrichissement de la langue par la multiplication des significations (grâce à la polysémie) au gré des multiples environnements de l'unité qui se fixent durablement dans son emploi : chaque signification sélectionne dans son environnement des paradigmes adéquats, en congruence sémantique avec elle. L'exemple de *bureau1* et celui de *bureau2* illustre bien la différence polysémique :

bureau1 : Paul a acheté un *bureau* tout neuf.
bureau2 : J'ai fait trois heures de *bureau*.

Une fois la congruence sémantique exigée par ces différentes significations complètement installée dans l'usage, une attraction lexicale s'effectue entre unité monolexicale et d'autres unités avec lesquelles une fixité de plus en plus importante régit les rapports jusqu'à en faire un groupe de mots solidaires. C'est ainsi que naît la phraséologie au sein du discours[35]. Si l'on reprend l'exemple de *bureau*, tout comme pour sa polysémie, on va retrouver des séries phraséologiques rattachées à ses différentes significations :

- *Articles, fournitures de bureau*
- *Bureau d'assurance, de tabac*
- *Jouer à bureaux ouverts/fermés*
- *Bureau de l'Assemblée, du Sénat, du syndic*
- *Bureau de vote, de dépouillement,*
- etc.

Tout comme pour l'existence de réseaux sémantiques, nous postulons celle des réseaux phraséologiques. Si le premier type participe d'une économie d'emploi fondée sur la variation des environnements de l'unité monolexicale, le second, qui s'inscrit en réalité dans cette variation, finit par produire des groupements syntagmatiques solidaires qui fonctionnent soit comme des unités polylexicales ayant leurs propres significations, soit comme des manières appropriées de concaténer les items lexicaux ; ce qui est une façon de prendre en charge le caractère idiomatique des associations syntagmatiques. Un verbe comme *asséner* est décrit par le *TLFi* comme suit :

[le sujet du verbe désigne une personne ; l'objet premier désigne un complément : une volée de coups, l'objet second une personne ou une partie du corps].

Puis il donne les collocations courantes dans lesquelles le verbe *asséner* signifie « diriger avec violence et de manière à frapper juste (un coup) vers quelqu'un, dans l'intention de le mettre à mal » : *asséner un coup (de poing, de bâton) sur la nuque, dans la figure ; asséner une raclée à quelqu'un*. Dans une autre signification, propre au domaine de la pêche, on a la collocation, plus figée que les précédentes, *asséner les filets*, c'est-à-dire « les développer sur le rivage pour accueillir la pêche ».

Partant de ces considérations, notre hypothèse est que le dictionnaire se structure, parallèlement aux réseaux sémantiques, selon des réseaux autres, les réseaux phraséologiques par exemple. L'entreprise est donc de générer automatiquement ces réseaux (voir § 3.3).

Pour ce faire, nous avons choisi comme corpus la neuvième édition du *Dictionnaire de l'Académie* disponible sur la toile, qui n'est pas encore achevée[36].

3. Chaîne de traitement

Le traitement textométrique visant à révéler les caractéristiques sémiotiques qui échappent aux analyses structurelles des productions langagières, ne consiste pas uniquement dans la représentation des données sous une forme mathématique, il y a notamment une technique d'automatisation susceptible de traiter une quantité importante de donnés textuelles, dont la répétition et la régularité sont les principales caractéristiques. Le dictionnaire étant en effet un type de données qui est par définition classé (par ordre alphabétique), catégorisé (par catégorie grammaticale, classe sémantique, etc.) et organisé, son traitement tend à tirer profit des séparations inhérentes et des classifications du lexique dans le dictionnaire. Des travaux textométriques du dictionnaire (Dodge 1993, Lebrat et Salem 1994, Béjoint 2007) présentent plusieurs méthodes de classification parmi lesquelles se trouve la classification des formes et des textes par filtrage (Lebrat et Salem, 1994) dont la logique consiste à classer les données textuelles selon leurs premiers facteurs significatifs. Dans le cadre de notre traitement, trois types de facteurs sont déterminants : le premier est celui des symboles visibles tels que la ponctuation, les marques alphanumériques, les formes algébriques, etc., ainsi que les symboles typographiques qui sont négligés ou invisibles pour des humains, comme les espaces (sécables et insécables), les sauts de ligne, etc. ; le deuxième est celui des typographies du dictionnaire qui traduisent la conception de l'organisation chez les élaborateurs (par exemple, les entrées sont en majuscule) et la hiérarchisation

[34] *Syntaxe et sémantique*, n° 5, Polysémie et polylexicalité, Caen, 2003.
[35] D'autres sources de phraséologismes existent évidemment : la dénomination polylexicale est la plus importante.
[36] Les données extraites sont strictement destinées à des fins de recherche et ne comportent aucune indication renvoyant à des personnes.

des contenus et des concepts ; le troisième est celui de la structuration globale du dictionnaire et de ses entrées (voir § 1).

Concernant les technologies utilisées, nous avons recours au langage de programmation Python pour les opérations d'automatisation et au logiciel d'analyse et de visualisation des réseaux GEPHI pour mettre en évidence les résultats. Des scripts ont été créés pour automatiser les traitements d'analyse du dictionnaire en vue de transformer les données dictionnairiques en un texte brut, en des réseaux sémantiques et phraséologiques. Dans cette optique, la statistique, qui n'est pas la finalité de notre travail, est un moyen pour faire émerger toutes les facettes formelles du dictionnaire : elle nous aide à mieux appréhender le dictionnaire.

La Figure 1 retrace les différentes étapes du traitement complètement automatisé, allant de la collecte des données jusqu'à l'évaluation des résultats obtenus :

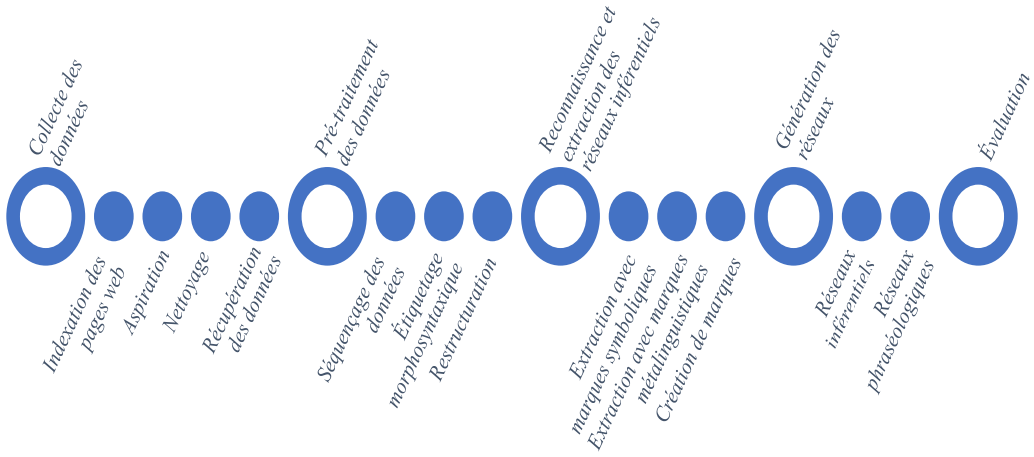

Figure 1- chaîne de traitement

3.1. La collecte des données

On a procédé à l'indexation du site https://www.dictionnaire-academie.fr (9e édition, de A à R). Les adresses URL sont générées automatiquement par un script informatique. Dès qu'une adresse URL générée est valide et accessible, le programme extrait le code source de la page web de l'URL et n'en garde que le texte brut. Parmi les difficultés rencontrées, nous retenons les trois suivantes :

- la numérotation des articles du dictionnaire étant souvent discontinue, cela provoque des interruptions dans l'extraction, qui peuvent être corrigées mais elles sont coûteuses ;
- une certaine hétérogénéité dans les données : le texte extrait du dictionnaire comporte d'autres types de données ;
- le décodage de certains symboles comme les accents, ‹œ› et ‹æ›.

Pour la récupération des données, comme les entrées du dictionnaire sont présentées sous forme de page web, nous avons créé un robot d'indexation pour en extraire le code source, et ajouté un script qui nettoie les balises pour ne préserver que les données textuelles en format brut.

3.2. Le prétraitement

Les données récupérées ne sont *a priori* segmentées que par retours chariots dans une structure originelle. Nous conservons cette segmentation en séparant les entrées de leurs articles. L'objectif est de faire de l'entrée du dictionnaire l'indice de l'article, et de stocker le tout dans un ensemble où ils sont interdépendants. Pour chaque entrée, nous segmentons les significations (ou sens) par les retours chariots et par le recours à des symboles.

Le résultat obtenu s'élève à 29 270 entrées.

3.3. Reconnaissance et extraction des phraséologismes

Deux procédures sont mises en place : le recours à des marqueurs lexicographiques et une méthode en l'absence de ces marqueurs.

L'économie du dictionnaire repose entre autres sur tout un appareil symbolique servant de marqueurs redondants. Il est formé d'abréviations, de symboles alphanumériques et de formes géométriques (comme les losanges ◆, les traits ━, etc.). Chaque marque joue le rôle d'un classifieur permettant d'inscrire les faits portant une marque quelconque dans l'une des catégories servant de grille d'analyse. Les marques des parties du discours (N, V, Adj., Adv., etc.) sont les plus récurrentes : elles sont systématiques et accompagnent chaque entrée. Tel n'est

pas le cas d'autres marqueurs qui ne sont ni systématiques ni suffisamment récurrents pour servir de filtres fiables. Les marqueurs domaniaux, indiquant le domaine où l'on emploie l'item lexical décrit ou une signification particulière, tout en étant redondants, ne couvrent ni tous les domaines ni tous les emplois. Cela ne signifie pas pour autant que leur emploi ne soit pas pertinent. Il y a d'autres marqueurs, qu'on pourrait appeler « métalexicaux », qui prennent en charge des catégories qui rendent compte du comportement des items décrits couvrant la syntaxe, la morphologie, la sémantique, la pragmatique, etc. En excluant les marqueurs de la nature grammaticale et ceux qui relèvent des domaines, au nombre de 96, et après modélisation de ces marqueurs, nous obtenons 49 marqueurs métalexicaux, qui totalisent 55 000 emplois dans le dictionnaire. Nous en fournissons un échantillon dans ce qui suit :

> *Fig., Par extension, Spécialement, Par métonymie, Expr., Par analogie, Fam., Loc., Vieilli, Anciennement, Absolument, Adjectivement, En apposition, Pop., Litt., Syn., Prov., Péj., Rare, Class., Elliptiquement, Très vieilli, Iron., Par exagération, Argot, En composition, Adverbialement, Par euphémisme, Régional, Par litote, Par antiphrase, Souvent péj., Vulgaire, Surtout fig., Argot militaire, Surtout pron., Parfois péj., Argot scolaire, Trivial, Didact., Par hyperbole, Souvent iron., En incise, Littérature, Abusivement, Par abréviation, Parfois iron., Excessivement, Surtout péj.*

La comparaison de la liste des abréviations avec la distribution effective dans le dictionnaire montre que certains marqueurs ne sont pas répertoriés.

Les deux figures suivantes (Figures 2 et 3) fournissent une idée sur leur couverture lexicale. Le croisement des deux types de formes (première lettre du marqueur en majuscule ou non) donne une idée précise sur la couverture globale et permet ainsi des recherches relativement précises sur chacune des catégories retenues[37] :

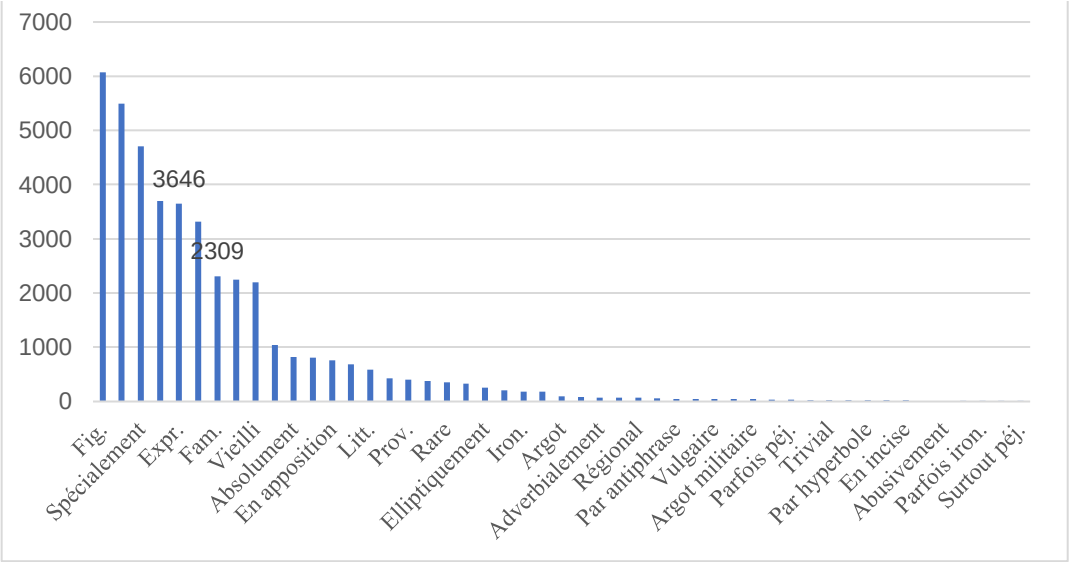

Figure 2

[37] Le résultat des requêtes sur les marqueurs domaniaux fournit des éléments intéressants, à développer dans un travail séparé.

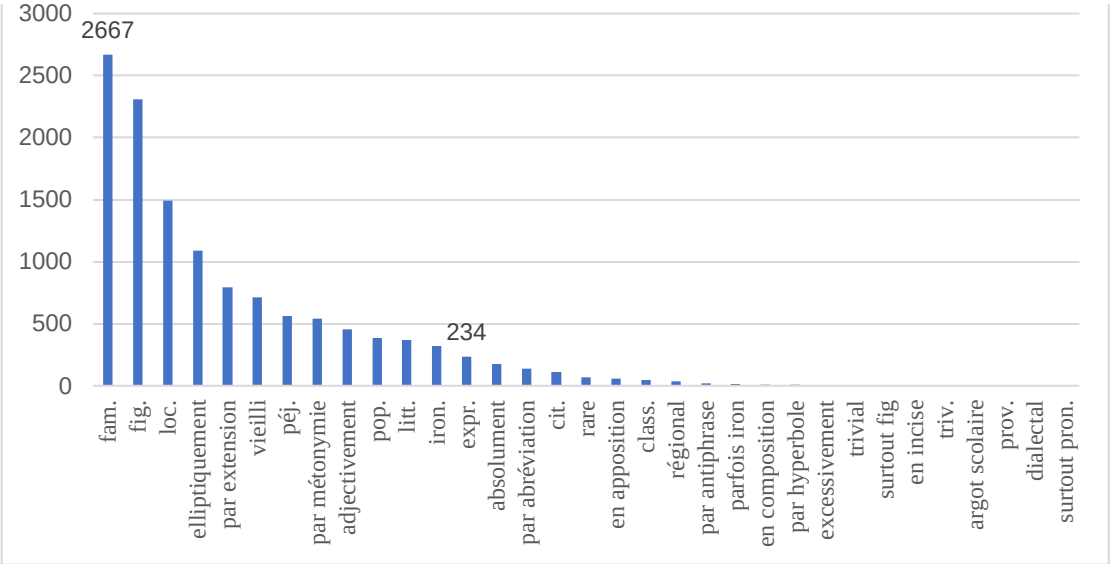

Figure 3

Nous notons que les croisements des deux formes des marqueurs *Fam./fam.* d'un côté, et ceux de *Expr./expr.* de l'autre, totalisent respectivement 4 976 (2 309 + 2 667) et 3 880 (3 646 + 234).

Plusieurs problèmes se sont posés lors de l'emploi de ces marqueurs, parce que leur présentation n'est pas normée : ils sont employés parfois seuls, parfois concaténés en deux ou trois marqueurs, juxtaposés ou coordonnés (exemples : *loc. pop.*, *Expr. Pop.*, *pop. et vieilli. et souvent péj*.). Cette dernière forme n'a pas été retenue dans ce travail. Pour les autres, il a fallu ajouter un ensemble de conditions formelles pour pouvoir les rendre fonctionnelles.

Pour détecter et extraire les phraséologismes moyennant les marqueurs méta-lexicaux et domaniaux, nous avons obtenu 20 768 phraséologismes, dont 14 591 sont repérés grâce aux premiers, 5 825 grâce aux seconds. Les figures 4 et 5 en font la synthèse :

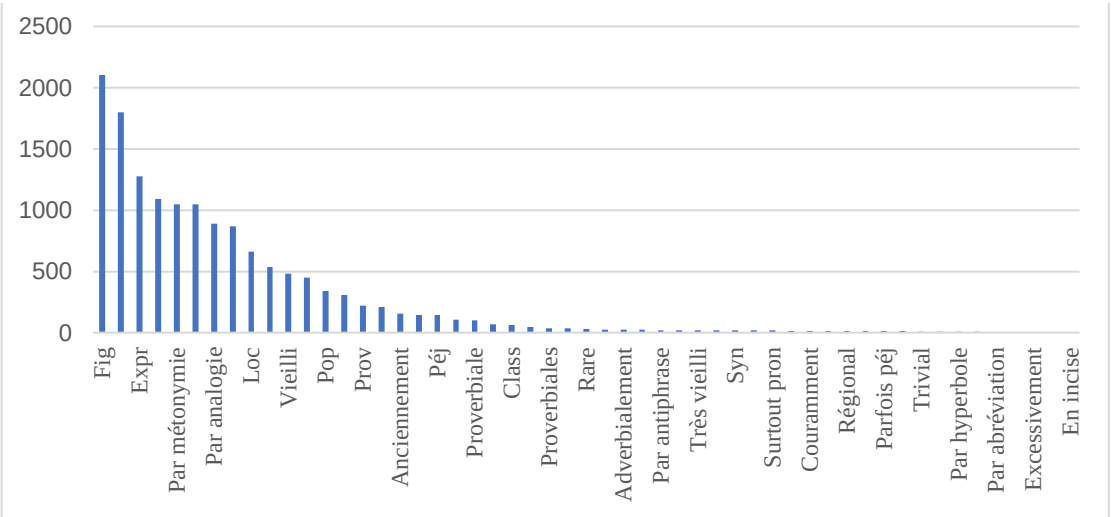

Figure 4

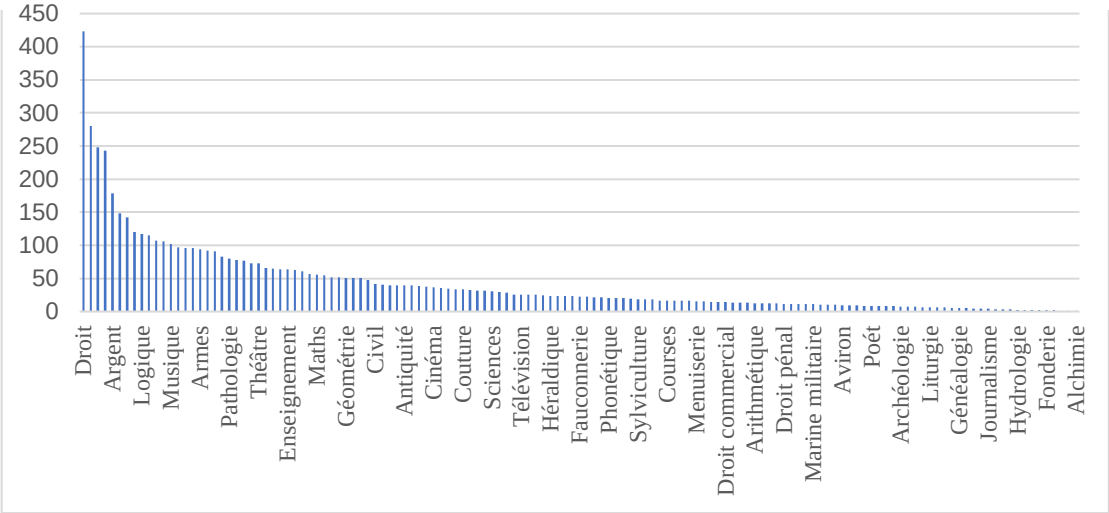

Figure 5

Nous fournissons l'extrait suivant des résultats obtenus :

```
12173   INTENSITÉ       Phonétique      phonétique.Accent d'intensité, qui renforce le son d'une syllabe par
rapport aux syllabes voisines. En français, l'accent d'intensité porte le plus souvent sur la dernière syllabe du
mot.
12174   INTENSITÉ       Physiologie     physiologie.Intensité liminale, seuil que doit atteindre un stimulus pour
entraîner la réponse d'une cellule sensorielle, d'un neurone. Intensité respiratoire, volume d'oxygène qu'absorbe
un être vivant par unité de masse et par unité de temps.
12175   INTENTION       Droit   droit.Intention frauduleuse. Intention de nuire.
12176   INTENTION       Expr    Expr.Être empli de bonnes intentions, vouloir bien faire. Faire à quelqu'un un
procès d'intention, lui prêter, généralement à tort, un dessein dont il n'y a aucune preuve qu'il l'ait formé.
12177   INTENTION       Proverbiales    proverbiales.C'est l'intention qui compte.L'enfer est pavé de bonnes
intentions,voir Enfer.
12178   INTERARMES      Armes   armes. L'École militaire interarmes.
12179   INTERCALAIRE    Astronomie      astronomie.Lune intercalaire, la treizième lune qui se trouve dans une
année, de trois ans en trois ans.
12180   INTERCALAIRE    Prosodie        prosodie.Vers intercalaire, vers qu'on répète plusieurs fois dans certains
poèmes, tels que les ballades, les rondeaux, etc.
12181   INTERCEPTER     Militaire       militaire.Intercepter un avion, un convoi militaire. Intercepter un
missile, l'empêcher d'atteindre son objectif.
12182   INTERCEPTER     Sports  sports.Intercepter le ballon, la balle.
12183   INTERCEPTEUR    Adjectivement   Adjectivement.Un avion intercepteur.
12184   INTERCEPTION    Sports  sports.L'interception d'un ballon lors d'une passe.
12185   INTERDICTION    Artillerie      artillerie.Tir d'interdiction, destiné à empêcher l'ennemi d'atteindre ou
d'occuper un objectif, une zone.
12186   INTERDICTION    Droit canon     droit canon.Interdiction d'un prêtre,voir Interdit.
12187   INTERDIT        Fig     Fig.Jeter l'interdit sur quelqu'un, sur quelque chose, exclure une personne d'un
groupe, d'une société, prohiber l'usage de quelque chose.
12188   INTÉRESSER      Par métonymie   Par métonymie.Intéresser le jeu, jouer de l'argent pour rendre le jeu plus
attrayant par l'appât du gain.
```

Figure 6

Pour obtenir la moyenne des phraséologismes, nous établissons le rapport suivant entre le nombre d'occurrences obtenues et celui des entrées du dictionnaire (20 768/29 270). Le résultat est 0,71 par entrée. Ce qui justifie la recherche des phraséologismes sans marqueurs.

Les phraséologismes non marqués appartiennent à trois catégories : des formes collocationnelles, des occurrences contenues dans des exemples et des séquences polylexicales suivies de définitions propres. Pour les collocations se dégagent deux cas de figure :

– la base, qui ne correspond pas à l'entrée de l'article, suivie du collocatif, qui représente l'entrée : *blanchiment, séchage du* ***papier****. Laver, laminer, calandrer le* ***papier****…* cela prend la forme d'une énumération ;
– la base (= l'entrée) suivie du collocatif ; la combinaison des deux se décline comme suit :

- base + collocatif énuméré, séparés par une virgule, la suite de l'énumération se terminant par un point : ***Papier*** *grand aigle, petit aigle, carré, cavalier…*
- base répétée + collocatifs énumérés séparés par une virgule, le tout se terminant par un point : ***Papier*** *paille,* ***papier*** *maïs.*
- base + collocatif sans énumération, les combinaisons étant séparées par un point : ***Papier*** *crépon.* ***Papier*** *pelure.* ***Papier*** *de soie.*

Le deuxième type concerne les phraséologismes contenus dans des exemples. Ils présentent la caractéristique d'être intégrés dans des contextes :

- *L'**abaissement** du mur nous permettra de jouir du paysage.*
- *On prévoit un **abaissement** sensible de la température au cours de la semaine.*

Quant au dernier type, il répond à la configuration suivante : phraséologisme + définition, les deux étant séparés par une virgule.

Exemple : **La boîte de pandore**, par allusion à un récit de la mythologie grecque, ce qui est source de grands malheurs.

La technique du repérage et de l'extraction consiste à tenir compte de la présence de l'entrée dans le phraséologisme, non dans l'explication.

Le résultat s'élève à 99 977 occurrences phraséologiques répertoriées comme suit :

- 1er type : 11 877
- 2e type : 58 753
- 3e type : 49 347

Ce qui distingue les deux premiers types et le troisième, c'est le critère sémantique : les deux premiers sont jugés transparents ; le dernier nécessite, par son opacité sémantique, une définition.

Les procédés d'extraction sont les suivants :

. Procédé 1 : repérer une entrée (ENTREE 1),
. Procédé 2 : extraire les phraséologismes et les définitions de l'entrée,
. Procédé 3 : isoler, extraire et lemmatiser les items du phraséologisme et ceux de la définition. (TreeTagger Python Wrapper, module Python : https://treetaggerwrapper.readthedocs.io/en/latest/)

Exemple : **ANCRE**

```
La          DET:ART le
maîtresse   NOM      maître|maîtresse
ancre       VER:pres ancrer[38]
.           SENT     .
Ancre       NOM      ancre
à           PRP      à
jas         NOM      jas
,           PUN      ,
à           PRP      à
bras        NOM      bras
escamotable ADJ      escamotable
.           SENT     .
Ancre-charrue NOM    <unknown>
.           SENT     .
Jeter       VER:infi jeter
l'          ADJ      <unknown>
ancre       NOM      ancre
```

Figure 7

Pour les phraséologismes transparents, chaque cas se décompose en mots formes séparés par des espaces :

[38] Le lemmatiseur a mal reconnu cet item : *ancre* est un nom ici. Les erreurs de lemmatisation constituent une grande partie des erreurs dans les résultats (voir § 5).

```
8343 - FEUILLE   Feuille d'épreuve.
8344 - FEUILLE   Une feuille locale.
8345 - FEUILLE   Une feuille de schiste.
8346 - FEUILLÉ   Le feuillé d'un paysage.
8347 - FEUILLER  Feuiller un linteau.
8348 - FEUILLETER        Feuilleter un dictionnaire.
8349 - FEUILLETON        Feuilleton hebdomadaire.
8350 - FEUILLETON        Lire le feuilleton du jour.
8351 - FEUILLETTE        Une feuillette de chêne.
8352 - FEUTRE    Feutre de laine.
8353 - FEUTRE    Un feutre mou.
8354 - FÈVE      Fève d'Égypte.
8355 - FÈVE      Fève de cacao.
8356 - FIASQUE   Une fiasque de chianti.
8357 - FIBRE     Fibre d'amiante.
8358 - FIBROMYOME        Fibromyome utérin.
8359 - FIBROSE   Fibrose pulmonaire.
8360 - FICELER   Ficeler un colis.
8361 - FICELER   Ficeler un rôti.
8362 - FICHE     Fiche à gond.
8363 - FICHE     Fiche de prise de courant.
8364 - FICHE     Fiche multiple.
8365 - FICHE     fiche comptable.
8366 - FICHE     Fiche médicale.
8367 - FICHER    Ficher un pieu en terre.
8368 - FICHIER   Un fichier alphabétique.
8369 - FICHIER   le fichier par matières.
8370 - FICHIER   Un fichier métallique.
8371 - FICHTREMENT       Il fait fichtrement froid.
8372 - FICHU     Un fichu brodé.
```

Figure 8

S'agissant des phraséologismes opaques, il est tenu compte de la chaîne de caractères et de l'ensemble des indications impliquées dans la hiérarchisation des significations dans l'article.

4. Génération

Deux types de relations seront générés : relations inférentielles et relations phraséologiques. Il s'agit de reconstituer, à partir des données automatisées, les réseaux inférentiels et les réseaux phraséologiques. Le programme consiste à traiter l'entrée « ARBRE ».

4.1. Génération des réseaux inférentiels

Il s'agit de considérer d'abord différentes définitions d'une entrée pour se servir des items qui les composent comme point de départ pour une recherche de leurs définitions respectives, lesquelles font l'objet du même traitement, et ainsi de suite.

Formalisme : la première phrase (une chaîne de caractères longue (>25 signes) qui se termine avec un point (avec ou sans virgule dans la phrase) .

Exemple : ARBRE

1. Végétal ligneux de grande taille dont la tige ne se ramifie qu'à partir d'une certaine hauteur.
2. Par analogie. Technique. Pièce maîtresse, axe d'une machine, d'un appareil.

Ces deux définitions contiennent les lemmes suivants (8 lemmes, segmentés, tokenisés et lemmatisés par Treetagger) :

VEGETAL (lettre V n'a pas été créée[39]), LIGNEUX, GRAND (il fait partie des mots considérés sans sens plein), TAILLE (lettre T n'a pas été créée), TIGE (lettre T n'a pas été créée), RAMIFIER, HAUTEUR, PIECE, MAITRE, AXE, MACHINE, APPAREIL[40] (voir Figure 9) :

[39] La 9e édition n'est pas encore achevée : elle s'arrête à la lettre R (version consultée le 24 mars 2019).
[40] Les items des définitions de la première entrée sont en majuscule.

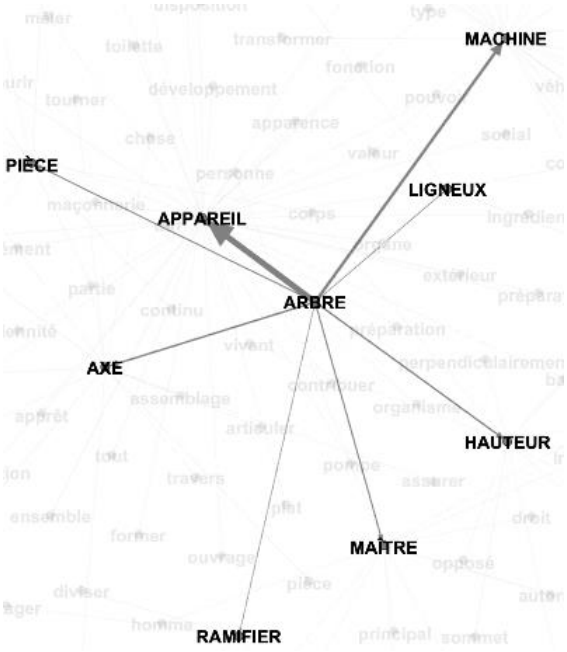

Figure 9

Ce graphe retrace les liens qui existent entre l'unité *arbre* et les items qui font partie des deux définitions. Il est obtenu automatiquement au moyen du logiciel GEPHI selon l'algorithme Fruchterman et Reingold (1991), qui tient compte de deux paramètres, la force et la distance : la force est la somme de l'attractivité et de la répulsivité des variables (en l'occurrence l'entrée *arbre* [variable1] et les items lexicaux définitoires [variables2]) ; la distance concerne l'espace d'affichage divisé par le nombre de liens (1991, p.1133). L'épaisseur des lignes traduit le degré de richesse des définitions des items pointés par les vecteurs (*appareil* est plus riche que *ligneux*). Les positions des mots dans le graphe n'ont pas de lien avec la proximité ou la distance lexicale ou sémantique entre les items. La disposition des nœuds que représentent les mots par rapport au noyau, en l'occurrence *Arbre*, est calculée selon la formule suivante : $\sqrt{\frac{zone\ d'affichage}{nombre\ de\ noeuds}}$[41], qui a pour seul but l'harmonisation de l'affichage du graphe. Afin d'illustrer les différents niveaux de relations sémantiques entre les items, nous mettons en majuscule les items définitoires de *Arbre*, les items définitoires de ces derniers étant affichés en minuscule. Dans la Figure 9, nous mettons en évidence le réseau du mot *Arbre*, constitué de ses items définitoires. Les réseaux des niveaux inférieurs, réseaux des items définitoires de *Arbre*, sont en grisé.

Afin de ne pas alourdir la présentation, nous nous limitons à la première définition de chaque item si ce dernier a plus d'une définition. Chaque item est réinjecté dans le dictionnaire pour sélectionner son entrée correspondante.

Exemple : L'adjectif LIGNEUX correspond à l'entrée LIGNEUX du dictionnaire, définie comme suit :

1. Qui a la nature ou la consistance du bois.
2. Dur comme du bois, qui a la consistance du bois.

Après la tokénisation et la lemmatisation des items de la définition de LIGNEUX, nous obtenons les items suivants : *nature, consistance, bois, dur*[42], qui forment de leur côté de nouveaux nœuds servant de points de départ pour d'autres items en fonction de leurs définitions respectives, ce qui donne la figure suivante :

[41] *Ibid.*
[42] Ces items sont en minuscule pour marquer la distance sémantique avec l'entrée initiale ARBRE.

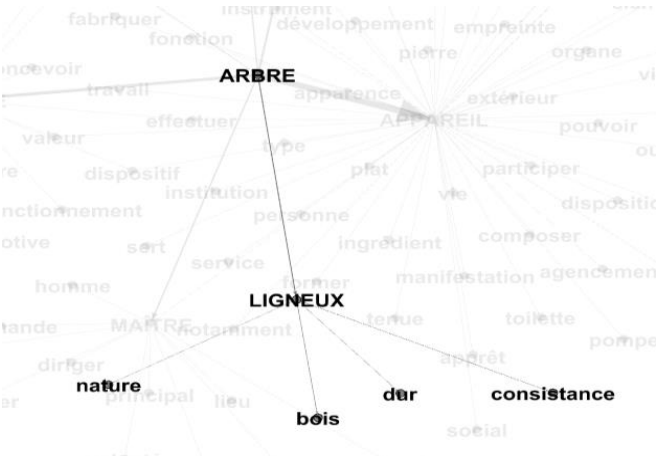

Figure 10

APPAREIL, de son côté, fournit (Figure 11) 7 définitions :

1. Apprêt préparatif à une manifestation empreinte de pompe de solennité.
2. Tout ce qui notamment dans la toilette, la tenue contribue à mettre en valeur l'apparence extérieure d'une personne.
3. Ensemble d'organes, d'éléments qui forment un tout et participent à une même fonction.
4. Ensemble de tous les organes et tissus pouvant concourir à une même fonction chez un être vivant.
5. Agencement et disposition des pierres dans un ouvrage de maçonnerie.
6. Préparation où sont mêlés plusieurs ingrédients et qui sert de base à un type de plat.
7. Ensemble des organismes qui concourent à la vie et au développement d'une institution sociale.[43]

Elles sont constituées des 45 items suivants : *apprêt, préparatif, manifestation, empreinte, pompe, solennité, toilette, tenue, contribuer, valeur, apparence, extérieur, personne, ensemble, organe, élément, former, participer, fonction, ensemble, tout, organe, tissu, pouvoir, concourir, fonction, vivant, agencement, disposition, pierre, ouvrage, maçonnerie, préparation, mêler, ingrédient, base, type, plat, ensemble, organisme, concourir, vie, développement, institution, social*

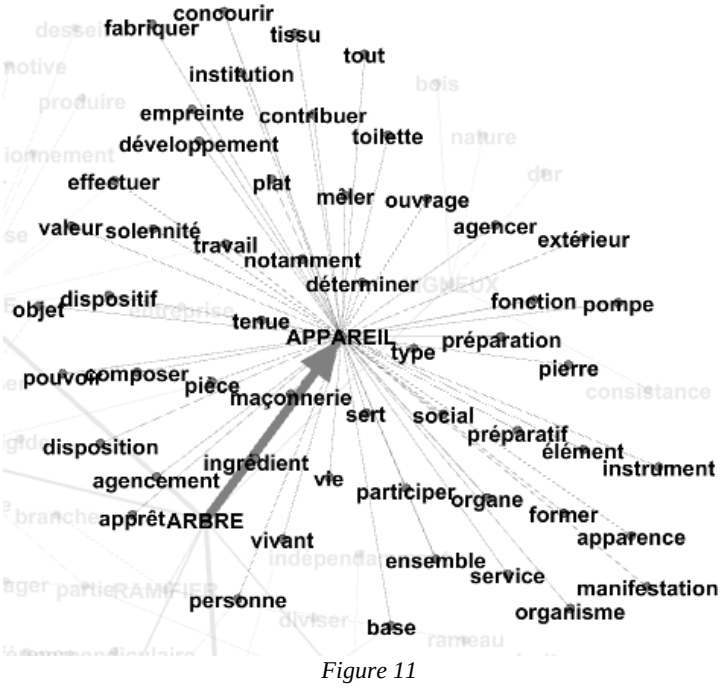

Figure 11

Au total, pour l'entrée ARBRE, on obtient 116 items qui découlent des définitions des 8 items définisseurs d'ARBRE :

[43] Les mots-outils ne sont pas pris en compte dans la lemmatisation.

nature, consistance, bois, dur, diviser, rameau, partager, branche, partie, morceau, portion, appartenir, séparer, corps, solide, continu, commande, dominer, droit, personne, diriger, volonté, homme, lieu, autorité, important, premier, principal, pièce, rigide, allonger, passe, travers, objet, tourner, assurer, assemblage, articuler, central, organiser, chose, dispositif, mécanique, complexe, produire, transformer, énergie, travail, effectuer, tâche, entreprise, œuvre, concevoir, dessein, destruction, moyen, propulsion, locomotive, véhicule, singulier, valeur, collectif, divers, système, fonctionnement, régulier, apprêt, préparatif, manifestation, empreinte, pompe, solennité, toilette, tenue, contribuer, apparence, extérieur, ensemble, organe, élément, former, participer, fonction, tout, tissu, pouvoir, concourir, vivant, agencement, disposition, pierre, ouvrage, maçonnerie, préparation, mêler, ingrédient, base, type, plat, organisme, vie, développement, institution, social, dimension, corps, base, sommet, vivant, taille, longueur, droite, abaissé, perpendiculairement, côté, opposé

4.2. Génération des réseaux phraséologiques

Nous générons des réseaux à partir des phraséologismes transparents, ceux qui ne sont pas définis par le dictionnaire.

Les phraséologismes formés à partir du constituant *arbre* s'inscrivent dans les différentes significations de cette unité :

- Arbre « végétal » : 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 à feuilles caduques ; 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 à caoutchouc ; arbre à pain ; arbre à vessies ;
- Arbre « mécanique » : arbre de moulin ; arbre d'un pressoir ; arbre d'une grue ; l'arbre à cames d'un moteur ; arbre de transmission ;
- Arbre « figuratif » : dresser un arbre généalogique.

Nous ne traitons que les phraséologismes relevant du premier sens, « végétal ». Nous utilisons les phraséologismes que nous avons récupérés dans § 3.3. Chaque phraséologisme est tokénisé en mots qui sont ensuite lemmatisés un à un. Ils sont projetés comme suit :

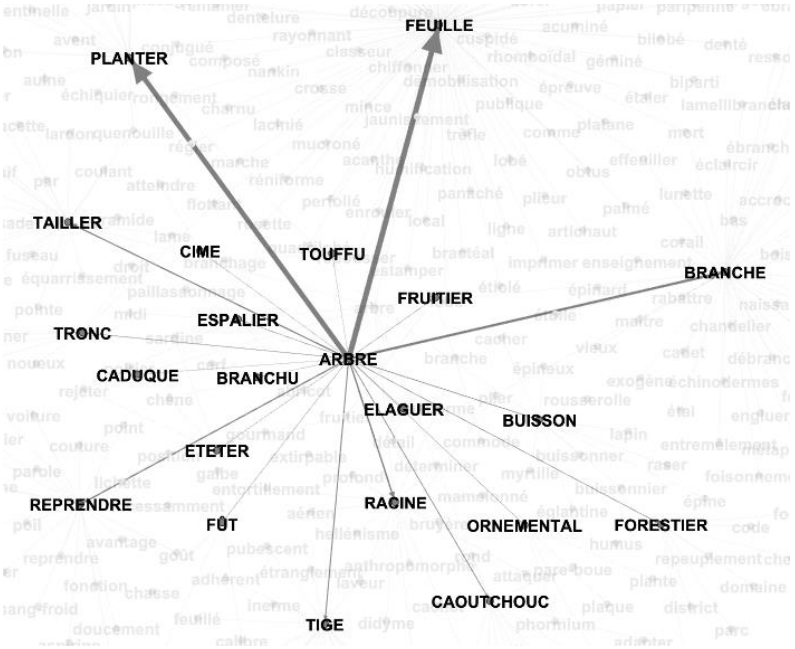

Figure 12

Prenons l'exemple de *Arbre branchu. Branchu* est également une entrée dans le dictionnaire qui dispose des phraséologismes suivants :

Un chêne branchu. *Un cerf branchu*. Ce qui donne un réseau restreint :

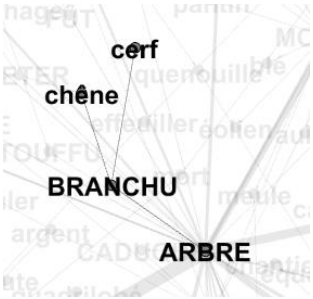

Figure 13

Le même cas de figure se présente avec CIME. Dans l'article de CIME, il n'y a qu'un phraséologisme : *atteindre la cime*.

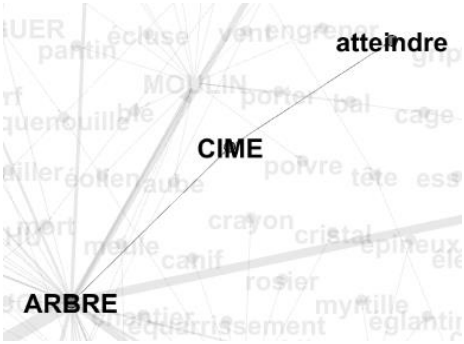

Figure 14

Nous relevons également des collocations verbales : *élaguer un arbre*. À partir de *élaguer un arbre*, nous extrayons les phraséologismes de *élaguer* : *élaguer des branches mortes*, *élaguer les détails*, *élaguer les gourmands*[44]. À partir de ces phraséologismes, on retient *branche*, *détail*, et ainsi de suite.

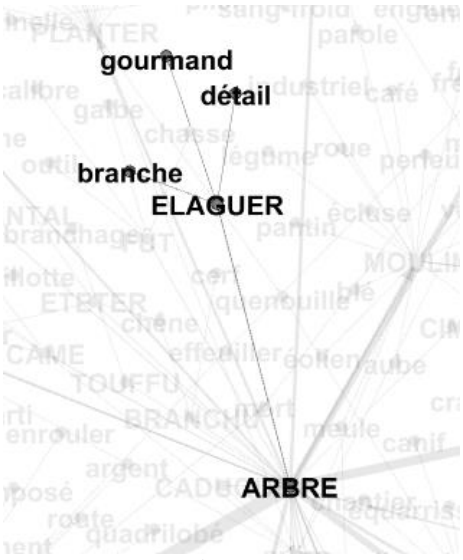

Figure 15

[44] Dans la définition de ***Gourmand*** : « Spécialement. Arboriculture. Branche gourmande, rameau d'arbre fruitier poussant au-dessous de la greffe, et ne donnant donc pas de fruits, qui se développe au détriment des autres rameaux. Subst. Couper un gourmand, des gourmands. ***Élaguer les gourmands***. – horticulture. Pois gourmands, ou pois mange-tout, qu'on mange avec la cosse. ». À noter : le phraséologisme *élaguer les gourmands* ne figure pas dans l'article *élaguer*.

Si nous partons de *feuille*, on aura par exemple :

- *les feuilles acuminées du cornouiller*
- *arbre à feuilles caduques* (arbre)
- *feuille d'argent* (argent)
- *feuille bilobée (bilobé)*
- *une feuille bipartie (biparti)*[45]
- *feuilles bractéales (bractéal)*
- *calandrer une feuille de matière plastique (calandrer)*
- *les feuilles charnues (charnu)*
- *chiffonner une feuille de papier (chiffonner)*
- *la chute des feuilles en automne (chute)*
- *feuilles ciliées (cilié)*
- *un classeur à feuilles perforées (classeur)*

Pour le traitement des phraséologismes longs, nous avons privilégié les structures « entrée + mot grammatical + X » ou « X + mot grammatical + entrée ». 111 items phraséologiques ont été obtenus :

acuminé, arbre, argent, bilobé, biparti, bractéal, calandrer, charnu, chéneau, chèvrefeuille, chiffonner, chute, cilié, classeur, cloqué, comme, composé, coricide, crosse, cuspidé, déchiqueté, découpé, découpure, démobilisation, denté, dentelé, dentelure, denticulé, dessèchement, digité, dorer, double, écrit, effeuiller, embrouiller, engainant, enrouler, envolement, estamper, étaler, étiolé, étoilé, extraire, fabrication, fabriquer, fer, conjugué, sèche, mort, trèfle, artichaut, rosette, lierre, acanthe, papier, température, marche, route, imprimer, épreuve, publique, local, schiste, flabellé, flottant, gâteau, géminé, humification, imposer, impôt, jaunissement, joseph, lacinié, ligne, ligulé, lobe, lobé, martelage, mince, modelage, mouilleur, mucroné, nankin, obtus, palmé, panaché, parchemin, paripenné, perfolié, placage, planimétrie, platane, plieur, puisard, quadrilobé, queue, raisin, ramette, rayonnant, refendre, régler, remanier, réniforme, replier, repousser, revers, rhomboïdal, rogner, rongement, rousseur, roussir.

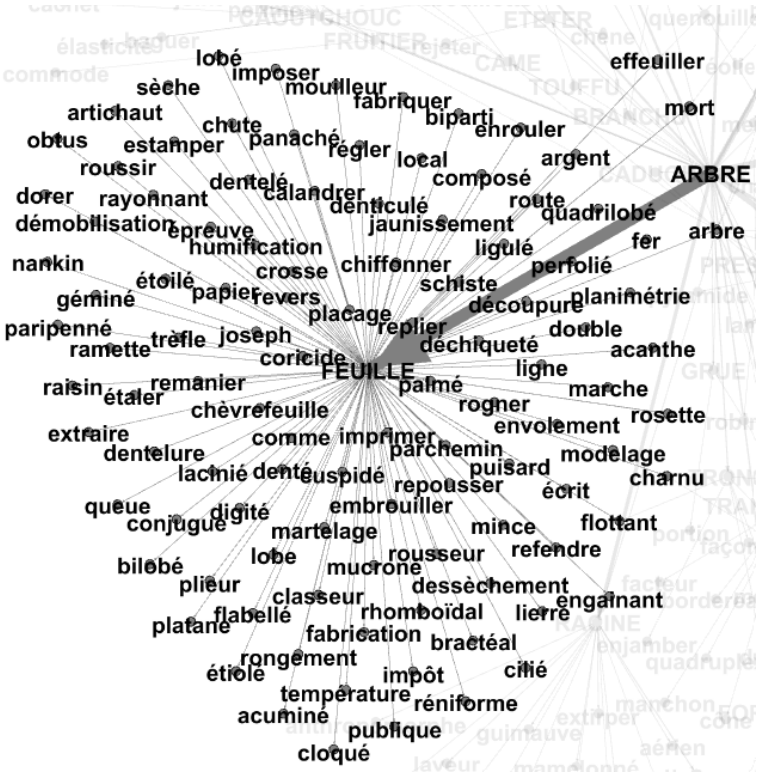

Figure 16

4.3. Génération des phraséologismes par un biais sémantique

En utilisant certains marqueurs métalinguistiques, nous pouvons également générer des phraséologismes par le biais sémantique.

Nous avons d'abord relevé des marqueurs métalinguistiques ayant les configurations suivantes : *se dit de, en parlant de, une espèce de, un type de,* ainsi que leurs variantes formelles (Figure 17). L'extraction et la reconstitution des collocations à partir de ces marqueurs se fait ensuite de la manière suivante : le programme isole l'entrée, qui est soit la base, soit le collocatif, suivie d'un marqueur métalinguistique, et il le couple avec un item pour composer

[45] Le phraséologisme est présent sous *biparti*.

la collocation. Les mots définisseurs suivis d'un marqueur métalinguistique nous permettent de générer sémantiquement des séquences collocationnelles : pour chaque mot définisseur, notre programme vérifie à nouveau s'il est présent dans chacune des définitions d'une entrée (différente de la première entrée). Si le mot définisseur figure effectivement dans une des définitions d'une entrée, notre programme couple la première entrée et la seconde entrée pour former une séquence. À titre d'exemple, l'entrée ABOUTIR (verbe intransitif), dont l'une des définitions est : « toucher par un bout. *La route aboutit à la mer. Riquet fit aboutir à Sète le canal du Midi. L'escalier aboutit au grenier. Cette rue aboutit au carrefour.* Par extension. En parlant d'une personne. Parvenir à. », contient le marqueur *en parlant d'*, qui est suivi du mot définisseur *Personne* (qui est un nom). Notre programme parcourt le dictionnaire afin de repérer les entrées dont la catégorie grammaticale est nominale, définies effectivement par le mot définisseur « personne ».

Notre programme repère l'entrée ACCRÉDITEUR qui est un nom ayant pour définisseur le mot « Personne » : « Personne qui donne sa garantie, sa caution en faveur d'un tiers. ». Il génère ensuite la séquence « Accréditeur aboutir »[46] selon la syntaxe du verbe intransitif suivante : Nom + Verbe. Ce faisant, nous révélons les combinatoires potentielles des mots du dictionnaire au moyen de leurs définitions et de leur syntaxe. Plus de 4 500 séquences du type collocationnel sont ainsi générées à partir des marqueurs métalinguistiques susmentionnés (Figure 17).

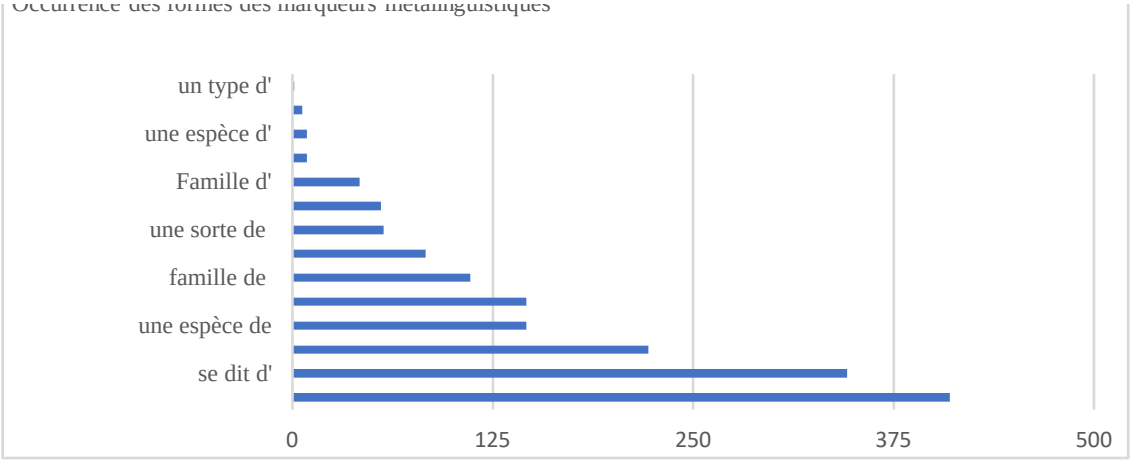

Figure 17

5. Évaluation

Concernant l'extraction des définitions, les premières acceptions sont toutes extraites (exceptée pour l'entrée *machine* : le programme a également extrait une définition de la deuxième acception).

Pour ce qui est de l'extraction des phraséologismes du mot *Arbre*, le taux de précision est de 85% (517/608), la plupart des erreurs provenant du tagger (« plante » [nom féminin désignant les végétaux] est lemmatisé comme « plant » (73 cas)). 10 cas ne sont pas figés, 3 cas d'erreur d'orthographe, 5 cas mal reconnus. Quelques difficultés subsistent :

1. Tous les tokens de définitions et de phraséologismes n'ont pas d'entrée dans le dictionnaire ;
2. Problèmes d'homonyme (*fût* : être – *fût* : tronc d'arbre, *être* : Nom – *être* : Verbe) ;
3. Problèmes d'orthographe (le programme trouve aussi *cimetière* à partir de *cime*, il trouve les phraséologismes formés de *Racine* (nom propre) au lieu de *racine* (nom commun)) ;
4. Imprécision du tagger : Treetagger considère que *plant* et *plante* ont le même lemme ;
5. Des cas qui posent problème : *ensemble, tout, objet* (qui sont des noms sommitaux).[47]

Conclusion

Il ne s'agit là que d'une approche globale qui n'a rien de définitif. Tous les éléments fournis n'ont qu'une pertinence méthodologique, dont l'objectif essentiel est de montrer que les dictionnaires informatisés comportent des données susceptibles d'être exploitées automatiquement. Cela permettrait aux chercheurs de travailler sur des données de plus en plus systématisées, de constituer des ressources de plus en plus exhaustives et de découvrir des liens qui ne sont pas accessibles directement au traitement humain, mais qui sont nécessaires à l'élaboration d'un dictionnaire entièrement automatisé à partir des dictionnaires disponibles.

Plusieurs difficultés sont à surmonter. Nous en retenons uniquement les trois suivantes :

[46] Les accords et les conjugaisons ne sont pas pris en compte dans la génération.
[47] Nous consacrerons un article à part aux retours théoriques et appliqués de cette expérimentation.

- La question théorique en rapport avec le caractère tridimensionnel des relations entre les unités lexicales ;
- La notion de nœud dans les graphes qu'il faut discuter en tant que lieu de continuité et de discontinuité sémantique et phraséologique entre les unités lexicales : il faut répondre entre autres à la question qui concerne le nombre de nœuds à partir desquels il y a rupture entre les éléments d'un réseau (voir Mejri 2020, à paraître) ;
- La question relative à l'élaboration concrète des champs sémantiques et lexicaux : les critères retenus, les délimitations et contours de chaque champ et les chevauchements entre champs.

Toutes ces questions seront développées dans des travaux ultérieurs.

Salah Mejri et Lichao Zhu
ssalah.mejri@gmail.com, lichao.zhu@gmail.com
TTN Sorbonne Paris Cité, Université Paris 13

Bibliographie

Autour d'un dictionnaire : le Trésor de la langue française (1990), Didier Érudition, CNRS.
BEJOINT Henri (2007), « Informatique et lexicographie de corpus : les nouveaux dictionnaires », in *Revue française de linguistique appliquée*, n° 1, pp. 7-23.
COLSON Jean-Pierre (2018), « Les traces du figement dans les corpus linguistiques : une étude de cas », *Le français moderne*, n°1, pp. 129-145.
CORBIN Danielle (1982), « Le monde étrange des dictionnaires : sur le statut lexicographique des adverbes en *-ment* », in *Lexique*, n° 1, Presses universitaires de Lille, pp. 149-158.
DODGE Yadolah (1993), *Statistique Dictionnaire encyclopédique*, Paris, Dunod.
ESCODA Xavier-Blanco et MEJRI Salah (2018), *Les pragmatèmes*, Paris, Classiques Garnier.
FAVAREAU Donald (2010), « Introduction : an evolutionary history of biosemiotics », *Essential Reading in biosemiotics*, Dordrecht, Springer.
FRANÇOIS Jacques (2017), *La genèse du langage et des langues*, Auxerre, Sciences Humaines Editions.
FRUCHTERMAN Thomas M. J. et REINGOLD Edward M. (1991), « Graph drawing by force-directed placement », in *Software-Practice & Experience archive*, vol. 21, n°11, pp. 1129-1164.
GOUDAILLER Jean-Pierre (1997), *Comment tu tchatches ! Dictionnaire du français contemporain des cités*, Paris, Édition Maisonneuve et Larose.
GREZKA Aude et ZHU Lichao (2017), « Du figement au défigement : la reconnaissance de néologismes polylexicaux », M.-H. Viguier et A. Grezka (dir.), *in Études de linguistique appliquée*, n°2, Klincksieck, pp. 181-191.
GUILBERT Louis (1971), « Fondements lexicologiques du dictionnaire », *Grand Larousse de la langue française*, Larousse, Paris.
KLINKENBERG Jean-Marie et *al.* (1997), *Une langue, une communauté. Le français en Belgique*, Louvain-la-Neuve, Duculot.
La Banque des mots (2015), n° 89, CILF, Paris.
LEBART Ludovic et SALEM André (1994), *Statistique textuelle*, Paris, Dunod.
MARTIN Robert (2001), *Sémantique et automate*. Coll. *Écritures électroniques*, Paris, PUF.
MARTINELLI Dario (2010), *A Critical Companion to zoosemiotics*, Dordrecht, Springer.
MEJRI Salah (1997), *Le figement lexical : descriptions linguistiques et structuration sémantique*, Publications de la Faculté des lettres de la Manouba, Tunis, Tunisie.
MEJRI Salah (2003), « Polysémie et polylexicalité », *in Syntaxe et sémantique*, n°5, Presses universitaires de Caen, pp. 13-30.
MEJRI Salah (2006), « Polylexicalité, monolexicalité et double articulation : la problématique du mot », *Cahiers de Lexicologie*, n° 89, pp. 209-221.
MEJRI Salah (2018), « La phraséologie française : synthèse, acquis théorique et descriptifs », *in Le Français Moderne*, CILF, Paris, pp. 5-32.
MEJRI Salah (2020, à paraître), Compte rendu *La genèse du langage et des langues*, Éditions Sciences Humaines, 2017, dans *Bulletin de la Société de linguistique de Paris*.
MILLER Philip H. & TORRIS Thérèse (1990), *Formalismes syntaxiques pour le traitement automatique du langage naturel*, Paris, Hermès.
PRUVOST Jean (2002), *Les dictionnaires de langue française*, collection *Que-sais-je*, n° 3622, Paris, PUF.
TOCCAFONDI Fiorenza (1999), « De Karl Bühler à Karl R. Popper », in *Philosophiques*, vol. 26, n°2, Société de philosophie du Québec, pp. 279-300.
ZHU Lichao (2019, à paraître), « Moule locutionnel lexicographique et traitement des phraséologismes », *in Cahiers du dictionnaire*, n° 11, Classiques Garnier.

ANNEXES

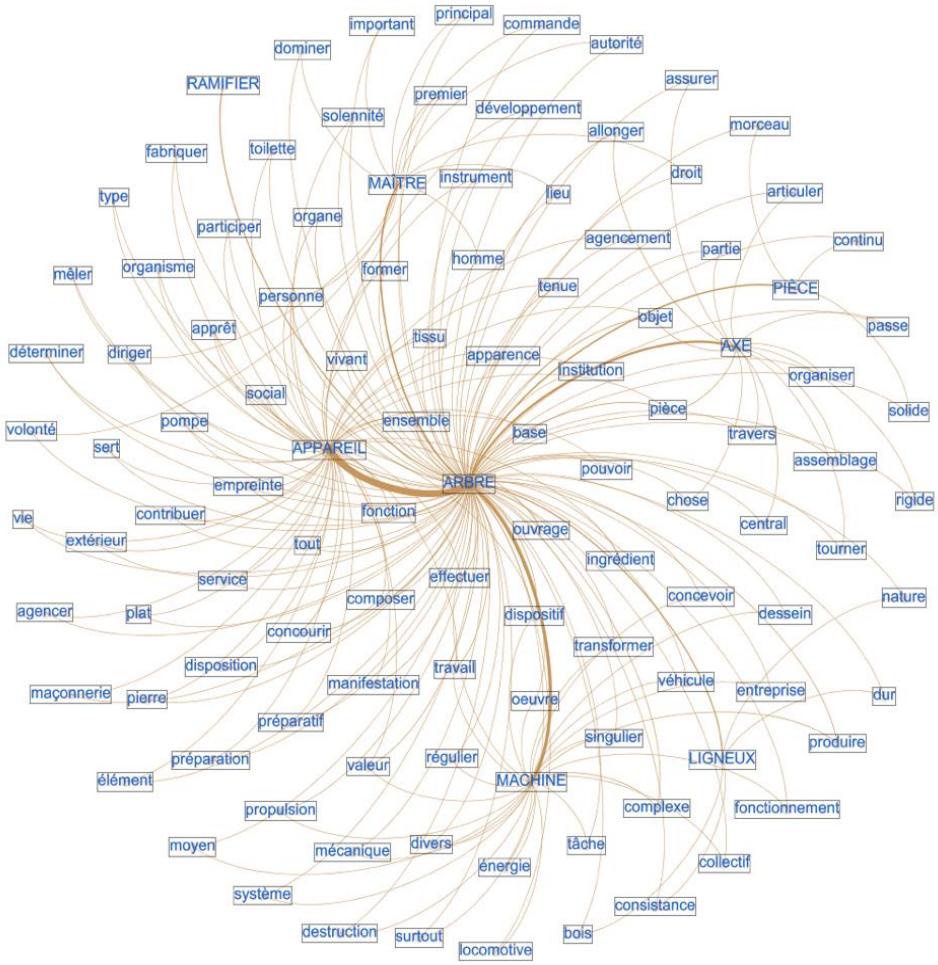

Figure 18. Extrait du réseau sémantique du mot Arbre

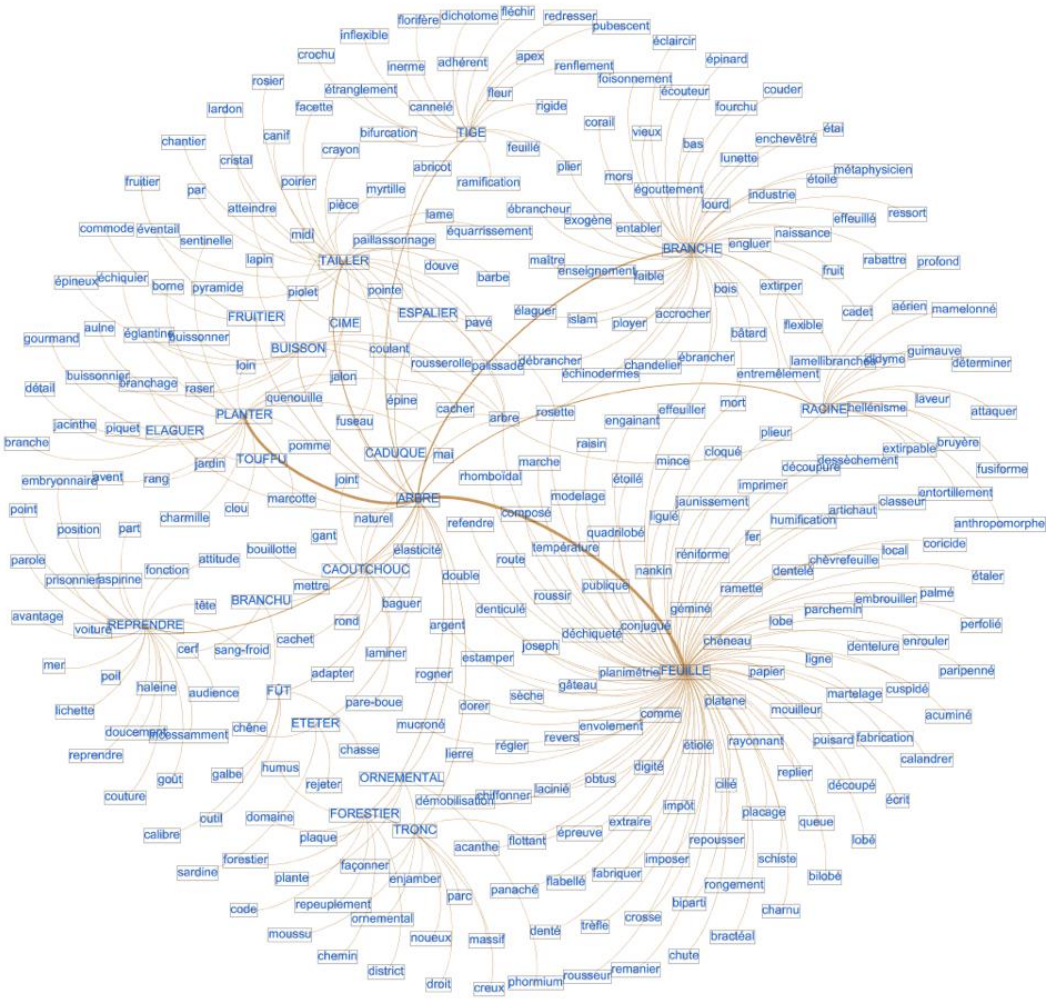

Figure 19. Extrait du réseau phraséologique du mot Arbre

Résumé

Nous partons de l'idée que le dictionnaire représente la première forme de simulation de la langue. C'est à ce titre qu'il revendique l'exhaustivité et la systématicité dans les descriptions. Partant des unités de la troisième articulation du langage, les unités lexicales, où se croisent tous les ingrédients de la langue (phonologie, sémantique, syntaxe, pragmatique), il essaie de rendre compte de l'extrême complexité du fonctionnement de la langue décrite.

L'informatisation des dictionnaires offre l'occasion d'exploiter les descriptions lexicographiques disponibles pour une éventuelle automatisation. Pour ce faire, nous avons élaboré une chaîne de traitement automatisée qui part des données de la neuvième édition du *Dictionnaire de l'Académie française*. Les données sont collectées, nettoyées et traitées en vue de deux opérations fondamentales : l'extraction des définitions et des phraséologismes, et la génération des réseaux référentiels et phraséologiques.

Les résultats obtenus permettent de confirmer l'hypothèse que tout dans la langue est structuré en réseaux. Même si une telle hypothèse est théoriquement bien établie, avec l'outil informatique, il est possible de la visualiser et d'avoir accès à des données de grande taille jusque-là peu accessibles.

Mots-clés

Dictionnaire informatisé, réseaux référentiels, phraséologie, réseaux phraséologiques, chaîne de traitement automatisée.

Abstract

The starting point of our article is that the dictionary represents the first form of simulation of the language. For this reason, it claims exhaustiveness and systematicity in the descriptions. Beginning with the units of the third articulation of language, the lexical units, where all the ingredients of the language (phonology, semantics, syntax,

pragmatics) intersect, the dictionary seeks to summarise the extreme complexity of the functioning of the described language.

The computerization of the dictionaries helps to exploit available lexicographical descriptions for a potential automation. To do this, we have developed an automated processing chain based on data of the ninth edition of the Dictionnaire de l'Académie française. Data is collected, cleaned and processed for two basic operations: extraction of definitions and phraseologisms as well as generation of referential and phraseological networks.

The results obtained confirm the hypothesis that all in the language is structured in networks. Even though the hypothesis is theoretically well-established, we are now able to visualize it with computer tool and to have access to big data hitherto hardly accessible.

Keywords

Computer dictionary, referential networks, phraseology, phraseological networks, automated processing chain.