Etude terminologique de la chimie en arabe dans une approche de fouille de textes

Baian Albeiriss

N° d'ordre NNT : 2018LYSE2057

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 484 Lettres, Langues, Linguistique et Arts

Discipline : Lexico Terminologie

Soutenue publiquement le 7 juillet 2018, par :

Baian ALBEIRISS

Étude terminologique de la chimie en arabe dans une approche de fouille de texte.

Devant le jury composé de :

Salam DIAB-DURANTON, Professeure des universités, COMUE Université Grenoble Alpes, Présidente

Abdenbi LACHKAR, Professeur des universités, Université Montpellier 3, Rapporteur

Abdelfattah BRAHAM, Professeur d'université, Université de Sousse, Rapporteur

Xavier LELUBRE, Professeur des universités, COMUE Université Grenoble Alpes, Examinateur

Sabrina BOULESNANE, Maître de conférences, Université Jean Moulin Lyon 3, Examinateur

Zoubeir MOUELHI, Maître de conférences, Université Paris 3, Examinateur

Mohamed HASSOUN, Professeur des universités, Ecole Nationale Sup. Sces de l'Information & Bibliothèque, Directeur de thèse

Contrat de diffusion

Université Lumière Lyon 2

École Doctorale : ED 484 - Lettres, Langues, Linguistique et Arts

Lexicologie et Terminologie Multilingues Traduction

ICAR - Interactions, corpus, apprentissages, représentations

Étude terminologique de la chimie en arabe

dans une approche de fouille de textes

Baïan ALBEIRISS

Sous la direction de Mohamed HASSOUN

7 Juillet 2018

Composition du Jury :

M. Mohamed HASSOUN, Professeur des Universités, ENSSIB, Lyon

M. Abdelfattah BRAHAM, Professeur, Université de Sousse, Tunisie

M. Abdenbi LACHKAR, Professeur des Universités, Université Montpellier 3

M. Zoubeïr MOUELHI, Maître de conférences, Université Sorbonne-Paris 3

Mme Salam DIAB-DURANTON, Professeure des Universités, Université Grenoble-Alpes

Mme Sabrina BOULESNANE, Maître de conférences, Université Jean Moulin Lyon 3

M. Xavier LELUBRE, Professeur honoraire, Université Grenoble-Alpes

Dédicace

À ma famille,

Mes grands-parents, Mohamed-Nour et Hiam,

Mes parents, Nudar et Majd,

Mes frères, Anas, Aïman et Homam,

Mes belles-sœurs, Oula, Nisrine et Claire,

Mes nièces, Arije, Joudy et Talia,

Mes cousines, Chahama, Aya et Tala,

À mes amis,

Loubna, Bayanne, Rasha, Hana

Nadia, Sarah, Samia, Khadida, Hanissa, Abir

Sakina, Eve, Pascaline, Neusa

François, Magali, Jacques, Éliane, Yvette, Geneviève, Claire

Nadjib, Saloua, Izabella, Ramzi

Abdelmajid, Jaber, Badera, Gulluzar, Catherine, Danièle

Mohamed, Hassan

À mes enseignants,

Imane, Karima, Saïda

Mathilde, Michel

Zoubeir, Mohamed, Ramzi, Hassan, Chérif

…Je dédie ce travail…

Remerciements

Je tiens à remercier les personnes qui m'ont aidée à réaliser ce travail, particulièrement :

Mon ancien directeur, Monsieur Xavier Lelubre, qui a bien voulu diriger cette thèse. Ses conseils, ses critiques constructives et ses encouragements ont permis à ce travail de voir le jour.

Mon nouveau directeur, Monsieur Mohamed Hassoun, qui a accepté de poursuivre la direction de cette thèse suite au départ en retraite de M. Lelubre. Je tiens à le remercier chaleureusement pour son encadrement, son soutien sans failles et sa disponibilité. Je le remercie de m'avoir faite confiance et de m'avoir donnée la chance de participer aux projets de l'équipe.

Tous mes remerciements également à Monsieur Abdelfattah Braham, Monsieur Abdenbi Lachkar, Monsieur Zoubeïr Mouelhi, Madame Salam Diab et Madame Sabrina Boulesnane pour avoir accepté de juger ce travail et pour l'intérêt qu'ils ont porté au sujet. Je les remercie pour leurs encouragements, leurs relectures, leurs critiques constructives et leurs conseils.

Je n'oublie pas de remercier tous mes collègues et tous mes professeurs de l'Université Lumière Lyon 2, de l'ENS de Lyon et de l'ENSSIB, pour leurs encouragements et leurs conseils.

Toute ma reconnaissance et toute ma gratitude à ma famille et à mes amis pour leur soutien moral et pour leur patience. Leur appui et leurs encouragements indéfectibles m'ont permise de mener à terme ce travail de thèse.

Système de transcription de l’arabe

Consonnes	
ء	ʾ
ب	b
ت	t
ث	ṯ
ج	ǰ
ح	ḥ
خ	ḫ
د	d
ذ	ḏ
ر	r
ز	z
س	s
ش	š

Consonnes	
ص	ṣ
ض	ḍ
ط	ṭ
ظ	ẓ
ع	ʿ
غ	ġ
ف	f
ق	q
ك	k
ل	l
م	m
ن	n
و	w

Consonnes	
ه	h
ي	y

Voyelles brèves	
◌َ	a
◌ُ	u
◌ِ	i

Voyelles longues	
ا	*â*
و	*û*
ي	*î*

Table des matières

Liste des tableaux

Liste des figures

Liste des annexes

Introduction générale

Suite à l'explosion des données textuelles sur Internet, la production de documents sous forme électronique s'accélère sans cesse. Or, pour produire, diffuser, rechercher et exploiter ces documents, les outils de gestion de l'information ont besoin de ressources lexicales et terminologiques telles que les bases de connaissances lexicales, les index, les réseaux lexicaux et les ontologies... Ces nouvelles productions lexicales et terminologiques entraînent des changements en profondeur, entre autres, de la pratique terminologique, conduisant essentiellement à une tâche d'analyse de corpus textuels, appelant ainsi à un renouvellement théorique de la terminologie (Bourigault et al., 1999)[1].

Terminologie et informatique se trouvent liées pour répondre au défi actuel : rendre accessibles ces informations le plus rapidement possible. Par conséquent, les outils d'analyse de corpus (Traitement Automatique des Langues) et la conception de méthodes d'analyse terminologique ou conceptuelle de textes (Linguistique et Ingénierie des Connaissances) ont connu ces dernières années une croissance importante.

Contrairement au français et à l'anglais, l'arabe a suscité un intérêt bien moindre jusque-là dans ces domaines. La relative complexité de cette langue rend quasiment inapplicables sur l'arabe les outils de traitement des langues généralistes. En effet, la langue arabe est considérée comme une langue complexe pour le traitement automatique du langage écrit (Dichy, 1990)[2]. Les résultats obtenus avec les outils d'analyse morphologiques et syntaxiques présentent beaucoup d'ambigüités qui sont dues, d'une part, à des phénomènes linguistiques tels que l'agglutination et la flexibilité de l'ordre des constituants, et d'autre part, à l'absence de voyellation et aux habitudes orthographiques à l'écrit (Abbes, 2004)[3].

Ces défis et ces ambiguïtés nous intriguent, nous intéressent, nous passionnent…

Problématique de la thèse

Notre formation initiale en chimie a rendu la question du domaine d'étude d'une évidence certaine. En effet, nous sommes titulaire d'un master en chimie, spécialité chimie

[1] Pour une terminologie textuelle

[2] L'écriture dans la représentation de la langue : la lettre et le mot en arabe

[3] La conception et la réalisation d'un concordancier électronique pour l'arabe

inorganique. D'autre part, la langue de la chimie est relativement peu étudiée (Peraldi, 2011)[4], en comparaison avec d'autres domaines scientifiques. Pourtant, elle est partie intégrante de notre vie à travers la nourriture, les vêtements, les technologies… De plus, la chimie est l'une des rares disciplines à bénéficier de nombreuses et constantes activités normalisatrices (Humbley, 1996)[5], comme le système international de la nomenclature qui permet de nommer les substances chimiques. Malgré l'importance de cette nomenclature internationale en chimie, il existe des ambiguïtés pouvant gêner la communication scientifique, notamment en raison du manque de communication entre experts, mais aussi entre spécialistes et non-initiés. Dans certains cas, cela complexifie la structuration conceptuelle et la transmission des connaissances spécialisées au sein de la discipline (Peraldi, 2012)[6].

Concernant l'arabe, ces ambiguïtés sont plus accentuées, puisque la terminologie scientifique et technique en général, la terminologie de la chimie en particulier, est tributaire du français et/ou de l'anglais. Par conséquent, l'arabe doit développer une forte créativité lexicale et terminologique en vue de compenser ce manque dans son lexique général et dans ses différents vocabulaires, y compris le vocabulaire de la chimie, à travers la création de différents organismes et l'optimisation de ceux qui existent, comme les académies de la langue arabe (Jaber, 2012)[7]. Analyser cette création lexicale s'avère intéressant ; cela l'est encore davantage dans les domaines de spécialité où il est plus précisément question d'étude et de création terminologiques. Avec un corpus de textes de chimie en arabe et des outils d'analyse de ce corpus, cette étude terminologique permettrait notamment de répondre aux questions suivantes, sans être pour autant exhaustives :

- Quels procédés de construction de termes sont les plus représentatifs dans la chimie en arabe ? Et par conséquent, quelle est la fréquence moyenne d'un procédé de construction morphologique en terminologie ?
- Quelles unités lexicales morphologiquement complexes ayant une base elle-même complexe peuvent être utilisées comme/dans des UTC ou UTS ?
- Quelle est la fréquence des verbes triconsonantiques pouvant servir la terminologie de la chimie ?

[4] Indétermination terminologique et multidimensionnalité dans le domaine de la chimie organique : analyse à partir d'un corpus spécialisé de langue anglaise

[5] La légitimation en terminologie, p. 123

[6] L'anglais de spécialité en chimie organique : entre indétermination terminologique et multidimensionnalité

[7] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours

- Quelle est la fréquence des noms par rapport aux verbes dans le domaine de la chimie en arabe ?
- Quels sont les termes dérivés des emprunts ?
- Quels sont les affixes et les formants de la chimie en arabe ?

À ces questions, nous proposons d'ajouter des questions qui préoccupent plus spécifiquement les chimistes ; par exemple :

- Quel composé chimique réagit avec telle ou telle molécule ?
- Quelles sont les méthodes d'analyse de tel ou tel produit chimique ?
- Quelle est la quantité de réactif à introduire ?

Cette étude terminologique de la chimie, les problèmes du traitement de l'arabe et les défis des systèmes d'information constitue un sujet de recherche prometteur.

Notre objectif est alors de mettre en place un outil de désambigüisation sémantique, pouvant être intégré à diverses applications du TAL telles que l'extraction d'information, la recherche d'information, la construction de dictionnaires terminologiques basés sur corpus, la construction d'ontologie et l'aide à la traduction spécialisée. Ces applications pourraient refléter la richesse de langue arabe, et son importance en termes de diffusion et de nombre de locuteurs (Abdul Hay, 2012)[8].

Notre travail vise à constituer un outil d'extraction des termes de la chimie en arabe et de leurs relations. Mais, une recherche pertinente en arabe passe obligatoirement par un système automatisé du traitement de la langue. Traiter automatiquement les corpus écrits en arabe ne peut se faire sans analyse linguistique. Cette analyse linguistique, plus précisément, cette étude terminologique est la base pour la construction des règles d'une grammaire d'identification afin de déterminer les termes de la chimie en arabe. La construction de cette grammaire d'identification nécessite la modélisation des patrons morphosyntaxiques à partir de leur observation en corpus et débouche sur la définition de règles de grammaire et de contraintes (Aussenac-Gilles et al., 2000)[9]. La mise au point de patrons morphosyntaxiques permet d'automatiser partiellement la recherche de traces de relations sur un corpus ; l'informatique a ceci de plaisant. Pour cela, il faut traduire ces patrons en language informatique de manière à en obtenir une représentation formelle. Cette représentation formelle est mise en œuvre par la

[8] Constitution d'une ressource sémantique arabe à partir de corpus multilingues alignés
[9] Les relations sémantiques : du linguistique au formel

machine pour obtenir le résultat recherché (Bachimont, 2000)[10]. Malgré les dernières avancées, la configuration d'un système d'extraction d'information est un travail d'ingénierie long et complexe (Toussaint, 2011)[11].

En d'autres termes, formaliser les caractéristiques terminologiques en étudiant les mécanismes de la construction morphosyntaxique des termes de la chimie en arabe. L'objectif général n'est pas seulement de proposer des patrons morphosyntaxiques de recherche automatique de ces termes. Le but consiste à mettre en évidence le lien qui existe entre les structures syntaxiques et sémantiques, et les connaissances du domaine en établissant des règles de grammaire qui régissent les termes du domaine. Ces règles seront implémentées par la suite dans un système d'analyse morphosyntaxique de l'arabe. L'outil demeure dans cette perspective un soutien nous permettant de rechercher et de valider de manière plus rapide les structures, mais aussi éventuellement d'élargir les résultats élaborés manuellement.

Pour déterminer les patrons syntaxiques des termes de la chimie en arabe, il faut analyser les modes et les moyens de la formation de termes scientifiques en arabe. Le modèle d'André Roman (1990)[12] a été largement testé dans un grand nombre de domaines de spécialité, notamment en physique (Lelubre, 1992)[13], a montré sa capacité à apporter de réponses satisfaisantes et a donné des éclaircissements pertinents. Cependant, l'identification de ces patrons syntaxiques, comme étant potentiellement des structures de termes, sont bruités, au sens où les candidats termes ainsi repérés ne correspondent pas tous à des termes (Toussaint, 2011). Par conséquent, nous construisons une classification des termes de la chimie en arabe de notre corpus que nous intégrons à la grammaire d'identification.

La mise en place de cet outil requiert et nécessite à la fois les compétences d'un expert du domaine, des compétences en modélisation des connaissances et en linguistique et des compétences en informatique, soit « un oiseau rare » (Bourigault et al., 2004)[14]. Cet oiseau rare, nous ne prétendons pas l'incarner. Par conséquent, nous mettons en place une collaboration entre acteurs de différentes spécialités. Nous concernant, nous possédons une certaine expertise du domaine de la chimie ; il reste alors le problème de langue et de sa connaissance afin de bien comprendre les spécifications de l'outil et être capable de les traduire

[10] Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en Ingénierie des connaissances

[11] Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances, p. 8

[12] Grammaire de l'arabe

[13] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation

[14] Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas

en langage informatique pour les implémenter et/ou savoir dialoguer avec les informaticiens qui le développent.

C'est dans cet élan d'ambition que nous abordons notre recherche. Elle a été initiée lors de notre stage de six mois dans le cadre du master SIMIL-TRA au sein de l'entreprise Techlimed[15], disposant d'outils de TAL pour l'arabe ; les premiers résultats obtenus sont très concluants. Cependant, lors de notre thèse, nous n'avons pas pu utiliser ces outils à notre grand regret, puisque cette société a disparu (dépôt de bilan), emportant avec elle ses secrets. D'autre part, une collaboration avec l'université libanaise à Beyrouth au Centre des sciences du langage et de la communication avec M. Ghassan Mourad a été envisagée. Mais elle n'a pas abouti, notamment en raison de la situation conflictuelle au Proche-Orient (zone déconseillée par le Quai d'Orsay). Sans la plateforme de Techlimed et sans collaboration, nous avons dû faire quelques concessions, notamment limiter notre travail de recherche à la conception de l'outil et non à sa réalisation. De plus, nous choisissons également de n'utiliser que des outils informatiques en accès libre, impliquant une certaine lenteur dans l'avancement des recherches. Ce choix est motivé, notamment par notre volonté de mette en ligne, par la suite, ce travail afin qu'il soit implémenté et validé par d'autres laboratoires disposant de ces outils d'analyse.

Plan de la thèse

Ce travail est constitué de trois parties qui s'étalent sur un total de sept chapitres.

Dans une première partie, nous étudions la terminologie de la chimie en arabe, sans viser une présentation exhaustive de la discipline et de son développement (prémices, évolution de la théorie, multiplication des domaines de recherche, etc.).

Le premier chapitre rappelle d'abord brièvement les différentes approches de la terminologie et ses acceptions, notamment la différence entre terminologie et nomenclature. Puis, nous passons en revue les principales théories, la richesse et la variété des sujets de recherche, les contacts stimulants avec les domaines de la terminologie qui font de cette discipline un carrefour interdisciplinaire privilégié. Enfin, nous traitons la terminologie scientifique à travers la langue de spécialité ainsi que la normalisation terminologique, notamment le cas de la terminologie scientifique arabe qui est tributaire des terminologies établies en anglais ou en français, avec ses différents organismes terminologiques.

[15] Fondé en 2009 par Ramzi Abbès, Techlimed était une société qui avait pour projet la création d'un moteur de recherche en langue arabe ainsi que le développement d'un logiciel de traduction automatique de la langue arabe sur le net.

Dans le deuxième chapitre, nous analysons le mode de formation des termes et leur comportement en contexte dans le cadre de la terminologie arabe de la chimie par une étude morphosyntaxique et sémantique des unités terminologiques simples et complexes, grâce notamment aux systèmes de nomination et de communication de la langue arabe par l'intermédiaire des procédés linguistiques de dérivation et de composition mais aussi par les transferts sémantiques et les emprunts, à partir des travaux de Roman (1990)[16] et de Lelubre (1992)[17].

La seconde partie est consacrée à la présentation de notre corpus de travail, un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages. Habituellement, cette phase met en évidence l'appareillage technique dû au développement des corpus. Il s'appuie sur des critères privilégiés (taille, source…) et en délaisse d'autres. Certains ne paraissent pas tellement évidents et méritent d'être développés ; nous avons choisi de les aborder, notamment la recherche d'un corpus en arabe (protocole de recueil) et les outils techniques pour son analyse (méthode d'analyse), critères participant à la qualité des corpus.

Le troisième chapitre explique les éléments essentiels qui font d'un recueil de données un corpus ainsi que les critères choisis tels que le format, le cadre et le canal, mais aussi les traitements manuel et semi-automatique réalisés sur ce corpus.

Confronté à un corpus spécialisé, le quatrième chapitre décrit le dépouillement des données en utilisant notamment des logiciels AntConc, Kawâkib et Xerox afin d'identifier les termes de la chimie en arabe.

Dans le cinquième chapitre, nous étudions le domaine de la chimie, depuis son origine à nos jours. Nous soulevons également les questions inhérentes aux frontières du domaine. En effet, les interférences sont nombreuses avec des disciplines voisines comme la physique et la biologie. Nous consultons également différentes classifications pour aboutir à notre classification des termes du domaine de spécialité.

Enfin, dans une dernière partie, au vu de notre étude terminologique exposée dans la première partie et de la classification construite dans la deuxième, nous proposons une modélisation des termes de la chimie en arabe à partir d'une grammaire d'identification à base de règles

[16] Grammaire de l'arabe

[17] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation

permettant de traduire en langage informatique les modèles de construction des unités terminologiques prenant en compte leurs structures internes, leurs usages et leurs environnements textuels.

Dans le sixième chapitre, nous décrivons le processus de l'extraction des unités terminologiques simples et complexes de la chimie. Nous présentons les différents patrons morphosyntaxiques de ces termes ainsi que les règles que nous avons établies pour leur identification. Ces règles s'appuient sur la reconnaissance des termes et sur les expansions en syntagmes, repérées à partir de ces termes que nous avons établis sur la base de leur caractérisation morphosyntaxique.

Le septième et dernier chapitre expose une des applications de ce travail : l'ontologie. Pour la terminologie, l'ontologie est une des perspectives les plus prometteuses. Nous utilisons l'environnement Protégé pour la construction des ontologies du domaine en raison de son accès libre et de la facilité d'utilisation. Cette ontologie propose une représentation des connaissances du domaine de la chimie, à travers ses termes, notamment les verbes spécialisés, et ses relations, comme les relations de hiérarchisation telles que la synonymie et l'hyperonymie…

Qu'il me soit permis de présenter ce travail de recherche, de la part d'une chimiste, reconvertie en linguiste et passionnée d'informatique.

Partie I : Étude terminologique

> *« Nous abordons le monde par perception, imagination, entendement. Nous relevons dans les objets des propriétés (couleur, structure moléculaire, masse atomique, rayonnement, etc.). Nous appréhendons les objets au travers des concepts que nous nous en faisons. Nous identifions par-là les objets en leur attribuant des caractères (en les caractérisant). Et nous désignons ces concepts grâce à des représentations symboliques (dessins, images, signes linguistiques, formules, équations, etc.). Voilà un schéma simplifié par lequel raisonnablement entrer en terminologie. C'est dire que cette présentation est phénoménologique et axée sur les choses. Car la terminologie est une discipline éminemment pratique et tournée vers le monde. Elle tient en cela de la démarche du scientifique, qui décrit les propriétés des objets en les conceptualisant et en les classant ». (Depecker, 2005)*[18]

Dans un travail de fouille de textes, il est nécessaire d'identifier les mots qui composent les textes recueillis pour le domaine étudié ainsi que leur formation ; cela correspond à l'étude terminologique du domaine étudié. Dans cette partie, nous présentons une analyse terminologique de la chimie en arabe afin d'identifier les patrons morphosyntaxiques des termes du domaine et les modéliser dans l'extracteur terminologique.

Dans un premier temps, nous rappelons les différentes approches de la terminologie en passant en revue brièvement ses principales théories. Puis, nous présentons la normalisation terminologique (avec ses différents organismes terminologiques) et le cas de la terminologie arabe, tributaire des terminologies établies en anglais ou en français. Enfin, nous analysons les moyens dont la langue arabe dispose pour la création et la formation des termes de la chimie.

[18] Contribution de la terminologie à la linguistique, p. 7

Chapitre 1 : État de l'art de la terminologie

Dans ce premier chapitre, nous nous intéressons à la notion de terminologie. Pour cela, nous définissons d'abord la terminologie de la terminologie, le terme ou encore la définition, afin d'expliquer les éléments essentiels d'une fiche terminologique. Puis, nous décrivons la terminologie comme un carrefour des sciences, notramment entre la linguistique et l'informatique. Ensuite, nous présentons les évolutions de la terminologie, de son origine jusqu'aux dernières approches. Enfin, nous exposons la notion de normalisation terminologique dans le contexte d'une langue de spécialité en arabe.

1.1 Terminologie : définitions

La terminologie peut être définie comme la science qui étudie les termes :

> *« La terminologie est la discipline ou science qui étudie les termes, leur formation, leurs emplois, leurs significations, leur évolution, leurs rapports à l'univers perçu ou conçu » (Gouadec, 1990)[19].*

Elle possède différentes acceptions :

> *« Le mot 'terminologie' désigne au moins trois concepts différents :*
>
> - *L'ensemble des principes et des fondements conceptuels qui régissent l'étude des termes ;*
> - *L'ensemble des règles qui permettent de réaliser un travail terminographique ;*
> - *L'ensemble des termes d'un domaine de spécialité donné.*
>
> *Si la première acception fait référence à la discipline et la deuxième à la méthodologie, la troisième désigne l'ensemble des termes d'un domaine donné ». (Cabré, 1998)[20]*

La terminologie peut désigner également « le vocabulaire des langues de spécialités, voc*abulaire spécialisé […] et d'autre part, les méthodes propres au travail terminologique* » (CST (Conférence des Services de traduction des États de l'Europe occidentale), 1990)[21].

[19] Terminologie. Constitution des données, p. 19
[20] La terminologie - Théorie, méthode et applications, p. 70
[21] Recommandations relatives à la terminologie, p. 12

Généralement, les deux acceptions les plus utilisées sont la terminologie dans le sens de vocabulaire, l'ensemble de termes se rapportant à un domaine particulier, et la terminologie au sens de la discipline qui étudie les termes.

De plus, la terminologie a fini par devenir une discipline à part entière :

> *« Née d'un besoin de précision et de clarté dans les langues de spécialité la terminologie s'est d'abord confondue avec la lexicologie et surtout la lexicographie pour s'en dégager par la suite en identifiant son objet propre et ses méthodes grâce à ses moyens empiriques ». (Rondeau, 1991)*[22]

Mais pour l'arabe, Hardane (1994)[23] considère que la terminologie n'est pas encore arrivée à se distinguer clairement de la lexicographie, notamment en raison de son attachement à cette tradition lexicographique et au caractère d'univocité exigé dans le travail terminologique qui est loin d'être atteint en arabe.

En ce qui nous concerne, nous considérons que la terminologie est la discipline qui étudie les termes et que le vocabulaire est l'ensemble de termes se rapportant à un domaine particulier.

1.1.1 Terme

Le mot est l'objet d'étude de la lexicologie et le terme est celui de la terminologie. Nous nous référons à Lehman et Martin-Berthet pour définir le mot :

> *« Le mot est reconnu comme une unité de base de la langue. L'identité d'un mot est constituée de trois éléments : une forme, un sens et une catégorie grammaticale ». (Lehmann et al., 2005).*[24]

Cet élément linguistique significatif, composé d'un ou de plusieurs phonèmes, est compris entre deux blancs et dénote un objet (substantif), une action ou un état (verbe), une qualité (adjectif), une relation (préposition). Quant au terme, il peut être défini comme étant « une unité linguistique désignant un concept, un objet ou un processus. Le terme est l'unité de désignation d'éléments de l'univers perçu ou conçu. Il ne se confond que rarement avec le mot orthographique » (Gouadec, 1990)[25]. Mais, le terme posséde un caractère référentiel qui doit être pris en compte :

[22] Introduction à la terminologie, p. 38
[23] Rôle du français dans l'élaboration terminologique arabe, p. 482
[24] Introduction à la lexicologie, p. 3
[25] Terminologie. Constitution des données, p. 19

« Dans un domaine de spécialité, le terme n'est que la dénomination d'une unité référentielle, extralinguistique. Le terme ne se définit pas par rapport à d'autres termes, mais par rapport à son référent ». (Lelubre, 1992)[26]

En effet, le terme est ici considéré comme le signe linguistique de Saussure comportant un signifié et un signifiant :

« Le terme est essentiellement un signe linguistique à double face (notion et dénomination) faisant partie d'un ensemble notionnel donné et se définissant par rapport à cet ensemble dans un domaine scientifique ou technique excluant la langue commune ». (Rondeau, 1991)[27]

Citons l'exemple 'eau' : employé dans le lexique, il devient un terme en chimie et renvoie à la molécule d'eau, à un liquide, à une solution… et il est également employé dans d'autres domaines avec d'autres acceptions comme en physique ou en géographie.

De plus, Le Guern parle plutôt de mot et de lexique et les oppose respectivement à terme et à terminologie par rapport à la relation avec les choses :

« Le lexique d'une langue est constitué de l'ensemble des mots de cette langue. (..) Les mots de la langue ne sont pas en relation immédiate avec les choses. Ils ont un signifié, mais n'ont pas de référence (..) le lexique de la langue ne contient pas de substantifs ; les mots que les lexicographes désignent comme substantifs ne sont en réalité que des prédicats, ils disent des propriétés et non des substances, des qualités et non des objets. Le signe d'une propriété est un prédicat, le signe d'un objet est un terme. Les termes ne font pas partie du lexique, c'est le discours qui les construit. (…) Pour designer des objets, le discours ne met pas seulement en œuvre le lexique, il a aussi besoin de la syntaxe. L'unité minimale de discours qui a la possibilité de signifier un objet est le syntagme nominal. (..) La fermeture du prédicat par le quantificateur le transforme en terme. » (Le Guern, 1989)[28]

L'arabe emploie la désignation « مصطلح = *muṣṭalaḥ* = terme » ; il s'agit du participe passif ; dérivé du nom d'action « اصطلاح = *iṣṭilâḥ* = convention », du verbe « اصطلح / يصطلح = *iṣṭa*la*ḥ*a / *yaṣṭaliḥ*u », qui a pour schème la 7ème forme de la tradition grammaticale des arabisants « افتعل / يفتعل= iftaʿala / yaftaʿilu » (à l'origine ; le « ط = ṭ » devait être un « ت = t » comme le schème l'indique ; mais pour des raisons phonologiques dues à la présence du « ص = *ṣ* » lettre emphatique, le « ت = t » a été remplacée par « ط = ṭ ») et dont la racine est « ص ل ح = *ṣ* l *ḥ* », signifiant le contraire de « فساد = fasâd = altération » (Talafheh, 2003)[29]. Cette acception du

[26] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 11
[27] Introduction à la terminologie, p. 43
[28] Sur les relations entre terminologie et lexique, p. 340
[29] La terminologie grammaticale complexe dans le Kitâb de Sîbawayhi, p. 18

terme correspond à son sens lexical ; son sens technique est « *l'expression de l'accord des gens* sur la nomination d'une chose au moyen d'un nom qui a été transposé de son premier objet » (Troupeau, 1994)[30]. De ce fait, « مصطلح = *muṣṭalaḥ* = terme » n'est pris que par « son aspect conventionnel, omettant ainsi ses autres propriétés, en l'occurrence l'univocité et l'appartenance à un domaine de spécialité » (Talafheh, 2003)[31].

D'autre part, le terme peut être composé d'un seul mot, c'est une unité terminologique simple (UTS) ; il peut être également composé de plusieurs mots, c'est une unité terminologique complexe (UTC). Cette UTC, appelée aussi mot composé, segment répété, collocation, collection de mots, syntagme terminologique, trouve de nombreuses définitions dans la littérature :

> *« Une unité de signification résultant d'un groupement ou d'une association, avec ou sans joncteur, d'un ensemble de lexèmes qui suivent les règles syntaxiques de phrase propres à une langue donnée et qui expriment, de manière univoque, un concept appartenant à un domaine déterminé de la connaissance » (Reguigui, 2002) [32].*

Une UTC serait alors une combinaison récurrente de mots qui se trouvent ensemble plus souvent que par le simple fait du hasard et qui correspond à une utilisation arbitraire (Smadja, 1993)[33].

Enfin, la plupart de ces unités terminologiques sont des substantifs :

> *« La très grande majorité des entrées 'd'un glossaire ou d'un vocabulaire scientifique et technique' est représentée par des substantifs. Ce type de distribution dans les catégories grammaticales est directement dépendant de l'aspect dénominatif de la signification [..] » (Guilbert, 1973)[34]*

Aux substantifs ajoutons les verbes et les adjectifs comme unités terminologiques :

> *« Effectivement, les termes adjectivaux et verbaux font partie intégrante de toute terminologie. Un grand nombre de termes adjectivaux interviennent d'ailleurs dans les unités terminologiques substantivales […]. Quant aux unités terminologiques verbales, elles interviennent dans tout texte de spécialité […] ». (Lelubre, 1992)[35]*

[30]La formation du vocabulaire scientifique et intellectuel
[31] La terminologie grammaticale complexe dans le Kitâb de Sîbawayhi, p. 20
[32] Anatomie des syntagmes terminologiques arabes : analyse formelle et quantitative, p. 210
[33] Retrieving collocations from text : Xtract, p. 147
[34] La spécificité du terme scientifique et technique, p. 17
[35] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 39

Concernant la question du statut du verbe en terminologie, il est à noter que le verbe a longtemps été mis de côté au profit du nom en terminologie justifiant ce désintéressement notamment par la place accordée aux objets et à leurs dénominations :

> *« La terminologie ne s'intéresse aux signes (mots et unités plus grandes que le mot) qu'en tant qu'ils fonctionnent comme des noms dénotant des objets et comme des « indicateurs de notions » (de concepts). Dans cette optique, les verbes sont des noms de processus, d'action » (Rey, 1979)*[36].

Ce n'est que récemment que les chercheurs, notamment L'Homme (1995)[37], Lerat (2002)[38], Costa et Silva (2004)[39] et Pimentel (2007)[40], ont redoré son blason en lui accordant une place dans les travaux terminologiques ; nous partageons cet avis, considérant notamment que le verbe peut soulever le problème de la compréhension. Mais, la question centrale est de savoir comment définir ces verbes spécialisés :

> *« La question est de savoir comment déterminer si le verbe est spécialisé ou non. On peut considérer que certains verbes peuvent véhiculer un sens spécialisé résultant d'un emploi divergeant d'emplois connus, de verbes n'existant que dans un domaine spécialisé et d'activités fortement associées à un domaine. Cela impliquerait que certains verbes sont plus spécialisés que d'autres et que le verbe n'exprime pas seulement un concept ou la caractéristique d'un concept. » (L'Homme, 2012)*[41]

Par exemple, le verbe « oxyder » est considéré comme un verbe spécialisé du domaine de la chimie du fait qu'il est exclusivement associé à des activités chimiques et non parce qu'il sert de supports linguistiques à des concepts (L'Homme, 2012)[42].

De ce fait, le verbe peut être un terme s'il respecte les critères de son attribution, à savoir, le lien qu'on peut établir avec des termes de nature nominale et la classe sémantique des verbes (Lorente, 2007)[43]. Mais, cela dépend aussi de son contexte, plus précisément de sa structure argumentale :

> *« (…) le verbe n'est pas spécialisé par lui-même. Étant une unité de nature prédicative (c'est-à-dire faisant appel à des participants appelés arguments ou actants selon le cadre utilisé pour les nommer),*

[36] La terminologie : noms et notions, p. 25

[37] Définition d'une méthode de recensement et de codage des verbes en langue technique : applications en traduction

[38] Vocabulaire juridique en schémas d'arguments juridique

[39] The verb in the terminological collocations. Contribution to the development of a morphological analyser MorphoComp

[40] O comportamento do verbo constituir em contexto de especialidade

[41] Le verbe terminologique : un portrait de travaux récents, p. 96

[42] Idem, p. 96

[43] Les unitats lèxiques verbals dels textos especialitzats. Redefinició d'una proposta de classificació

le verbe est spécialisé ou non dans la mesure où l'ensemble de sa structure argumentale est prise en compte. » (L'Homme, 2012)[44]

Par conséquent, nous prenons en compte la structure argumentale des verbes considérés comme des termes de chimie, permettant notamment d'identifier et/ou de valider les termes nominaux ; par exemple, le verbe « filtrer » utilise comme argument une large gamme de composés chimiques qui seront indexés, si ce n'est pas déjà fait.

En ce qui nous concerne, nous considérons le terme comme une unité linguistique désignant « toute entité décrite » (Depecker, 2005)[45].

1.1.2 Concept

Le concept est une notion centrale en terminologie :

« L'appréhension des objets du monde repose, en terminologie, sur le concept. Défini comme une 'unité de connaissance créée par une combinaison unique de caractères' (norme ISO 087), il permet de regrouper sous une même appellation les objets qui partagent des propriétés communes ». (Roche, 2007)[46]

Citons, par exemple, le concept de 'fonction chimique' dans lequel les termes 'acide carboxylique', 'alcool', 'ester' etc., partagent les mêmes propriétés.

Rappelons que la terminologie n'étudie pas un terme seul, mais le terme dans son contexte, sa formation et ses usages. Le terme est alors pris dans son sens linguistique en faisant allusion à son référent. Pour cela, nous nous référons au signe linguistique de De Saussure :

« Le signe linguistique unit non une chose et un nom, mais un concept et une image acoustique, le concept étant le signifié (représentation mentale qui correspond au Signifiant) et l'image acoustique, le signifiant (suite de sons articulés) ». (De Saussure, 1968)[47]

A cela Lelubre ajoute :

« La relation entre la chose – le référent (élément du réel appartenant à l'univers extralinguistique, associé au signe) - et le nom – le terme – ne peut alors être directe ; c'est une relation médiate, qui

[44] Le verbe terminologique : un portrait de travaux récents, p. 99
[45] Contribution de la terminologie à la linguistique, p. 8
[46] Le terme et le concept : fondements d'une ontoterminologie, p. 5
[47] Cours de linguistique générale, p. 98

passe par le concept, souvent appelé aussi notion, qui est alors le signifié, le terme étant le signifiant. » (Lelubre, 1992)[48]

Considérant que la terminologie suit la démarche du scientifique, à savoir, décrire les propriétés des objets en les conceptualisant et en les classant, le concept est son élément pivot :

« Pour correspondre à cette démarche, c'est de concept que l'on parle en terminologie plutôt que de notion. Concept induit conceptualiser et conceptualisation. Et bien sûr concevoir et conception. De plus concept renvoie à percept, notion répandue en psychologie et utile pour comprendre la liaison des langues aux réalités qu'elles décrivent. » (Depecker, 2005)[49]

Mais, parfois le concept peut être réduit tout simplement à la signification d'un terme :

« Le terme 'concept' est utilisé par les épistémologues de deux manières différentes. Souvent, ce terme désigne à la fois la dénomination (le terme) et le concept (la signification). Terme et concept sont alors équivalents ». (Hermans, 1989)[50]

Le terme peut être représenté par le triangle sémiotique du signe linguistique :

« Le terme, qui réunit (…) désignation et concept, renvoie à un objet ». (Depecker, 2002)[51].

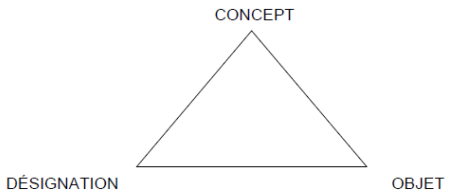


Figure 1 : Représentation graphique du terme

En ce qui nous concerne, nous utiliserons 'unité référentielle' et 'terme' à la place de 'référent' et 'signifiant' :

« L'objet de la terminologie est l'étude des relations entre des ensembles d'entités, que nous appelons unités référentielles, qu'il s'agisse d'objets (concrets) ou de concepts (abstraits), relatives à des domaines spécialisés de l'activité et de la connaissance humaines, et leur dénomination, les termes ». (Lelubre, 1992)[52]

De ce fait, pour chaque unité référentielle, il ne doit correspondre qu'un terme et un seul, sinon, il y a synonymie. De la même manière, pour chaque terme, il ne doit correspondre qu'une unité référentielle et une seule, sinon, il y a polyvalence (homonymie ou bien polyréférentialité).

[48] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 21
[49] Contribution de la terminologie à la linguistique, p. 8
[50] La définition des termes scientifiques, p. 531
[51] Entre signe et concept. - Eléments de terminologie générale, p. 21
[52] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation

En ce qui nous concerne, nous considérons le concept comme une « unité de pensée susceptible *d'être soumis à des* processus vérifiables » (Depecker, 2005)[53].

1.1.3 Dénomination

L'Homme a toujours naturellement eu besoin de nommer les objets du monde qui l'entourent, de les classer et d'établir des relations entre eux, mais surtout de comprendre les équivalences entre les nominations que chaque communauté a choisi pour ces mêmes objets afin de pouvoir échanger et communiquer :

> *« La démarche de nomination a probablement été l'une des premières démarches scientifiques de l'esprit humain. Nommer les objets, c'est les classer. Les classer selon des traits distinctifs en groupes ou ensembles organisés, cela signifie structurer ces objets, et par conséquence structurer le monde auquel ils appartiennent et qui entoure l'individu ». (Loffler-Laurian, 1984)*[54]

Dans un travail terminologique, les objets du domaine étudié doivent être nommés et désignés :

> *« Les besoins premiers, en matière de vocabulaire et dans la perspective terminologique, sont évidemment ceux de la nomination, de la désignation au moyen de signes du langage les plus appropriés (les noms) ». (Rey, 1979)* [55]

En effet, la dénomination des choses et des objets du monde est essentielle comme le disait le célèbre Linné : « Nomina si nescis, perit et cognitio rerum : Si l'on ignore le nom des choses, on en perd aussi la connaissance ».

Ainsi, la connaissance est liée à la nomination en terminologie par plusieurs aspects :

> *« Les besoins terminologiques [..] se manifestent à l'intérieur de ce cadre général du besoin de nommer. Ils concernent essentiellement la nécessité de […]*
>
> - *Construire les connaissances par les termes appropriés. Il s'agit là de l'élaboration des terminologies scientifiques, techniques et institutionnelles, avec les caractères propres qui éventuellement les opposent. [..]*
> - *Contrôler les connaissances et surtout le rapport correct entre les unités du langage, les 'termes' qu'ils expriment et les 'concepts' qu'ils sont chargés de délimiter et de véhiculer. C'est le rôle de la normalisation, entendue son sens le plus large.*

[53] Contribution de la terminologie à la linguistique, p. 8
[54] Vulgarisation scientifique : formulation, reformulation, traduction, p. 109
[55] La terminologie : noms et notions, p. 14

- *Communiquer les connaissances ainsi constituées en systèmes et contrôlées quant à leur expression, suppose enfin, plusieurs activités […] ». (Rey, 1979)*[56]

L'acte de nomination relie une unité référentielle à une unité de nomination :

> *« L'opération de nomination est essentiellement la mise en relation, idéalement biunivoque, de deux entités [...] (qui) sont toujours, nécessairement, une unité linguistique, l'unité de nomination et une unité extra- linguistique, l'unité référentielle ». (Roman, 1999)*[57]

En ce qui nous concerne, nous appellerons dénomination mais aussi désignation ou appellation le vocabulaire de la chimie tels que 'benzène' ou 'acide acétique'.

1.1.4 Nomenclature

La nomenclature s'est distinguée de la terminologie par le fait qu'il « faut comprendre par terminologie seulement la désignation des concepts, comme cristal, minéral, roche, et par *nomenclature la dénomination d'*objets, comme minéraux, sortes de roches) » (Philipsborn, 1955)[58]. Cette distinction était pratiquée surtout par les terminologues soviétiques et allemands des années 1960 et 1970. Elle peut se résumer par « nomenclature : ensemble des dénominations ; terminologie : ensemble des concepts » (Morgenroth 1994)[59]. Cette différence entre nomenclature et terminologie est aussi soulignée par Lerat (1995)[60], par l'exemple de la formule chimique qui n'a pas de morphologie, ni de fonction syntaxique et syntagmatique caractéristique de la langue naturelle en expliquant que la nomenclature est l'ensemble des moyens (symboles et des formules) propres à une science, tandis que la terminologie est l'ensemble des termes, des noms et des notions. Quant à Kocourek, il expliquait qu'une certaine distinction pourrait être envisagée, considérant la terminologie comme « *l'ensemble des unités* désignant les notions (*abstraites) d'une science, alors que la nomenclature est l'ensemble des* noms des choses concrètes, plantes, animaux, composés chimiques » (Kocourek 1991)[61]. Mais, de plus en plus, la nomenclature et la terminologie se voient considérées comme équivalentes et ne se distinguent plus :

[56] Idem, p. 18
[57] La création lexicale en arabe, ressources et limites du système de nomination dans une langue humaine naturelle
[58] Cité par Morgenroth, le terme technique, p. 245
[59] Le terme technique
[60] Les langues spécialisées, p. 53
[61] La langue française de la technique et de la science. Vers une linguistique de la langue savante, p. 182

> *« Les distinctions entre terminologies et nomenclature ont conservé une valeur essentiellement théorique, puisque les critères développés pour délimiter ces deux catégories n'ont pas abouti à des définitions rigoureuses » (Khaioutine et al., 1996)*[62]

En ce qui nous concerne, nous partageons l'avis d'Humbley que « la nomenclature ne *représente pas toute la terminologie, mais elle est significative par l'attention qu'elle porte* aux problèmes de dénomination » (Humbley, 2001)[63]. Dans la même lignée, Morgenroth développe cette distinction dans le domaine des sciences :

> *« En effet, dans les sciences naturelles, surtout en médecine, en chimie et en biologie, on distingue un système de noms génériques (ou généraux), la terminologie, de différents systèmes de mots spécifiques (ou individuels), les nomenclatures. […] Les mots répertoriés dans les nomenclatures constituent de pures énumérations, ils représentent les objets de la science donnée, tandis que les mots de la terminologie font partie du système conceptuel et méthodologique de la science » (Morgenroth, 1994)*[64].

De de fait, nous considérons que la nomenclature chimique est l'ensemble des termes désignant des composés chimiques, comme 'acide sulfurique', 'oxygène', 'dichlorométhane'. Par conséquent, le vocabulaire du domaine de la chimie se compose de la 'nomenclature chimique' et des autres du domaine comme 'filtration', 'combustion', 'bécher'…

Les termes de la nomenclature chimique doivent indiquer par leur dénomination le nom exact du composé chimique afin que la dénomination lui corresponde conformément. De plus, la dénomination doit renseigner la présence des éléments chimiques et des liaisons chimiques ainsi que leurs quantités respectives. Elle doit informer également de la structure et du type de répartition des molécules et des atomes. De ce fait, chaque composé chimique devrait avoir une et une seule dénomination mais, en réalité, il y en a plusieurs attestées à l'usage. Par exemple, la substance '1,3,5,7-tétraazatricyclo[3.3.1.1^{3,7}]décane' (nom systématique) est plus connue sous le nom 'hexaméthylènetétramine' (nom courant) ou par son abréviation 'htma'. La première dénomination est assez complexe ; mais, elle indique avec précision la structure du composé chimique alors que les deux autres dénominations facilitent la communication technique et sont conformes à la règle de l'économie du langage :

> *« […] il peut y avoir plusieurs façons de nommer un composé ou une espèce, mais aucune n'est supérieure aux autres. La complexité des noms dépend de la quantité d'information à fournir. Un nom*

[62] Tenants et aboutissants des études sur les notions de terminologie et nomenclature dans la linguistique russe et soviétique

[63] Quelques enjeux de la dénomination en terminologie

[64] Le terme technique

basé sur la composition, par exemple, sera plus simple qu'un nom fondé sur la structure (ou constitution) qui contient des données sur la disposition des atomes dans l'espace » (Leigh et al., 2001)[65].

A cela, ils ajoutent :

« Les spécialistes de la nomenclature admettent deux catégories de termes :

- *Les noms arbitraires (y compris les noms des éléments tels que sodium ou oxygène) ainsi que les abréviations usuelles (telles que diphos ou LithAl), dit noms triviaux, cet adjectif n'étant ni péjoratif ni méprisant ;*
- *Les noms de nomenclature systématique, ensemble de règles elles-mêmes arbitraires. Ces règles sont codifiées de tel sorte que chacun puisse les utiliser pour représenter les corps purs, tout comme on utilise l'alphabet pour représenter les mots ». (Leigh et al., 2001)*[66]

Rappelons que lors de la mise en place des premières classifications des composés chimiques, les scientifiques ont attribué généralement des noms triviaux en absence de la connaissance de la structure moléculaire. Par la suite, les chimistes ont gardé ces dénominations anciennes qui restent encore ancrées dans la nomenclature chimique actuelle.

La complexité de la dénomination de certains composés chimiques, comme nous venons de le voir, exige que la nomenclature soit normalisée. Dans cette perspective, les chimistes ont classé la nomenclature chimique selon différents types de systèmes (Leigh et al., 2001)[67].

1.1.4.1 Nomenclature binaire

La nomenclature binaire permet de nommer des composés dont la structure est peu ou pas renseignée, mais dont nous connaissons au moins les constituants électropositifs et électronégatifs. De ce fait, la dénomination est formée du nom de chaque constituant du composé chimique en ajoutant parfois des préfixes et des suffixes multiplicatifs. Nous rencontrons la nomenclature binaire généralement en chimie inorganique. Voici un exemple :

[65] Principes de nomenclature de la chimie. Introduction aux recommandations de l'IUPAC, p. 26
[66] Idem
[67] Principes de nomenclature de la chimie. Introduction aux recommandations de l'IUPAC, p. 26

le sel binaire ‘chlorure de sodium NaCl’ est composé de l’ion de ‘chlorure Cl^-’ et de l’ion de ‘sodium Na^+’.

1.1.4.2 Nomenclature de coordination

La nomenclature de coordination est un système additif créé pour nommer les composés de coordination qui sont constitués d’un atome central entouré de ligands. De ce fait, la dénomination est formée de la combinaison des noms des différents constituants du composé chimique. Cette nomenclature est aussi utilisée pour des oxacides et leurs anions. Citons par exemple le complexe de coordination ‘pentacyanocarbonylferrate (II) $[Fe(CN)_5CO]^{3-}$’ : il est formé d’un atome de fer Fe^{2+}, de cinq ligands anioniques ‘cyano CN^-’ et un ligand neutre ‘carbonyl CO’ ; la charge de l’atome central est indiquée en chiffres romains ‘(II)’ et puisque la charge totale du complexe est négative, il s’agit alors d’un anion complexe, exprimé par le suffixe ‘ate’ et attribué à l’atome central ‘ferrate’.

1.1.4.3 Nomenclature substitutive

La nomenclature substitutive est la principale nomenclature utilisée en chimie organique. La dénomination consiste à la modification par substitution d’une molécule fondamentale formelle, généralement un hydrure. Cette nomenclature se base sur l'utilisation des groupes fonctionnels (préfixes et/ou suffixe) que l'on cite à l'aide d'indices par rapport à un hydrure en suivante les règles suivantes :

- Déterminer la structure de base (chaîne ou cycle)
- Déterminer la fonction principale (suffixe), en suivant l'ordre de priorité
- Nommer les substituants
- Numéroter la chaîne
- Assembler les noms des substituants selon l’ordre alphabétique

Par exemple, ‘acide hydroxyéthanoïque $C_2H_4O_3$’ :

- Cette molécule est une chaîne qui contient deux atomes de carbone : -éthan-
- Il y a deux groupes fonctionnels dans cette molécule : un acide carboxylique et un alcool. L’acide carboxylique est prioritaire devant l’alcool ; il est alors la fonction principale : acide ..... -oïque
- L’alcool est donc considéré comme un constituant : hydroxy-

Ainsi, les différents groupes fonctionnels sont classés selon l'ordre de priorité permettant de déterminer la fonction principale. Elle est désignée par le suffixe correspondant. Tous les autres groupes sont désignés par des préfixes. Mais, certains groupes caractéristiques doivent toujours être cités comme préfixes alors que d'autres peuvent être utilisés comme préfixes ou comme suffixes[68].

D'autres types de systèmes de nomenclature sont également utilisés bien qu'ils sont moins importants, tels que la nomenclature acide, la nomenclature de remplacement, la nomenclature de classe fonctionnelle, la nomenclature soustractive, la nomenclature des composés organométalliques, la nomenclature des polymères…

Des préfixes et des suffixes sont employés dans la nomenclature chimique ainsi qu'en chimie dans les noms de processus, de matières, de caractères des substances formant ainsi des termes. En effet, la normalisation des préfixes et des suffixes en chimie a été initiée par les chimistes de Morveau, Lavoisier, Berthollet et de Fourcroy en 1787 et est complétée par des chercheurs comme Darmesteter (1877)[69] et plus récemment comme Kocourek (1982)[70] et Bonnet (2001)[71] dans leurs travaux portant notamment sur les préfixes et les suffixes dans le discours scientifique. Nous analysons plus en détail la formation de ces préfixes et ces suffixes dans le chapitre suivant (cf. 2.2.2.1.1 Dérivation).

1.1.5 Définition

La définition peut-être considérée comme l'élément fondamental dans un travail terminologique :

> *"Au plan notionnel, pour qu'un nom ait droit au titre de terme, il faut qu'il puisse, en tant qu'élément d'un ensemble (une terminologie), être distingué de tout autre. Le seul moyen pour exprimer ce système de distinctions réciproques est l'opération dite définition. » (Rey, 1979)*[72]

Cependant, elle est le centre du problème en terminologie, ce qui implique que ce n'est pas toujours évident. Cette définition terminologique est différente de la définition lexicographique :

[68] http://www.chemexper.com/misc/iupac/
[69] De la création actuelle de mots nouveaux dans la langue française et des lois qui la régissent
[70] La langue française de la technique et de la science. Vers une linguistique de la langue savante
[71] La construction d'une langue savante en Europe du Ve au XIXe siècle : le latin et le grec dans les sciences
[72] La terminologie : noms et notions, p. 23

> *« […] concerne seulement les signes d'une langue : elle explicite des signifiés tout en essayant de distinguer, non des concepts et des classes de choses, mais des sens et des classes d'usages (d'emplois) des signes." (Rey, 1979)*[73]

Les différences entre la définition lexicographique générale et la définition terminologique (du vocabulaire scientifique et technique) se trouvent souvent dans le choix des incluants et des différences spécifiques.

La définition terminologique, à la différence du développement encyclopédique « s'arrête quand elle a donné toutes les informations permettant de situer et de différencier un concept à l'intérieur d'un système conceptuel » (De Bessé 1990)[74]. Elle consiste à :

> *« Déterminer l'ensemble des caractères entrants dans la compréhension d'un concept. Le résultat de cette opération est une proposition énonçant une équivalence entre un terme, le défini, et l'ensemble des caractères qui le définissent » (De Bessé, 1990)*[75]

La définition consiste alors à déterminer l'ensemble des caractères entrant dans la compréhension d'un concept, puisqu'elle est donc « un système de concepts (exprimés sous la forme de caractères) qui décrit un concept en le plaçant dans le système de concepts du domaine considéré » (Depecker, 2002)[76]. Par conséquent, la définition terminologique est « un compromis entre définition lexicographique et description encyclopédique, destiné à améliorer l'usage des noms pour leur permettre de fonctionner comme termes » (Rey, 1979)[77]. De ce fait, la définition terminologique doit indiquer les traits de substance de l'unité référentielle : la classe d'objets à laquelle elle appartient, ses fonctions principales, ses traits pertinents, afin d'être adaptée à l'opération de nomination.

Il existe plusieurs types de définition : la définition par compréhension, la définition par extension et la définition par contexte ; la première énumère les caractères d'une notion donnée en partant du terme général le plus proche appelé le générique, et en dressant la liste de tous les caractères restrictifs qui la distingue d'autres notions au même niveau d'abstraction. Quant à la seconde, elle énumère toutes les espèces d'un générique au même niveau d'abstraction, ou toutes les parties d'un tout. Pour la dernière, il s'agit de l'occurrence du terme à définir dans une phrase.

Les critères essentiels d'une définition sont :

---

[73] La terminologie : noms et notions, p. 23
[74] La définition terminologique
[75] Idem
[76] Entre signe et concept : éléments de terminologie générale
[77] La terminologie : noms et notions, p. 43

- *« Fixer la relation entre le terme et le concept lors de sa mise au point,*
- *Identifier un terme en vérifiant qu'il lui correspond une définition qui lui soit propre,*
- *Expliquer la signification d'un concept pour les utilisateurs de banques de données terminologiques comme les traducteurs et les spécialistes du domaine et éventuellement les hommes de loi. » (Lelubre, 1992)*[78]

Pour notre travail, la définition en terminologie doit être précise, claire et succincte, en mettant en évidence les traits significatifs propres au terme traité, afin de délimiter les concepts de la chimie ; il s'agit du moyen le plus important pour la détermination des termes, éléments appartenant à un système terminologique.

1.1.6 Fiche terminologique

La fiche terminologique représente « *l'objet sur lequel portent consécutivement l'activité du terminographe et l'activité du terminologue* » (Gouadec, 1990)[79]. Il s'agit d'un des outils utilisés par le terminologue et/ou le terminographe, qui constitue un support sur lequel sont consignées les données terminologiques et terminographiques. Une fiche terminologique contient différentes informations inhérentes aux termes étudiés (linguistique, encyclopédique, indexation…). Voici les informations que nous indiquons dans nos fiches terminologiques :

- Le terme : c'est l'entrée de la fiche terminologique ; présentée sous sa forme lemmatisée : noms au singulier, adjectifs au masculin, verbes à la troisième personne du singulier masculin de l'accompli actif.
- L'équivalent : pour chaque terme arabe, nous avons donné son équivalent en français. Pour trouver l'équivalent d'un terme en français, il faut chercher dans la littérature de la chimie ou le cas échéant dans les dictionnaires ; en ce qui nous concerne, nous avons proposé les termes du domaine mais nous avons vérifié nos connaissances à chaque fois pour confirmer.
- La classe : chaque terme est étiqueté par une classe et/ou une sous-classe.
- La catégorie lexicale et grammaticale : chaque terme comporte des informations sur sa catégorie lexicale (nom, adjectif, verbe, adverbe) et la catégorie grammaticale (genre, nombre).

[78] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 53
[79] Terminologie. Constitution des données, p. 37

- La définition : chaque terme est décrit par les traits de substance de l'unité référentielle, c'est-à-dire, la classe d'objets à laquelle elle appartient, ses fonctions principales, ses traits pertinents.
- Le contexte : chaque terme est accompagné de son contexte ; il s'agit de l'énoncé qui entoure le terme, exprimant une idée complète, avec la référence du corpus et/ou du chapitre afin d'attester des termes.
- Les variantes terminologiques : dès le début de la lecture de notre corpus, nous avons remarqué l'existence de plusieurs formes pour exprimer une seule notion ; nous considérons ces variantes comme des synonymes.

1.2 Terminologie, domaine interdisciplinaire

La terminologie est étroitement liée à d'autres branches de la linguistique :

> *« Par son aspect linguistique, la terminologie touche à la sémantique lexicale, ou lexicologie, car les termes sont des faits de langue au même titre que les autres unités lexicales ; ils ont cependant leurs caractères propres, qui permettent de distinguer le terme et le non-terme (...)*
>
> *La terminologie touche également à la syntaxe en ce que les unités terminologiques sont souvent syntagmatiques et peuvent atteindre un degré de complexité considérable ; elle touche enfin à la sémiologie en ce sens que les termes sont de véritables signes dans lesquels le signifie a prépondérance sur le signifiant. » (Rondeau, 1991)*[80]

La terminologie est « une branche de la linguistique, à côté d'autres disciplines, comme la lexicologie et la lexicographie » (Lelubre, 1992)[81] et peut être aussi considérée comme un carrefour interdisciplinaire :

> *« [...] Ainsi, elle (la langue) emprunte des éléments et des concepts à la morphologie, à la lexicologie et à la sémantique. Toutefois, bien que la langue soit le fondement même de la terminologie, il serait difficile de la considérer comme une autre branche de la linguistique, que ce soit la phonologie, la morphologie, la lexicologie ou la syntaxe. La terminologie est un domaine interdisciplinaire dont les mots spécialisés de la langue naturelle constituent l'objet d'étude premier. En plus de la linguistique, d'autres disciplines scientifiques contribuent à la terminologie ». (Cabré, 1998)*[82]

Étant la science qui étudie les termes, la terminologie se voit intervenir dans d'autres domaines de la linguistique, notamment la lexicologie ; mais, elle intervient également dans d'autres

[80] Introduction à la terminologie
[81] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation
[82] La terminologie - Théorie, méthode et applications

domaines littéraires, comme la sociologie, ainsi que le domaine de l'informatique lors du traitement des termes, comme la linguistique de corpus et l'ontologie.

1.2.1 Terminologie et lexicologie

Certains considèrent la terminologie comme une branche de la linguistique, d'autres la voient comme « une variante de la lexicologie (ou étude du lexique) à laquelle elle emprunte ses méthodes et ses analyses pour les appliquer à un objet constitué par l'ensemble des terminologies » (Gouadec, 1990)[83]. En effet, l'objet d'étude est différent ; la lexicologie traite les mots de la langue commune alors que la terminologie traite les termes de la langue de spécialité :

> *« La lexicologie a pour matière le lexique, ensemble dont les éléments sont des lexies ou unités lexicales de la langue commune - c'est-à-dire qui ne sont pas relatifs à un domaine spécialisé - tandis que la terminologie a pour matière des systèmes terminologiques, ensembles dont les éléments sont les termes ou encore les unités terminologiques ». (Lelubre, 1992)*[84]

Le lexicologue utilise le lexique général pour son étude alors que le terminologue se limite au vocabulaire de spécialité (Guilbert, 1973)[85]. De plus, la méthodologie n'est pas le même ; « la terminologie part du concept, la lexicologie de la désignation » (Cabré, 1998)[86]. En effet, la terminologie adopte une démarche dite 'onomasiologique' qui consiste à trouver le ou les termes utilisés (attestés) ou possibles (néologismes créés) qui désignent un concept connu. Quant à la lexicologie, elle adopte une démarche dite 'sémasiologique' qui consiste à partir d'un mot connu pour en déterminer la signification.

Pour notre étude terminologique de la chimie en arabe, nous adoptons les deux démarches ; nous partirons d'abord de l'approche lexicologique, soit la démarche sémasiologique afin de repérer les termes du corpus, puis nous suivrons l'approche terminologique, soit la démarche onomasiologique afin de déterminer quelle(s) dénomination(s) représente(nt) la notion étudiée.

1.2.2 Terminologie et sociologie

La terminologie, comme la linguistique, possède des marques d'usage, des « marques *sociolinguistiques : jargons d'atelier, registre technico*-scientifique, registre commercial »

[83] Terminologie. Constitution des données, p. 28
[84] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation
[85] La spécificité du terme scientifique et technique, p. 6
[86] La terminologie - Théorie, méthode et applications

(Dubuc, 2002)[87] ; appelée socioterminologie, c'est la science étudiant « *l'information scientifique et les termes en relation avec la société, l'action et, plus particulièrement, avec l'identité sociale dans la p*luralité de ses manifestations » (Abi Ghanem-Chadarevian, 2016)[88] c'est-à-dire, que la socioterminologie fait la relation entre le discours et la pratique (Barna, 2014)[89]. Cette approche s'est développée progressivement à partir des années 1980 et représente un moyen de récupérer la dimension sociale de la terminologie (Gaudin, 1993)[90], qui est peut-être considérée comme partie intégrante de la théorie terminologique (Gambier, 1991)[91]. Mais la socioterminologie a des limites et cela doit être pris en considération avec modération dans la normalisation des termes :

> *« Les travaux de terminologie visent alors la mise au point, la normalisation, la diffusion et l'implantation des vocabulaires de spécialité dans la langue nationale. Lorsque la langue est commune à plusieurs pays, par exemple le français, l'anglais, le basque, le néerlandais, on constate que des faits de culture ou de procédure introduisent nécessairement une certaine variation dans les terminologies. De ce fait, la normalisation des terminologies implique un certain compromis entre deux obligations : introduire des variantes mais ne pas s'éloigner inutilement de l'usage commun à tous les spécialistes de la même langue. » (Corbeil, 1999)*[92]

Dans le domaine de la chimie en arabe, des marques socioterminologiques peuvent être identifiées telles que la catégorie sociale (étudiant, professeur, chercheur), la forme de discours (texte écrit, communication orale, cours magistraux, travaux pratiques…), mais aussi l'espace géographique (pays du Maghreb, du Machrek et du Golfe). De ce fait, nous ne pouvons pas ignorer les marques socioterminologiques pour notre étude terminologique.

1.2.3 Terminologie et linguistique de corpus

Suite à l'explosion des corpus numériques et face aux besoins des entreprises qui doivent gérer une documentation considérable, les corpus sont devenus le principal matériau de la linguistique, donnant naissance à la linguistique de corpus, ainsi que de la terminologie. La terminologie basée sur corpus, en opposition à la vision wüsterienne (Wüster, 1981)[93],

[87] Manuel pratique de terminologie
[88] Socioterminologie et interactions langagières en arabe
[89] Divergences et convergences dans la terminologie médicale vétérinaire pour les vertébrés domestiques entre le roumain et le français
[90] Pour une socioterminologie : des problèmes sémantiques aux pratiques institutionnelles
[91] Travail et vocabulaire spécialisé : prolégomènes à une socioterminologie
[92] La terminologie : une discipline au service d'objectifs multiples
[93] L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses

permet d'identifier les termes et leurs relations, formant ainsi un réseau terminologique (Condamines, 2005)[94]. Cette nouvelle approche, appelée terminologie textuelle (Slodzian, 2000)[95], possède une relation privilégiée avec l'informatique, puisqu'elle a besoin de plusieurs logiciels pour son exploitation tels que le traitement de texte, la base de données, l'ontologie…, impliquant des compétences informatiques inhérentes à la formation du terminologue.

Pour notre travail, nous avons constitué un corpus à partir duquel nous avons étudié la terminologie de la chimie en arabe à l'aide d'outils informatiques (cf. 3.2 Démarche de la recherche de corpus et 4.2 Analyse des formes).

1.2.4 Terminologie et ontologie

La terminologie et l'ontologie possèdent un rapport privilégié avec le concept, puisque l'essence du terme est concentrée dans le concept, une unité de connaissance ; d'une part, la terminologie étudie les termes et d'autre part, l'ontologie, étudie les concepts. Elles utilisent toutes les deux l'approche de l'objet et l'approche conceptuelle, les relations entre les concepts et les propriétés de concepts, ainsi que l'analyse de la désignation et les méthodes classificatoires (Depecker, 2007)[96]. De ce fait, une nouvelle approche s'est constituée sous le nom d'ontoterminologie :

> *« L'ontoterminologie, terminologie dont le système notionnel est une ontologie formelle, insiste d'une part sur l'importance des principes épistémologiques qui président à la conceptualisation du domaine – c'est l'ontologie dans sa définition première –, et d'autre part sur la nécessité d'une approche scientifique de la terminologie où l'expert joue un rôle fondamental – c'est l'ontologie dans ses définitions plus récentes » (Roche, 2008)[97].*

L'ontoterminologie a l'avantage de fournir la qualité des produits terminologiques, puisque les terminologies sont structurées en schémas conceptuels, à l'aide des principes et des méthodes disponibles pour la structuration des ontologies.

Pour notre travail, nous avons établi l'ébauche d'une ontoterminologie de la chimie en arabe.

[94] Linguistique de corpus et terminologie
[95] L'émergence d'une terminologie textuelle
[96] Linguistique, terminologie et ontologie
[97] Quelle terminologie pour les sociétés de l'information ?

1.3 Évolution de la terminologie

En retraçant brièvement l'historique de la terminologie, nous apercevons que depuis longtemps nommer les objets est une pratique usuelle ; mais, elle n'était pas encore appelée terminologie. Cela concerne toutes les langues, notamment l'arabe ; les linguistes et grammairiens arabes de l'époque classique ont donné une place très importante à la lexicologie et à la terminologie (Jaber, 2012)[98].

Suite au développement des sciences et des techniques ainsi que la circulation de l'information à l'échelle mondiale, la terminologie permet de classer et d'organiser les domaines de connaissance notamment dans les sciences grâce aux travaux du naturaliste Linné pour la botanique (Systema naturæ, 1735), Lamarck pour la biologie (Flore française, 1778), et des chimistes Lavoisier, Berthollet, Guyton-Morveau et de Fourcroy, pour la systématisation de la chimie (Méthode de nomenclature chimique, 1787). Vers la fin du 19ème siècle, la terminologie n'est plus restreinte aux sciences et commence à s'appliquer aux techniques.

Au début du 20ème siècle, la terminologie moderne prend son essor et se développe ; quatre étapes se distinguent : les origines (de 1930 à 1960), caractérisées par la création des méthodes de travail terminologique, l'étape de structuration (de 1960 à 1975), marquée par le développement de l'informatique, des techniques documentaires et l'apparition de la normalisation des langues, la période d'éclatement (de 1975 à 1985), caractérisée par l'apparition de la micro-informatique et de l'aménagement linguistique et l'étape des larges horizons (depuis 1985), où les méthodes informatiques, les instruments de travail terminologique et les ressources se développent, tout comme les industries de la langue et la création des réseaux internationaux (Cabré, 1998)[99].

À l'origine, des recherches sur la terminologie sont effectuées par trois écoles de pensée classique qui deviennent les trois grands centres d'activité terminologique, à savoir, Vienne, Moscou et Prague. De ces écoles est sorti Wüster, le père de la Théorie Générale de la Terminologie (TGT), « *à qui nous devons d'immenses efforts pour la reconnaissance* disciplinaire et politique de la terminologie » (Cabré, 2007)[100] ; sa thèse intitulée Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik (Normalisation internationale de la langue électrotechnique) est considérée comme le premier travail théorique sur la

[98] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours
[99] La terminologie. Théorie, méthode et application, p. 27
[100] La terminologie, une discipline en évolution : le passé, le présent et quelques éléments prospectifs

terminologie et son dictionnaire de la machine-outil vérifie en pratique ses théories. Sous son influence, la Fédération internationale des associations de normalisation (ISA), voit le jour.

Dans les années 1960 à 1975, la terminologie s'étend vers d'autres centres d'influence comme le Canada, la Belgique et la Suisse (pays officiellement bilingues ou plurilingues). La France s'intéresse aussi à la terminologie et « *devient alors l'un des enjeux majeurs d'une politique linguistique qui ne cesse d'évoluer à mesure que la francophonie se construit sur la scène* internationale » (Depecker, 2005)[101] avec notamment Gilbert, Rey, et De Bessé. La terminologie fait également son entrée dans les grandes organisations internationales et les organismes à vocation plus technique, comme la création par Wüster en 1971 d'INFOTERM (International Information Centre for Terminology). On assiste aussi à l'apparition des premières grandes banques de données terminologiques et au début de l'organisation de la coopération terminologique mondiale.

À partir des années 70, on assiste à un aménagement linguistique, c'est-à-dire, une politique conduite par un État ou une organisation internationale à propos d'une ou plusieurs langues parlées dans les territoires relevant de sa souveraineté, comme le cas de du Canada (Corbeil, 2007)[102].

Depuis 1985, la terminologie a fait son entrée dans le monde universitaire en tant que discipline autonome. Par ailleurs, elle s'est avérée utile pour les besoins de commercialisation des biens et des services dans la société mondiale ainsi que dans le développement des outils d'ingénierie linguistique et du traitement du langage naturel. Suite à l'explosion des corpus numériques, aux développements de la linguistique de corpus et de l'ingénierie linguistique, une redéfinition de la terminologie et une révision de sa théorie a été nécessaire pour répondre aux nouveaux besoins s'orientant à présent vers une terminologie descriptive.

Après avoir connu son premier essor au début du siècle, notamment avec la théorie wüstérienne, la terminologie a été de nouveau au centre des occupations recherchant constamment de nouveaux modèles théoriques.

[101] Présentation, p. 3

[102] Le rôle de la terminologie en aménagement linguistique : genèse et description de l'approche québécoise, p. 94

1.3.1 Fondements théoriques de la terminologie

Pendant les années 1930, les fondements théoriques de la terminologie ont été développés par plusieurs écoles de pensée : l'école autrichienne, l'école tchèque et l'école soviétique.

1.3.1.1 Ecole autrichienne

Fondé par Wüster, l'école autrichienne est basée sur sa théorie, la Théorie Générale de la Terminologie (TGT) ; elle est la première à rendre compte des termes, de leurs fonctions et de leurs relations. Cette théorie est basée sur les éléments suivants :

- *L'importance du concept, des relations entre les concepts et de la relation entre le terme et le concept,*
- *La nécessaire démarche onomasiologique,*
- *L'analyse située au niveau du terme,*
- *L'approche essentiellement synchronique,*
- *La définition de la terminologie comme étant un domaine interdisciplinaire. (Cabré, 2007)*[103]

La terminologie est soit considérée comme une discipline autonome à caractère interdisciplinaire, soit centrée sur la philosophie (organisation des notions et des connaissances), soit appartenant aux langues de spécialité. Cependant, cette théorie a des limites tant par sa conception idéaliste que par l'impasse qu'elle fait sur certains principes de la terminologie :

- La phraséologie en réduisant la terminologie à l'analyse du terme seulement (« combinatoire des termes » (Cabré, 2007)[104]),
- Les aspects pragmatiques, sémantiques et sociologiques de la terminologie en minimisant la distinction entre les termes et les unités lexicales de langue générale « conception linguistique » (Cabré, 2007)[105]),
- Les autres applications de la terminologie comme la traduction, la rédaction technique, la communication en général… en se concentrant sur la documentation et la normalisation.

[103] La terminologie, une discipline en évolution : le passé, le présent et quelques éléments prospectifs
[104] Idem
[105] Idem

La théorie wüsterienne est donc considérée comme une « approche prescriptive, conceptuelle et onomasiologique » en opposition aux nouvelles théories qui sont considérées comme une « *approche descriptive, linguistique et sémasiologique, basée sur l'étude de textes spécialisés* » (Bertels, 2009)[106]. Mais cette théorie a influencé et influence toujours les travaux de la terminologie et de l'ontologie jusqu'à nos jours (Budin, 2007)[107].

1.3.1.2 Ecole soviétique

Suite à la traduction de la théorie wüsterienne en langue russe, Lotte fonde la seconde école terminologique ; ce dernier observe l'importance du discours et du contexte, au sein desquels le terme reçoit sa valeur ainsi que la reconnaissance du caractère essentiellement polysémique du vocabulaire. C'est « une conception beaucoup plus linguistique de la terminologie ; une influence moins profonde de la philosophie que chez les chercheurs germano-autrichiens » (Rondeau, 1991)[108]. Les sujets les plus importants traités par l'école de Moscou étaient : la notion, le découpage du terme, la différence entre terminologie et nomenclature (cf. 1.1.4 Nomenclature) et la place de la terminologie en tant que discipline, puisque « c'est en U. R. S. S. qu'est née la terminologie comme discipline scientifique, au moment où s'élaboraient en Autriche des méthodes de traitement des données terminologiques » (Rondeau, 1991)[109].

1.3.1.3 Ecole tchèque

À la base des travaux théoriques de l'école de Prague se trouve la conception de la langue comme un système possédant une fonction et une finalité, celle d'exprimer et de communiquer. Elle ouvre les portes à la linguistique structurelle ou fonctionnelle ; le structuralisme est devenu la base théorique de cette école, considérant les langues de spécialité comme un style fonctionnel de la langue, un style professionnel. La standardisation terminologique internationale est centrale dans les travaux de cette école.

1.3.2 Les nouvelles théories de la terminologie

Plusieurs théories ont vu le jour et peuvent se résumer en quatre points de vue :

[106] Etudier la sémantique des termes techniques : des théories à la pratique

[107] L'apport de la philosophie autrichienne au développement de la théorie de la terminologie : ontologie, théories de la connaissance et de l'objet

[108] Introduction à la terminologie

[109] Idem, p. 7

- *« La terminologie comme un besoin social,*
- *La terminologie comme une pratique destinée à satisfaire les besoins engendrés par cette exigence ;*
- *La terminologie comme une application ou ensemble de ressources générées par la pratique,*
- *La terminologie comme un domaine du savoir ». (Cabré, 2007[110])*

1.3.2.1 Théorie Communicative de la Terminologie

La Théorie Communicative de la Terminologie (TCT) peut être définie comme « une théorie linguistique des unités terminologiques à composante cognitive et communicative » (Cabré, 2007)[111]. Voici quelques principes fondamentaux de cette théorie :

- L'unité terminologique est l'élément central de la terminologie,
- Les unités terminologiques ont des propriétés linguistiques, cognitives et socio-communicatives,
- L'analyse des unités terminologiques doit se faire à partir du corpus.

1.3.2.2 Terminologie textuelle

La terminologie textuelle peut être définie comme une « analyse de corpus textuels » (Bourigault et al., 1999)[112]. Les fondamentaux de cette théorie sont :

- Le texte est le point de départ de la description à construire. On va du texte vers le terme. Les bases théoriques de la terminologie doivent être ancrées dans une linguistique textuelle.
- Le terme est le produit d'un travail d'analyse, réalisé à partir du corpus en identifiant les structures lexicales qui présentent des caractéristiques à la fois spécifiques et stables et à partir de l'application en vérifiant leur utilité.

[110] La terminologie, une discipline en évolution : le passé, le présent et quelques éléments prospectifs, p. 80
[111] Idem, p. 96
[112] Pour une terminologie textuelle

D'autres approches ont fait l'objet de travaux comme la terminologie sociocognitive de Temmerman (2000)[113], la terminologie culturelle de Diki-Kidiri (2007)[114] ou encore la Frame-Based Terminology de Faber (2014)[115].

1.4 Terminologie scientifique

Notre intérêt se porte sur la notion de normalisation terminologique dans le contexte d'une langue de spécialité en arabe. Dans un premier temps, nous décrivons les caractéristiques des langues de spécialité. Dans un second temps, nous présentons la normalisation terminologique, à travers ses organismes et ceux du monde arabe avec ses disparités.

1.4.1 Langue de spécialité

La problématique de la langue de spécialité n'est pas une nouveauté et remonte au siècle dernier. En effet, avec les progrès techniques, le vocabulaire évolue, s'enrichit et s'élargit demandant alors aux spécialistes de la langue la description de ces nouveaux concepts pour répondre aux besoins communicatifs des gens de science. La terminologie, comme d'autres domaines, est au cœur des recherches de la langue de spécialité. La terminologie permet alors la description sémantique des termes et la description des mécanismes de leur formation mais aussi la systématisation du vocabulaire dans la langue de spécialité. Cette étude de la langue de spécialité consiste à déterminer son comportement en analysant le vocabulaire et la syntaxe utilisés. Ces deux caractéristiques varient selon la situation de communication étudiée et spécifient la langue de spécialité :

> *« 'Expression générique pour désigner les langues utilisées dans des situations de communication (orales ou écrites) qui impliquent la transmission d'une information relevant d'un champ d'expérience particulier ». (Galisson et al., 1976)*[116]

Cette définition met en évidence la spécificité de la langue de spécialité qui est appelée aussi « langue spécialisée, langues spécialisées, technolectes, langages de spécialité et langages spécialisés » (Vicente, 2009)[117]. Et afin de saisir la langue toute entière dans sa diversité, une

[113] Une théorie réaliste de la terminologie : le sociocognitivisme
[114] Éléments de terminologie culturelle
[115] Frames as a Framework for Terminology
[116] Dictionnaire de didactique des langues, p. 511
[117] La didactique du concept de langue spécialisée : vers une approche traductologique de la question

division en sous-systèmes plus homogènes peut-être utile (Kocourek, 1991)[118]. Ainsi, la langue de spécialité serait un sous-ensemble de la langue, fondamentalement pragmatique, de la langue dans son sens global (Cabré, 1998)[119]. Elle serait alors « une langue naturelle considérée en tant que vecteur de connaissances spécialisées ». (Lerat, 1995)[120]

Au vu de ces définitions, nous pourrions dire qu'un texte est composé d'un système d'expressions, constitué par une première 'couche', la langue toute entière, sur laquelle vient s'appuyer une seconde 'couche', la langue de spécialité.

La langue de spécialité intéresse aussi bien les linguistes (interprètes, traducteurs, lexicologues, lexicographes) que les scientifiques (spécialistes et auteurs techniques, rédacteurs des revues spécialisées) (Kocourek, 1982)[121]. En effet, elle fait intervenir autant le spécialiste que le vocabulaire spécialisé, variant les types de communication que ce soit entre spécialistes ou entre spécialiste et non-spécialiste. La communication entre spécialistes constitue la langue de spécialité. Compris par ses pairs, le spécialiste « condense » au plus haut niveau son vocabulaire en utilisant des termes du domaine sous la forme de symboles ou de formules ou encore de schémas et pourquoi pas des éléments de code non verbaux. Cela lui permet d'éliminer toutes les ambiguïtés de la langue courante et d'atteindre le plus haut degré de précision et de synthèse dans sa description. Quant à la communication entre spécialiste et non-spécialiste, elle permet d'expliquer les concepts d'un domaine sans employer un vocabulaire spécialisé. Ainsi, le spécialiste doit utiliser des mots et des expressions de la langue courante pour être compris. Il s'agit là de la vulgarisation, comme dans les articles de revues non spécialisées qui traitent de sujets techniques.

Que la communication soit entre spécialistes ou entre spécialiste et non-spécialiste, la langue de spécialité possède des caractéristiques.

1.4.1.1 Variante de la langue générale

La langue de spécialité n'est ni autonome, ni complète et elle ne peut exister qu'avec la langue générale. En effet, la grammaire et la syntaxe restent plus au moins inchangées ; seuls les termes utilisés, c'est-à-dire, le vocabulaire spécialisé, sont spécifiques à la langue de spécialité. Il s'agirait d'une simple variante linguistique de la langue générale (Cabré, 1998)[122],

[118] La langue française de la technique et de la science. Vers une linguistique de la langue savante, p. 13
[119] La terminologie - Théorie, méthode et applications, p. 119
[120] Les langues spécialisées, p. 20
[121] La Langue française de la technique et de la science
[122] La terminologie - Théorie, méthode et applications, p. 119

composée d'éléments concrets qui s'inscrivent dans un domaine de vie, comme par exemple la chimie. De plus, la langue de spécialité n'est pas une langue littéraire, permettant par exemple la recherche de synonymes, mais elle se caractérise par d'autres formes ; variée, elle oscille entre courante et soutenue où elle peut être plus ou moins formelle selon son lieu de présence et est fortement liée à la culture de la société qui l'utilise (Kocourek, 1982)[123].

Le vocabulaire de de la chimie, comme pour d'autres domaines de spécialité, possède les mêmes mécanismes de création de la langue générale et ses règles de grammaire restent valables. En effet, le but est de nommer de nouvelles découvertes, impliquant au passage l'accroissement de son vocabulaire, afin de répondre aux besoins de la langue de spécialité. Ainsi, la langue de spécialité crée sa propre base lexicale et terminologique dans laquelle s'inscrivent aussi bien les mots employés dans la langue générale que des termes du domaine, comme des processus (filtration), des noms d'instruments (bécher), des substances (chlorure de sodium) (Peraldi, 2012)[124].

1.4.1.2 Fonction communicative

La langue de spécialité ne se limite pas à une compétence linguistique ; elle possède autant une compétence linguistique qu'une compétence spécialisée, notamment dans le cadre de son enseignement, longtemps limité à l'apprentissage de son vocabulaire spécialisé :

> *« La théorie didactique prégnante, à cette époque [1960], était que l'étudiant devait d'abord posséder la langue usuelle, puis compléter le vocabulaire courant par une terminologie technoscientifique appropriée ». (Gentilhomme, 2000)[125]*

Cette double compétence, linguistique et spécialisée, facilite la diffusion des langues de spécialités et son enseignement, puisque sa fonction principale est la communication :

> *« C'est sans doute la fonction communicative qui a un rôle fondamental en langue de spécialité. Nous supposons que, en langue de spécialité et surtout en langue de technique et de la science, la divergence entre l'intention de l'émetteur (de l'auteur) et la compréhension du récepteur (du destinataire) est réduite au minimum. Par conséquent, il sera possible de n'aborder les problèmes des participants qu'indirectement, par le truchement des concepts qui sous-tendent quelques fonctions classiques, à savoir les fonctions cognitives, monologique/dialogique, émotive, conative, métalinguistique et esthétique. » (Kocourek, 1982)[126]*

[123] La Langue française de la technique et de la science, p. 24
[124] L'anglais de spécialité en chimie organique : entre indétermination terminologique et multidimensionnalité
[125] Termes et textes mathématiques. Réflexions linguistiques non standard, p. 58
[126] La Langue française de la technique et de la science, p. 19

Cette perspective communicative de l'enseignement va jusqu'à créer un nouveau terme pour désigner la langue de spécialité 'Communication sur Objectifs Spécifiques (COS)', prenant en compte les paramètres complets de la situation de communication (déictiques personnels, temporels, spatiaux) aussi bien que des implications ayant un caractère interculturel, socioculturel et stratégique (Binon et al., 2000)[127].

Pour les textes scientifiques en général, pour la chimie, en particulier, la langue de spécialité présente des manifestations graphiques comportant des signes, comme les lettres grecques, les indices numériques et les symboles ; mais, il ne s'agit pas de « *limiter l'approche des langues* de spécialité à une lexicologie des racines grecques, latines et autres » (Lerat, 1995)[128].

Ainsi, la langue de spécialité dispose de moyens qui aident à connaître le monde, en facilitant cette découverte par l'expérience des spécialistes et des domaines déterminés par les recherches spécialisées.

D'autre part, la langue de spécialité se caractérise par une recherche de la monosémie et en terminologie, l'univocité notion-terme est primordiale ; mais inversement pour la langue générale, face à un degré de spécialisation moins élevé, la monosémie peut être moins respectée, et nous pouvons nous trouver face à des cas de synonymie, tout au moins partielle :

> *« La morphologie lexicale adéquate est celle qui conduit d'une base univoque à des formes plus complexes, mais également monosémiques. La langue générale est plus capricieuse, mais en langue spécialisée on observe des séries terminologiques homogènes sémantiquement mais fortement sujettes à l'allomorphie du fait du recours aux formants grecs anciens, latins et anglais ». (Lerat, 1995)*[129]

De plus, la langue de spécialité propose une fonction esthétique, par une clarté et une précision non négligeable :

> *« Bien que l'on puisse affirmer que la langue technique et scientifique tend à neutraliser l'aspect esthétique, il serait inexact de dire que l'aspect esthétique au sens large lui fait entièrement défaut. Ce serait négliger le fait qu'il existe en techniques et en sciences un certain idéal linguistique, par exemple la précision du contenu, la concision de la forme, l'élégance et la systématique de la terminologie employée, la réduction de la synonymie et de l'ambigüité, en bref toutes les propriétés idéales liées à l'intellectualité et à la particularité de cette langue. La littérature technique et*

[127] Les langue(s) de spécialité(s) : mythe ou réalité ? Lexicographie et langue(s) de spécialité(s), p. 40
[128] Les langues spécialisées
[129] Idem, p. 12

scientifique possède des caractères spécifiques qui complètent le portrait de la créativité linguistique, dont la littérature esthétique au sens étroit ne constitue qu'une partie ». (Kocourek, 1982)[130]

De ce fait, les langues de spécialité dovent être standardisées en mettant en place une normalisation terminologique.

1.4.2 Normalisation terminologique

Avec le développement des sciences et des technologies, les langues s'enrichissent face au besoin de dénomination ; les termes fleurissent de partout :

« Les terminologies naissent souvent dans le désordre, au hasard des besoins et des créations, selon les régions géographiques où elles s'implantent » (Dubuc, 2002)[131].

La prolifération de ces termes n'est pas compatible avec la nécessité de précision des discours scientifiques, et se voit confrontée à des enjeux normatifs urgents et stratégiques, afin de 'mettre de l'ordre'. En effet, la formation de nouveaux termes peut créer des cas de synonymie, de polysémie… qui sont autant d'ambiguïtés qui peuvent être à l'origine d'une perte d'information.

La normalisation terminologique est un processus complexe permettant de faciliter la communication spécialisée et de réduire au maximum ces ambiguïtés :

« Un processus complexe comprenant diverses opérations : l'unification des notions et des systèmes de notions, la réduction de l'homonymie, l'élimination de la synonymie, la stabilisation des dénominations » (Cabré 1998)[132].

Cette normalisation terminologique cherche à optimiser la communication spécialisée :

« Elle [la normalisation] enferme (...) les notions dans des étiquettes ou des dénominations qui permettront à une variété d'interlocuteurs de percevoir un message identique » (Rondeau, 1991)[133].

De ce fait, la normalisation terminologique est basée sur plusieurs critères fondamentaux (Rondeau, 1991)[134]. D'une part, elle doit prendre en compte des facteurs d'ordre sociolinguistique (usage établi, milieu d'implantation, les besoins des usagers, etc.) et certaines valeurs psycholinguistiques (esthétique, motivation, habitudes des locuteurs, résistance au

[130] La Langue française de la technique et de la science
[131] Manuel pratique de terminologie, p. 73
[132] La terminologie- Théorie, méthode et applications, p. 245
[133] Introduction à la terminologie, p. 93
[134] Idem

changement, etc.). D'autre part, elle doit améliorer la qualité de la communication spécialisée et doit être une référence stable pour les usagers, autant à un niveau national et international, à l'aide de coordinations et de collaborations avec les spécialistes des domaines d'une diffusion aussi large que possible. Ces critères fondamentaux devraient garantir la réussite de l'adaptation d'une terminologie normalisée. D'autres caractéristiques peuvent être ajoutées, telles que la motivation, la simplicité, la brièveté, le parallélisme par rapport à l'anglais, dans le cas d'un terme venant en remplacement d'un emprunt, la paradigmisation ou l'importance d'établir des séries de termes, le caractère imagé, la résonance non technocratique (Depecker, 1994)[135].

Cependant, il est parfois difficile d'implanter un terme normalisé sans prendre en compte les aspects sociolinguistiques et psycholinguistiques. En effet, lorsque le terme imposé par la norme va à l'encontre de l'usage, ses chances de succès sont assez faibles, voir quasiment nulles (Rondeau, 1991)[136]. Il s'agit d'une certaine résistance de la part des locuteurs face au remplacement d'un terme déjà existant et pouvant être très ancré dans l'usage ; il faudrait replacer ces termes en discours pour clarifier la communication plutôt que de l'opacifier (Depecker, 1994)[137]. Cela peut provenir notamment d'une distance trop importante entre les normalisateurs et les usagers, puisque la norme relevant du pouvoir, les usagers n'appartiennent pas à la hiérarchie des centres de décision, ce qui conduit au « dirigisme linguistique » (Rondeau, 1983)[138].

Pour la chimie, comme pour d'autres domaines, la normalisation terminologique est nécessaire. Pour cela, une négociation entre experts et usagers permettrait de ne pas imposer des modèles trop éloignés des pratiques langagières réelles mais « plutôt de convaincre en proposant des terminologies harmonisées » (Depecker 1994)[139] afin de satisfaire à l'objectif d'optimisation de la transmission de l'information scientifique et technique que se fixe la normalisation.

1.4.2.1 Organismes terminologiques

Les normes ISO sont considérées comme le guide du travail terminologique, « une base *de référence sur l'acquis en matière de terminologie, sur lequel s'accorde aujourd'hui la*

[135] L'aménagement terminologique : de l'usage à la décision, p. 12
[136] Introduction à la terminologie, p. 116
[137] L'aménagement terminologique : de l'usage à la décision, p. 13
[138] Introduction à la terminologie, p. 115
[139] L'aménagement terminologique : de l'usage à la décision, p. 13

communauté internationale » (Depecker, 2002)[140]. Mais, l'ISO n'est pas le seul organisme dont le but est d'unifier les méthodologies employées dans la création de nouveaux termes ainsi que de trouver la politique adéquate afin de rendre ces termes inventés facilement utilisables. D'autres organismes manifestent également leur intérêt pour la terminologie pendant les trente dernières années, l'activité terminologique devenant vraiment importante, ouvrant « de larges horizons de la terminologie moderne (depuis 1985) » (Cabré,1998)[141] et créant de nombreuses associations, réseaux et centres de terminologie ; l'Association Européenne de Terminologie a réalisé le répertoire 'Entités impliquées dans le travail terminologique trilingue français / anglais / espagnol', rassemblant des informations sur la structure et les activités de ces organismes, répertoire publié sur leur site internet (http://eaft-aet.net/fileadmin/files/Directory.pdf)[142]. Dans ce répertoire, nous remarquons l'absence d'entité issue du monde arabe alors qu'il y a 5 entités latino-américaines, 2 asiatiques et 2 nord-américaines.

Pour notre étude terminologique de la chimie, nous nous référons principalement aux normes terminologiques de l'ISO, à l'Union internationale de chimie pure et appliquée (UICPA) et au bureau de la traduction pour la normalisation terminologique.

1.4.2.1.1 Normes ISO en terminologie

L'ISO (Organisation internationale de normalisation) est une organisation internationale non gouvernementale, indépendante, dont les 162 membres sont les organismes nationaux de normalisation. Par ses membres, l'organisation réunit des experts qui mettent en commun leurs connaissances pour élaborer des normes internationales d'application volontaire, fondées sur le consensus, pertinentes pour le marché, soutenant l'innovation et apportant des solutions aux enjeux mondiaux.

Cet organisme élabore des normes sous forme de documents définissant les exigences, les spécifications, les lignes directrices ou des caractéristiques à utiliser systématiquement pour assurer l'aptitude à l'emploi des matériaux, produits, processus et services. Ces normes servent de base aux technologies que nous utilisons et en assurent la qualité voulue, en apportant des avantages réels et mesurables dans pratiquement tous les domaines imaginables. Plus de 21991 normes internationales ont été publiées ; 32 concernent les normes terminologiques (Barna,

[140] Entre signe et concept : éléments de terminologie générale
[141] La terminologie : théorie, méthode et applications, p. 27
[142] Entités impliquées dans le travail terminologique

2014)[143], notamment les normes ISO 704 ‘Travail terminologique -- Principes et méthodes’ (ISO 704, 2009)[144], ISO 860 ‘Travaux terminologiques – Harmonisation des concepts et des termes’ (ISO 860, 2007)[145], ISO 1087-1, ISO 1087-1, ‘Travaux terminologiques -- Vocabulaire -- Partie 1 : Théorie et application’ (ISO 10877-1, 2000)[146].

1.4.2.1.2 Union internationale de chimie pure et appliquée (UICPA)

L'Union internationale de chimie pure et appliquée (UICPA) de l’anglais International Union of Pure and Applied Chemistry (IUPAC) est l'autorité mondiale en matière de nomenclature et de terminologie chimiques, y compris la désignation de nouveaux éléments dans le tableau périodique, des méthodes normalisées de mesure et de nombreuses autres données évaluées de manière critique. Organisation scientifique, l'UICPA a été fondée en 1919 par des universitaires et des industriels qui partageaient un objectif commun : réunir une communauté chimique fragmentée mondialement pour l'avancement des sciences chimiques par la collaboration et le libre échange d'informations scientifiques. Ainsi, l'UICPA crée un langage commun et normalise les processus et les procédures à travers un système de projet formel. Cela a contribué à la diversité et à l'interdisciplinarité de la chimie en fournissant un langage commun pour la chimie et en préconisant l'échange d'informations scientifiques[147].

1.4.2.1.3 Bureau de la traduction. Normalisation terminologique.

Responsable de Termium, l’une des banques de terminologie les plus exhaustives au monde, le Bureau de la traduction du gouvernement du Canada est une référence en terminologie, en raison de la structure multilingue et multiculturelle de la société canadienne. Il offre un large éventail de services à des clients du gouvernement qui doivent communiquer dans plusieurs langues ou organiser des événements à l’intention de participants de langues différentes. Le Bureau est renommé pour la grande qualité de ses services, qu’il s’agisse de traduction, d’interprétation, de services linguistiques ou de terminologie. Il propose notamment un ouvrage condensé ‘Précis de terminologie’, permettant de mieux structurer la démarche terminologique, facilitant de façon encore plus soutenue la collaboration interorganisations et

[143] Divergences et convergences dans la terminologie médicale vétérinaire pour les vertébrés domestiques entre le roumain et le français
[144] ISO 704 : Principes et méthodes de la terminologie
[145] ISO 860 : Travaux terminologiques -- Harmonisation des concepts et des termes
[146] ISO 1087-1 : Travaux terminologiques -- Vocabulaire -- Partie 1 : Théorie et application
[147] https://iupac.org

visant l'excellence dans la gestion harmonisée des fonds terminologiques de provenances diverses, mais néanmoins complémentaires (Pavel et al., 2001)[148].

1.4.2.2 Terminologie dans le monde arabe

À l'échelle du monde arabe, il n'existe pas une instance reconnue en matière de création terminologique ; mais, ce sont soit des instances soit des organismes qui établissent les terminologies.

À la fin de la première moitié du 19éme siècle, suite au mouvement de la « نهضة = *nahḍa* = renaissance arabe », la création d'une académie de la langue arabe a été tentée à plusieurs reprises avant de voir le jour à Damas en 1918 (Hamzaoui, 1965)[149] ; s'ensuit d'autres académies, comme le Caire en 1932, l'Irak en 1947, la Jordanie en 1976… Ces académies de la langue arabe ont pour objectif commun de préserver l'intégralité de la langue arabe et de l'adapter aux besoins de la vie contemporaine, en réorganisant l'enseignement, en développant la langue, en créant de nouveaux dictionnaires et en l'enrichissant de termes techniques et scientifiques (Jaber, 2012)[150].

Entre-temps, en 1962, le Bureau de Coordination de l'Arabisation est créé et sera affilié par la suite en 1970 à l'ALECSO (Organisation Arabe pour l'Éducation, la Culture et les Sciences) ; son rôle consiste à améliorer les domaines de l'éducation, de la culture et de la science et la coordination entre les États arabes, notamment en effectuant des recherches en arabisation et en normalisation terminologique, afin de promouvoir la langue arabe pour suivre et participer au progrès scientifique[151].

D'autres organismes et d'autres organisations, comme AIDMO (Organisation Arabe pour le Développement Industriel et des Mines), organisation arabe spécialisée dans le domaine de l'industrie, des mines et de la normalisation, affiliée à la ligue des Etats arabes[152], ou encore SASO (Organisation Saoudienne de Normalisation, Métrologie et qualité), organisme de référence distingué dans tous les domaines de la normalisation et de la qualité à l'échelle nationale, régionale et internationale[153] se développent dans le monde arabe.

[148] Précis de terminologie
[149] L'Académie arabe de Damas et les problèmes de la modernisation de la langue arabe
[150] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours, p. 139
[151] http://www.alecso.org/newsite/
[152] http://www.aidmo.org/
[153] http://www.saso.gov.sa/

Mais, malgré les efforts de ces académies, ces organisations, ces organismes et ces individus, l'arabe scientifique et technique se retrouve sujet à des créations personnelles diverses et variées de la part de ses locuteurs, soit en accord avec les instances du pays, soit en accord avec le système international. En effet, dans la plupart des pays arabes, si l'enseignement scientifique et technique se fait en arabe pour l'enseignement secondaire, il est suivi dans l'enseignement supérieur généralement en français et/ou en anglais.

Nous pensons que l'absence de normalisation terminologique arabe vient d'un manque de consensus entre les différentes instances / organismes, d'une part, et spécialistes / professionnels d'autre part. Cette faible coordination institutionnelle régionale, c'est-à-dire, le manque de cohérence stratégique entre les pays et entre les spécialistes ne favorise pas l'unification et l'harmonisation de la langue arabe afin d'obtenir une seule référence pour les travaux scientifiques.

Cette langue arabe est une langue sacrée, savante, technique, administrative, médiatique... que le monde arabe partage depuis le VII^{ème} siècle. Elle transcende ses disparités de cultures linguistiques dialectales :

> *« Des disparités taxonomiques et sémantiques sont encore très courantes entre les régions (Machreq/Maghreb), dans les pays (villes et campagnes) et dans les structures (langue littéraire ou administrative/langue dialectale ou populaire) ». (Hudrisier et al., 2017)*[154]

L'importance du phénomène de variations régionales, variations plus ou moins importantes selon les domaines, est un défi pour les instances terminologiques arabes concernées ; il ne s'agit pas simplement d'une diglossie de l'arabe, « *opposant un arabe 'littéral' ou 'littéraire' (voire 'classique') à un arabe 'dialectal', le premier s'associerait à un usage écrit formel ou* académique, le second à une pratique orale quotidienne ou vernaculaire » (Dichy, 2017)[155], mais plutôt d'une polyglossie de l'arabe, due à la diversité des parlers et des arabes moyens (Dichy, 2010)[156].

Cette diversité culturelle n'est pas sans conséquences sur la terminologie et sa normalisation :

> *« L'empreinte culturelle de chaque communauté est bien présente dans la conceptualisation même de domaines fortement techniques, comme le montre par exemple la classification du domaine du*

[154] Normalisation de la langue et de l'écriture arabe : enjeux culturels régionaux et mondiaux

[155] Polyglossie de l'arabe et subsidiarité : au-delà des confusions entrainées par la notion de 'diglossie'

[156] La polyglossie de l'arabe, illustrée par deux corpus d'époques et de natures différentes : un échange radiophonique syrien et un conte des Mille et Une Nuits

système hormonal, apparemment universel, mais qui révèle des différences de conceptualisation fondamentales d'une langue à l'autre (Bagge, 1999)[157]

De plus, une relation de contact de langues où un ensemble exerce une certaine 'influence' linguistique sur l'autre affecte la langue arabe beaucoup plus que les autres langues (Bianchini et al., 2008)[158].

Pour le domaine de la chimie comme pour d'autres domaines scientifiques, il n'est pas question de polyglossie, qui ne concerne que la langue générale ; il s'agit plutôt d'influence étrangère. De ce fait, la terminologie scientifique arabe en général, la terminologie de la chimie en particulier, est tributaire des terminologies établies en anglais et/ou en français.

Dans ce chapitre, nous avons défini la terminologie, comme une discipline à part entière, à la frontière des sciences par sa nature interdisciplinaire. Pour le domaine de la chimie, le vocabulaire est composé de termes issus de la nomenclature chimique, normalisée par l'UICPA et ainsi que d'autres termes. L'ensemble de ces termes est tributaire des terminologies établies en anglais et/ou en français.

A présent, nous allons voir les moyens dont la langue arabe dispose pour la création et la formation des termes de la chimie.

[157] Analyse sémantique comparative des vocabulaires scientifiques anglais et français
[158] Les mots de l'eau : entre terminologie spécialisée et analyse

Chapitre 2 : Ressource de l'arabe pour la création et la formation des termes de la chimie

Dans ce chapitre, nous examinons les moyens dont la langue arabe dispose pour la création et la formation des termes de la chimie afin de les modéliser par des patrons morphosyntaxiques. Pour cela, nous présentons dans un premier temps les systèmes de la langue arabe, notamment les possibilités offertes par les systèmes de nomination et de communication à partir de l'approche de Roman (1990)[159]. Dans un second temps, nous expliquons les différents procédés linguistiques mis en œuvre pour la création des termes de la chimie en arabe à partir de l'approche de Lelubre (1992)[160].

2.1 Système de la langue arabe

La langue arabe est un système de sous-systèmes, interdépendants, composé de quatre sous-systèmes :

- Le sous-système phonologique
- Le sous-système syllabique
- Le sous-système de nomination
- Le sous-système de communication

Les deux premiers sous-systèmes constituent le matériau phonétique dont toute langue humaine dispose, contenant des consonnes (C) et des voyelles (V) pour le sous-système de phonèmes, et des syllabes (S) pour le sous-système de syllabes. Le sous-système de nomination est construit dans les langues sémitiques sur des combinaisons de consonnes (C) et sur des racines de consonnes (C). Le sous-système de communication est construit sur les voyelles (V) (Roman, 1990)[161].

[159] Grammaire de l'arabe
[160] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation
[161] La création lexicale en arabe, p. 21

2.1.1 Système syllabique de l'arabe

Le système syllabique de l'arabe est composé de syllabes, de consonnes et des voyelles, la « materia prima phonétique » (Roman, 1999)[162], permettant de former trois ensembles :

- L'ensemble des consonnes {C}
- L'ensemble des voyelles {V}
- L'ensemble des syllabes {S}

Ce système syllabique implique la répartition de l'ensemble des consonnes et des voyelles en deux ensembles disjoints et l'ensemble des syllabes, comme la combinatoire des ensembles de consonnes et de voyelles. Autrement dit, les consonnes et les voyelles appartiennent à deux ensembles différents mais peuvent être utilisées indépendamment les unes des autres :

> *« [ce] système syllabique particulier à la langue arabe a déterminé un fait fondamental pour l'organisation de cette langue : la séparation totale entre ses consonnes et ses voyelles » (Roman, 1999)[163]*

La langue arabe compte 28 consonnes (C) comme بـ=b et 6 voyelles (V) comme ـِ=i ; les syllabes peuvent être seulement de deux types, soit S = {CV} comme لَـ=la, soit S = {CVC} comme باب=bâb où S = syllabe, C = consonne et V = voyelle :

> *« [un] tel système syllabique détermine une disjonction de l'ensemble des consonnes }C} et de l'ensemble des voyelles }V} : S=}CV, CVC} → }C} M }V}=O. Cette disjonction, dès lors que les consonnes et les voyelles peuvent être utilisées indépendamment les unes des autres, permet l'attribution systématique de tâches différentes aux consonnes et aux voyelles » (Roman, 1990)[164]*

Autrement dit, la combinatoire des ensembles de consonnes et de voyelles n'est possible que selon ces deux syllabes à moins d'y être contraint phonétiquement ou syntaxiquement :

> *« En arabe classique encore, tout arrangement de consonnes et de voyelles qui, dans le cadre d'une syllabe, ne serait ni « CV » ni « CVC » constitue une syllabe a- systématique produite par une contrainte phonétique ou par une pause syntaxique » (Roman, 1999)[165]*

De ce fait, les consonnes et les voyelles sont alors disponibles dans le système de la langue pour réaliser deux tâches fondamentales :

---

[162] Grammaire de l'arabe, p. 10
[163] Idem, p. 117
[164] Idem, p. 6
[165] La création lexicale en arabe, p. 18

- L'une assurée par le sous-système de nomination de la langue, construit sur des racines de consonnes, pour permettre aux hommes de nommer des entités extralinguistiques,
- L'autre assurée par le sous-système de communication de la langue, construit à l'origine sur des voyelles, pour permettre aux hommes de communiquer entre eux.

2.1.2 Système de nomination de l'arabe

Le système de nomination de l'arabe est construit sur la disjonction de l'ensemble des voyelles et des consonnes par des racines consonantiques :

> *« La langue arabe, comme toute autre langue sémitique, s'est constituée en exploitant cette disjonction : elle a établi son système de nomination sur une combinatoire de consonnes, ses unités de nomination sont construites sur des racines de consonnes » (Roman, 1987)*[166]

En effet, langue sémitique, l'arabe dispose par la répartition de ses voyelles et de ses consonnes d'une structure fondamentale basée sur des racines consonantiques, structure qui apparaît de la manière la plus claire en arabe. Cette combinatoire apparaît fondamentalement comme une combinatoire de consonnes du fait du nombre important de consonnes :

> *« En effet, les langues sémitiques ont construit leur système de nomination sur des arrangements de consonnes, la combinatoire de leurs consonnes étant plus puissante du fait de leur nombre largement plus élevé, toujours, que celui de leurs voyelles » (Roman, 1999)*[167]

Généralement, les racines consonantiques sont construites sur trois consonnes permettant de produire en nombre suffisant les arrangements constituant les racines, mais des racines biconsonantiques et quadriconsonantiques sont aussi utilisées pour répondre aux besoins de nomination. Ces racines consonantiques représentent la plupart des unités de nomination et sont appelées unités fléchies. Ces unités fléchies peuvent être considérées comme les images linguistiques des res et des modus inventés par les hommes et l'arabe en a fait les deux unités indispensables du noyau de la phrase :

> *« Les res [sont] des entités du monde imaginées par lui [l'homme] hors temps, comme étrangères au temps, dont le temps n 'est pas une composante [...] Les res linguistiques dénotent précisément soit un objet « enfant » soit une idée « grandeur. » Dans le premier cas de figure, la res est concrète et dans le deuxième cas, elle est abstraite.*
>
> *[Quant aux modus, ce sont] des entités du monde imaginées par l'homme dans le temps, comme s'inscrivant dans un déroulement apparent du temps, dont le temps est l'une des composantes [...]*

[166] Les formes infinitives de l'arabe dans l'organisation générale de la langue arabe, p. 212
[167] La création lexicale en arabe, p. 19

Les modus linguistiques dénotent précisément soit une action - un mouvement, par exemple « enfanter » ou un changement, par exemple « grandir », soit une actualisation, c'est-à-dire, un état résultant d'une action ou d'un changement, par exemple « enfantin » ou « grand ». (Roman, 1990)[168]

Quant aux autres unités de nomination, qui ne sont pas construites sur des racines, elles sont appelées unités amorphes ; elles comprennent les modalités qui font partie du système linguistique et les unités amorphes hors système, notamment les emprunts et les sigles :

« Les signifiants de ces modalités sont d'une part les voyelles, dont le système syllabique impose l'emploi et dont c'est la position par rapport aux consonnes radicales qui établit le signifié qui leur est propre, et d'autre part des unités libres, comptant au moins une syllabe. » (Roman, 1990)[169]

De ce fait, ces unités interviennent sur les termes par des modalités déterminant soit une res comme le défini, par l'article défini « ال = al », soit un modus, comme l'agentivité pour les verbes (Lelubre, 1992)[170]. Ces res et ces modus sont construits sur des racines ; elles peuvent être générales, particulières ou se combiner. Il n'existe que deux modus qui soient des unités de nomination générale ; construits sur une racine monoconsonantique, le modus 'faire', qui a deux signifiants /s/ et /?/ toujours en position préfixale, et le modus 'être' qui a un seul signifiant /y/ toujours en position suffixale. Mais, il existe différents modus qui sont des unités de nomination particulière, notamment les modus personnels ou verbes et les modus impersonnels tels que les noms d'action, les participes, … (Lelubre, 1992)[171].

2.1.2.1 Modus personnel ou verbe

La majorité des verbes arabes sont formés d'une racine triconsonantique et éventuellement tétraconsonantique, associés à différents schèmes. Nous analysons les verbes de notre corpus :

- Verbe à racine triconsonantique simple : ce sont des verbes simples construits sur une racine triconsonantique √CCC ; ils correspondent à 1ère forme de la tradition grammaticale des arabisants. Par exemple le verbe « غسل = ġasala = rincer ».
- Verbe à racine triconsonantique augmenté : ce sont des verbes construits sur une racine triconsonantique √CCC avec addition de modalité ; il s'agit de la 2ème forme de la

[168] Grammaire de l'arabe, p. 3
[169] Idem, p. 6
[170] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 129
[171] Idem, p. 138

tradition grammaticale des arabisants, correspondant à une valeur itérative, voire une valeur intensive (Dichy et al., 1999)[172]. Par exemple le verbe « شرّح = *š*arra*ḥ*a = filtrer ».

- Verbe à racine triconsonantique connecté à la 2ème forme de la tradition grammaticale des arabisants : ce sont des verbes construits sur une racine triconsonantique √CCC ; ils correspondent à 5ème forme de la tradition grammaticale des arabisants. Par exemple le verbe « تشكّل = ta*š*akkala = se former ».
- Verbe à racine quadriconsonantique : ce sont des verbes, construits sur une racine quadriconsonantique √CCCC ; il s'agit « de formes expressives, onomatopées, verbalisation de formules ou de termes techniques empruntés » (Lelubre, 1992)[173]. Par exemple le verbe « بلور = balwara = cristalliser ».

Dans les langues de spécialités en général, en chimie en particulier, la nominalisation est beaucoup plus fréquente que le verbe correspondant. Nous avons ainsi étudié plusieurs dizaines de couples nom/verbe ; par exemple pour le couple « حضّر/تحضير = *ḥ*a*ḍḍ*ara/ta*ḥḍî*r = préparation/préparer », le nom d'action, nom morphologiquement relié au verbe, est largement plus employé dans notre corpus. Cette nominalisation correspond au modus impersonnel.

2.1.2.2 Modus impersonnel

Les modus impersonnels sont les noms d'action et les participes.

2.1.2.2.1 Nom d'action

Le nom d'action, appelé également nom verbal, est le plus important en terminologie et correspond au *maṣd*ar de la tradition grammaticale arabe ; il représente l'action faite par le verbe, dénommant de nombreux termes abstraits, dénotant des processus, des phénomènes.

- Nom d'action de la 1ère forme : la forme simple du verbe : « ضغط = *ḍ*a*ġṭ*= pression ».
- Nom d'action de la 2ème forme : taf`îl : « تحريك = ta*ḥ*rîk = agitation ».
- Nom d'action de la 4ème forme : ’if`âl : « إرجاع = ’irjâ‘ = réduction ».
- Nom d'action de la 5ème forme : tafa`ul : « تشكّل = ta*š*akkul = formation ».
- Nom d'action de la 6ème forme : tafâ`ul : « تفاعل = tafâ`ul = réaction ».
- Nom d'action de la 7ème forme : infi`âl : « انصهار = in*ṣ*ihâl = fusion ».
- Nom d'action de la 8ème forme : ifti`âl : « امتصاص = imti*ṣ*â*ṣ* = absorption ».

[172] Les verbes arabes

[173] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 142

- Nom d'action de la 10ème forme : istif âl : « استخدام = isti*ḥd*âm = utilisation ».

2.1.2.2.2 Participe actif

Le participe actif correspond à l'ism al-fâ*'il* de la tradition grammaticale arabe construit sur le schème fâ*'il*. Il désigne celui qui exécute l'action pour être tantôt utilisé comme un adjectif (ou un nom), tantôt avec sa valeur verbale. Il est souvent présent comme adjectif dans les unités terminologiques complexes. Par exemple « ناتج= nâtij = produit ».

2.1.2.2.3 Participe passif

Le participe passif correspond à l'ism al-maf'ûl de la tradition grammaticale arabe construit sur le schème maf'ûl. Il désigne celui qui subit l'action ou parfois le résultat de l'action. Par exemple « مركب = murakkab = composé ».

2.1.2.2.4 Nom de lieu

Le nom de lieu correspond à l'ism al-makân de la tradition grammaticale arabe, construit sur les schèmes maf'al et maf'il. Par exemple « مخبر= ma*ḥ*bar = laboratoire ».

2.1.2.2.5 Nom d'instrument

Le nom d'instrument correspond à l'ism al-âla de la tradition grammaticale arabe, construit sur les schèmes mif'al, mif'ala et mif'âl. Par exemple « مقياس= miqyâs = jauge ».

2.1.2.3 Recours à des affixes et à des formants

Le système de nomination dispose également d'autres unités, les affixes et les formants (Lelubre, 1992)[174].

Les premiers ne peuvent apparaître que comme constituants d'une unité de nomination, étant placés avant une base, les préfixes, ou bien après, les suffixes. Par conséquent, ils n'ont aucune autonomie et ils représentent un ensemble fermé, correspondant en arabe classique aux racines monoconsonantiques. Par exemple, « حمضية = *ḥamḍiyyat* = acidité ».

Quant aux seconds, ce sont des éléments constitués à partir de lexies, prises telles quelles ou bien tronquées ; certains peuvent être syntaxiquement autonomes. Ils peuvent être créés à partir d'emprunts faits à d'autres langues, notamment des lexies latines ou grecques pour la formation des termes scientifiques. Certains de ces formants sont en position préfixale (antéposés), d'autres en position suffixale (postposés). Contrairement aux affixes, les formants constituent

[174] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 164

une liste a priori ouverte, susceptible en fonction des besoins de nomination d'être augmentée pour tel ou tel domaine ; mais l'arabe est une langue très pauvre en formants.

Pour la terminologie scientifique en général, pour celle de la chimie en particulier, l'arabe est confronté à l'emploi massif d'affixes et de formants ; ces affixes et ces formants sont rendus en arabe de plusieurs façons (cf. 2.2 Création lexicale en arabe).

Toutes ces unités sont des unités du système de nomination de l'arabe correspondant aux unités lexicales simples constituées par un seul mot, « le mot étant défini comme une unité morphologique syntaxique autonome » (Roman, 1999)[175] ; elles peuvent être des termes. Nous les appelons 'unités terminologiques simples' (UTS).

Le système de nomination de l'arabe permet de créer des unités terminologiques simples mais face à l'important nombre de termes qu'il faut créer, le système de nomination se trouve rapidement saturé :

> *« L'opération de nomination ne peut rester dans le cadre du seul système de nomination « alors [que] le nombre maximum des unités de nomination sera égal au nombre maximum des formes qui peuvent être générées par ce système […] Aussi par la force des choses, l'opération de nomination n'est-elle pas restée réduite aux seules formes du système de nomination : elle a combiné, linéairement, des formes différentes dans une même unité terminologique qu'elle a ainsi constituée en syntagme, c'est-à-dire en constituants de phrases faits de « n » unités de nomination.. » (Roman, 1999)*[176]

Par conséquent, le système de communication prend le relai afin d'assurer l'opération de nomination.

Dans un domaine tel que la chimie, le besoin de précision et d'explication dans la dénomination de ses concepts témoignent de l'impossibilité de la nomination en dépendant du seul système de nomination, d'où le recours au système de communication.

2.1.3 Système de communication de l'arabe

Le système de communication est considéré comme étant l'un des quatre sous-systèmes de la langue arabe permettant aux hommes de communiquer entre eux (Roman, 1999)[177]. Sa fonction principale est de former la phrase permettant ainsi de créer des unités terminologiques complexes (UTC) :

[175] La création lexicale en arabe, p. 137
[176] Idem, p. 102
[177] Idem, p. 21

« Le rôle fondamental du système de communication est d'assurer l'insertion des unités de nomination produites par le système de nomination au sein de la phrase, la phrase étant l'unité maximale du système de communication. Mais le système de communication joue aussi un autre rôle : […] (il) permet, au même titre que le système de nomination et concurremment à lui, de créer des unités terminologiques, non pas formées d'un seul constituant (UTS) comme celles formées par le système de nomination, mais comprenant alors plusieurs constituants (UTC). » (Lelubre, 1992)[178]

Ce système de communication offre la possibilité de former des unités terminologiques complexes par le biais d'une combinatoire linéaire des unités terminologiques simples crée par le système de nomination (Lelubre, 1992)[179].

2.1.3.1 Unité de communication

Le système de communication possède comme ressources fondamentales les voyelles brèves comme voyelles désinentielles à la fin de chaque unité de nomination :

« Ces voyelles désinentielles étaient les pièces primitives du sous-système de communication de l'arabe classique » (Roman, 1999)[180]

À ces voyelles désinentielles considérées comme 'les fonctionnels liés au système', le système de communication dispose d'autres moyens : des fonctionnels autres que les voyelles désinentielles et des coordonnants.

2.1.3.1.1 Fonctionnel

À l'exception des voyelles désinentielles casuelles, les fonctionnels sont tous des morphèmes libres :

- Les fonctionnels non spécifiés sémantiquement, notamment la voyelle de l'accusatif et l'identité de la voyelle désinentielle entre l'expansion et sa base que la tradition grammaticale arabe nomme l'accord.
- Les fonctionnels spécifiés sémantiquement, qui introduisent des syntagmes auxquels ils donnent une situation syntaxique propre ; ils correspondent aux prépositions, parmi lesquelles certaines sont thermogènes comme « من = min = de » dans « جو من الأزوت = jaww min al'azût = atmosphère d'azote ».
- D'autres fonctionnels supplémentaires que l'arabe s'est donné par figement d'unités fléchies à l'accusatif, avec la voyelle « ◌َ = a » ; parmi lesquels ceux qui interviennent

[178] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 188
[179] Idem, p. 25
[180] La création lexicale en arabe, p. 20

dans la formation des UTC, notamment « تحت = ta*ḥ*ta = au-dessous » dans « إشعاع تحت الحمراء = ʾišʿâʿ *taḥta alḥamrâʾ* = rayonnement infrarouge ».

2.1.3.1.2 Coordonnant

Les autres unités du système de communication sont les coordonnants ; ils permettent de coordonner des phrases ou des syntagmes à l'aide des coordonnants « وَ = wa = et », « أَو = aw = ou » et le coordonnant 'zéro', comme dans « أكسدة إرجاع = ʾaksada ʾirjâʿ = oxydo-réduction ».

Avec ces unités de communication, le système de communication permet de créer au sein de la phrase les unités terminologiques complexes comprenant plusieurs constituants, correspondant aux unités référentielles à dénommer.

2.1.3.2 Formation des unités terminologiques complexes (UTC)

Une unité terminologique complexe est formée de deux éléments, une base et une extension. Chacune d'entre elles peut être simple, indécomposable, ou bien elle-même être complexe, décomposable à son tour en base et extension (Roman, 1999)[181] :

- Une base
  - Simple et indécomposable
  - Complexe et décomposable
- Une extension
  - Simple et indécomposable
  - Complexe et décomposable

Les formes possibles d'une UTC sont au nombre de quatre (cf. Tableau 1 : Différentes formes possibles des UTC) :

[181] La création lexicale en arabe, p. 87

Eléments composant l'UTC	UTC en arabe	UTC en français
Base simple Extension simple	تكـاثف الدولي	condensation aldolique
Base simple Extension complexe	طيف الرنين النووي	spectre de résonance nucléaire
Base complexe Extension simple	الهيدروكربون الحلقية المشبعة	hydrocarbure aromatique saturé
Base complexe Extension complexe	طيف الامتصاص للأشعة تحت الحمراء	spectre d'absorption des rayons infrarouges

Tableau 1 : Différentes formes possibles des UTC

Cela peut aboutir à la formation des unités terminologiques complexes, très longues, et qui sont largement utilisées dons notre corpus. Ces unités terminologiques complexes sont formées sur la relation base - extension ; c'est une relation binaire qui met en jeu deux éléments et pas plus, en l'occurrence, l'élément de la base et celui de l'extension (Roman, 1999)[182].

Il existe en arabe deux types d'extension, selon la nature de la relation univoque qui la réunit à sa base : extension par coordination et extension par subordination. Cette base peut être soit substantivale, l'UTC sera une UTC substantivale, soit adjectivale, l'UTC sera une UTC adjectivale. Pour notre corpus, nous avons repéré seulement des UTC substantivales.

Notre analyse des UTC se fait en fonction de la nature de ses éléments, c'est-à-dire, la structure de la base et celle de l'extension. Nous précisons également s'il s'agit d'une base simple ou complexe, et d'une extension simple ou complexe (Albeiriss, 2017)[183]. Chaque cas est illustré par des exemples de notre domaine d'étude et nous utilisons comme notation les symboles suivants afin d'exprimer la nature des éléments de l'UTC :

- S : unité lexicale substantivale
- A : unité lexicale adjectivale
- BS : base annective substantivale
- AB : base annective adjectivale
- P : préposition, fonctionnel
- F : formant
- & : Coordonnant

2.1.3.2.1 Extension par coordination

La coordination est « une relation égale, non hiérarchisée, [...] qui jumelle à sa base une extension de même statut syntaxique que celle-ci, les coordonnants en sont les différents

[182] La création lexicale en arabe, p. 86

[183] Modélisation des unités lexicales complexes d'une langue spécialisée : le cas de la chimie en arabe

moyens » (Roman, 1999)[184]. Dans ce type d'extension, les deux éléments peuvent être interchangeables et l'un des deux éléments se trouve avant l'autre sur l'axe syntagmatique ; par exemple, « أكسدة إرجاع = ’aksada ’irjâ ‘ = oxydo-réduction ».

Dans notre corpus, ce cas n’a pas été répertorié ; cependant, nous identifions des extensions de coordination en combinaison avec des extensions de subordination que nous allons présenter.

2.1.3.2.2 Extension par subordination

La relation de subordination est « une relation inégale, hiérarchisée, [et] qui rattache une extension à sa base, l'extension recevant le statut syntaxique qui lui octroie le fonctionnel qui signifie son rattachement » (Roman, 1999)[185]

On distingue en arabe quatre types d'extension par subordination, ou encore expansion, qui participent à la formation des unités terminologiques complexes : l'expansion d'identification, l'expansion annective, l'expansion complétive et l'expansion modale.

2.1.3.2.2.1 UTC substantivale à expansion d’identification ou épithétique

L’expansion d’identification ou d’identité correspond l'épithète « alna`t wa albadal = النعت والبدل » de la tradition grammaticale arabe, et est très utilisée dans la formation des UTC.

L’UTC substantivale à expansion épithétique est composée :

- D’une base
  - Simple ; il s’agit d’un substantif
  - Complexe ; il peut s’agir d’une UTC que nous noterons UTC’
- D’une expansion
  - Simple ; il s’agit d’un adjectif
  - Complexe ; il peut s’agir d’une UTC que nous noterons UTC’

Voici la formation des UTC à expansion épithétique repérées dans notre corpus :

- UTC = Substantif + Adjectif : cette UTC est formée d’une base substantivale simple S1, pouvant correspondre à une unité terminologique simple, et d’une expansion d’identification A1, pouvant correspondre à une unité terminologique simple. Par exemple, « تكـاثف الدولي = takâ*ṯuf* ’aldûlî = condensation aldolique » ; voici son schéma

[184] La création lexicale en arabe, p. 180

[185] Idem, p. 181

morphosyntaxique (cf. Figure 2 : Schéma morphosyntaxique d'une UTC de type S1A1) :

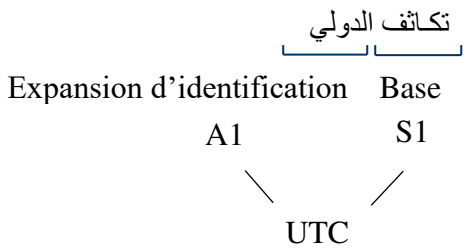


Figure 2 : Schéma morphosyntaxique *d'une* UTC de type S1A1

- UTC = UTC' + Adjectif : cette UTC est composée de l'UTC', formée d'une base substantivale simple S1, pouvant correspondre à une unité terminologique simple, et d'une expansion d'identification A1, pouvant correspondre à une unité terminologique simple, suivie d'une expansion d'identification A2, pouvant correspondre à une unité terminologique simple. En d'autres termes, cette UTC possède deux adjectifs qui ne sont pas interchangeables ; l'élément A2 est une expansion d'identification de la base complexe (S1 A1), qui est elle-même composée de la base substantivale S1 et de l'expansion d'identification A1. Par exemple, « تفاعل سلسلي متفرع = tafâ'ul silsilî mutafarri' = réaction en chaîne ramifiée » ; voici son schéma morphosyntaxique (Figure 3 : Schéma morphosyntaxique d'une UTC de type S1A1A2) :

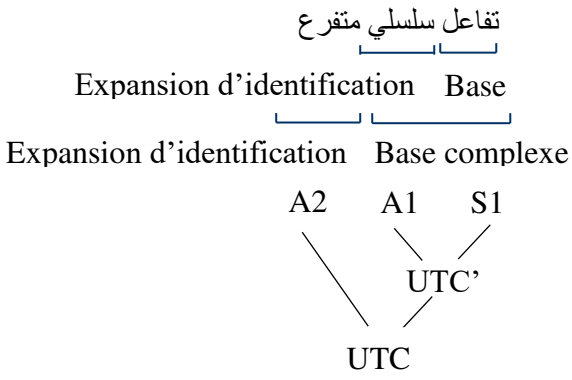


Figure 3 : Schéma morphosyntaxique *d'une* UTC de type S1A1A2

Cette UTC peut avoir deux interprétations : soit l'analyse présentée ci-dessus, soit une base substantivale simple, suivi d'une expansion d'identification complexe. Nous n'avons pas retenu la deuxième interprétation, considérant que la seconde expansion d'identification est liée à la base complexe.

- UTC = Substantif + Adjectif + Adjectif : cette UTC est composée d'une base substantivale simple S, correspondant à une unité terminologique simple, suivie d'une expansion d'identification complexe A1 A2, correspondant à deux unités

terminologiques simples. En d'autres termes, cette UTC possède deux adjectifs reliés par le coordonnant « zéro ». Par exemple, « هيدروكربون أليفاتي مشبع = hîdûkarbûn alifatiya mu*šaba*ʿ = hydrocarbure aliphatique saturé » ; voici son schéma morphosyntaxique (cf. Figure 4 : Schéma morphosyntaxique d'une UTC de type SA1A2) :

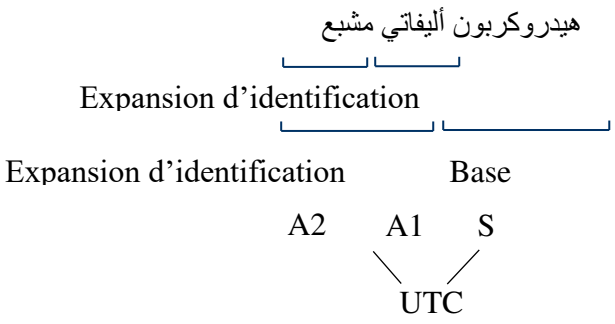


Figure 4 : Schéma morphosyntaxique d'une UTC de type SA1A2

Cette UTC peut avoir deux interprétations : soit l'analyse présentée ci-dessus, soit une cette UTC est composée de l'UTC', formée d'une base substantivale simple et d'une expansion d'identification, suivie d'une expansion d'identification. Nous n'avons pas retenu la deuxième interprétation, considérant que les deux adjectifs sont des expansions d'identification de la base substantivale.

2.1.3.2.2.2 UTC substantivale à expansion annective

L'expansion annective ou d'annexion correspond complément d'annexion « mudâf 'ilayhi = المضاف اليه », très fréquente dans les UTC.

L'UTC substantivale à expansion annective est composée :

- D'une base
  - Simple ; il s'agit d'un substantif
  - Complexe ; il peut s'agir d''une UTC que nous notons UTC'
- D'une expansion
  - Simple ; il s'agit d'un substantif
  - Complexe ; il peut s'agir d'une UTC que nous notons UTC'

Voici la formation des UTC à expansion annective repérées dans notre corpus :

- UTC = Substantif + Substantif : cette UTC est formée d'une base substantivale simple S1, correspondant à une unité terminologique simple, et d'une expansion d'annexion

S2, correspondant à une unité terminologique simple. Par exemple, « طريقة العمل = ṭarîqat al'amal = mode opératoire ». Voici son schéma morphosyntaxique (cf. Figure 5 : Schéma morphosyntaxique d'une UTC de type S1S2) :

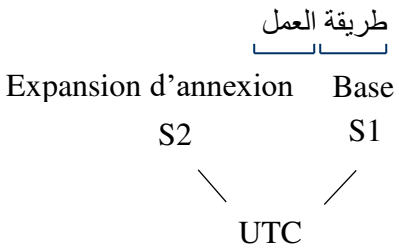


Figure 5 : Schéma morphosyntaxique *d'une* UTC de type S1S2

- UTC = UTC' + Substantif : cette UTC est composée de l'UTC' formée d'une base substantivale simple S1, correspondant à une unité terminologique simple, et d'une expansion d'annexion S2, correspondant à une unité terminologique simple, suivie d'une expansion d'annexion S3, correspondant à une unité terminologique simple. En d'autres termes, l'élément S3 est une expansion d'annexion de la base complexe (S1 S2), qui est elle-même composée de la base substantivale S1 et de l'expansion d'annexion S3. Par exemple, « نواة ذرة الكربون = nawât ḏarrat alkarbûn = noyau de l'atome de carbone ». Voici son schéma morphosyntaxique (cf. Figure 6 : Schéma morphosyntaxique d'une UTC de type S1S2S3) :

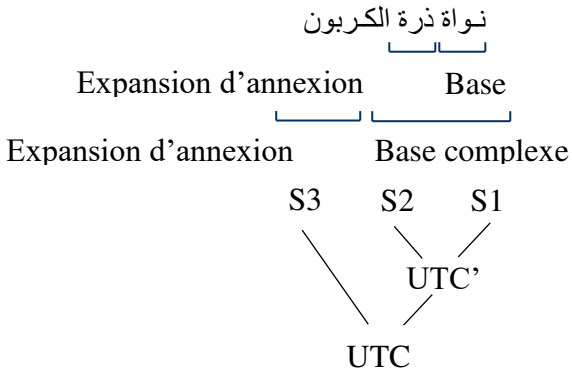


Figure 6 : Schéma morphosyntaxique *d'une* UTC de type S1S2S3

- UTC = Substantif + UTC' : cette UTC est composée d'une base substantivale simple S1, correspondant à une unité terminologique simple, suivi de l'UTC' formée d'une base substantivale simple S2, correspondant à une unité terminologique simple et d'une expansion d'identification A2, correspondant à une unité terminologique simple. En d'autres termes, la base complexe (S2A2) est une expansion d'annexion de la base simple S1. Par exemple, « طيف الرنين النووي = ṭayf alranîn alnawawî = spectre de

résonance nucléaire ». Voici son schéma morphosyntaxique (cf. Figure 7 : Schéma morphosyntaxique d'une UTC de type S1S2A2) :

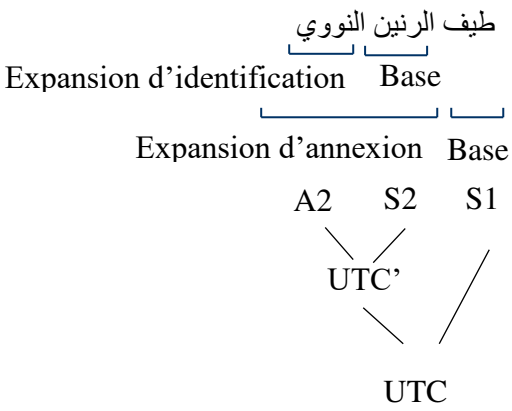


Figure 7 : Schéma morphosyntaxique *d'une* UTC de type S1S2A2

2.1.3.2.2.3 UTC substantivale à expansion complétive

L'expansion complétive correspond au « jârr wa majrûr = جار ومجرور » de la grammaire arabe traditionnelle ; elle est introduite par une préposition, appelée aussi joncteur dans la littérature terminologique.

L'UTC substantivale à expansion complétive est composée :

- D'une base
  - Simple ; il s'agit d'un substantif
  - Complexe ; il peut s'agir d'une UTC que nous notons UTC'
- D'une préposition
- D'une expansion
  - Simple ; il s'agit d'un substantif
  - Complexe ; il peut s'agir d'une UTC que nous notons UTC'

Cette description montre la complexité des UTC formées pour cette expansion. Voici la formation des UTC à expansion complétive repérées dans notre corpus :

- UTC = Substantif + Préposition + Substantif : cette UTC est formée d'une base substantivale simple S1, pouvant correspondre à une unité terminologique simple, suivie d'une préposition P1 et d'une expansion complétive S2, pouvant correspondre à une unité terminologique simple. Par exemple, « قطرة بقطرة = *qaṭra*t *biqaṭra*t = goutte à goutte ». Voici son schéma morphosyntaxique (cf. Figure 8 : Schéma morphosyntaxique d'une UTC de type S1P1S2) :

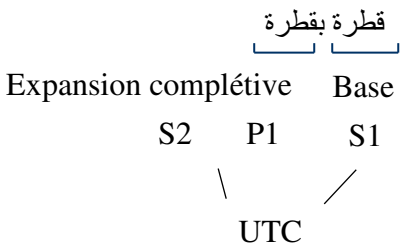


Figure 8 : Schéma morphosyntaxique *d'une* UTC de type S1P1S2

- UTC = Substantif + Préposition + UTC' : cette UTC est formée d'une base substantivale simple S1, correspondant à une unité terminologique simple, suivie d'une préposition P1 et de l'UTC', composée d'une base substantivale simple S2, correspondant à une unité terminologique simple et d'une expansion. Par exemple, « طريقة للتفاعل الألدولي = *ṭ*arîqat litafâ ul al'aldûlî = synthèse par la réaction aldolique ». Voici son schéma morphosyntaxique (cf. Figure 9 : Schéma morphosyntaxique d'une UTC de type S1 P1S2) :

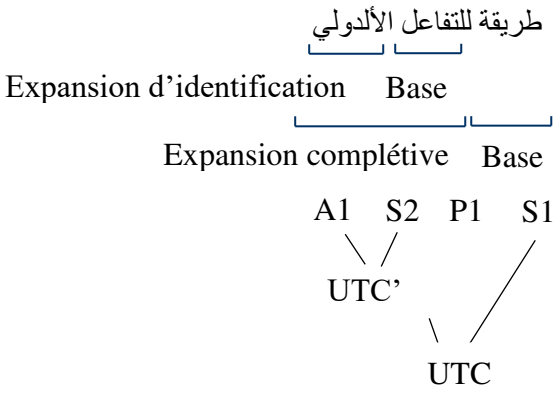


Figure 9 : Schéma morphosyntaxique *d'une* UTC de type S1 P1S2

- UTC = UTC' + Préposition + UTC' : cette UTC est formée de l'UTC', composée d'une base substantivale simple S1, correspondant à une unité terminologique simple et d'une expansion, suivie d'une préposition P1 et de l'UTC'', composée d'une base substantivale simple S2, correspondant à une unité terminologique simple et d'une expansion. Par exemple, « محلول مشبع من كلوريد الصوديوم = *maḥl*ûl mu*šb*a min klurîd alṣûdyûm = solution saturée de chlorure de sodium ». Voici son schéma morphosyntaxique (cf. Figure 10 : Schéma morphosyntaxique d'une UTC de type S1A1P1S2S3) :

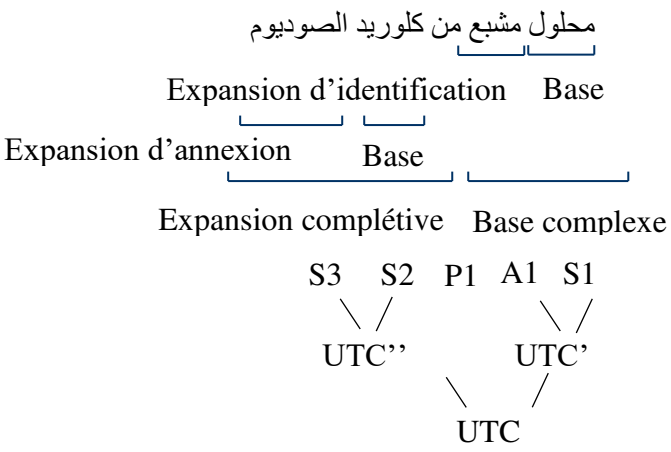


Figure 10 : Schéma morphosyntaxique *d'une* UTC de type S1A1P1S2S3

2.1.3.2.2.4 UTC substantivale à expansion modale

L'expansion modale correspond au « tamyîz = spécificatif », le « hâl = complément d'état » et le « mafûl mutlaq = complément absolu » de la tradition grammaticale arabe ; « elle dénote dans la phrase une certaine actualisation de sa base » (Roman, 1999)[186]. Ce cas n'a pas été repéré dans notre corpus.

2.1.3.2.3 Cas particulier

Dans notre corpus, nous observons la combinaison de différents patrons morphosyntaxiques, les UTC à base annective, mais également des constructions syntaxiques qui ne sont pas répertoriées dans la tradition grammaticale arabe, les UTC de la nomenclature chimique.

2.1.3.2.3.1 UTC à base annective

Les UTC à base annective contiennent des éléments qui interviennent comme base d'annexion. Ces éléments sont appelés par hypallage 'bases annectives' (Lelubre, 1992)[187] ; ces bases annectives peuvent être adjectivales ou substantivales.

- Base adjectivale « ذو = *ḏû* = qui a » : elle est utilisée uniquement comme une base d'expansion annective et l'élément d'expansion est alors substantival. Elle se décline au masculin en ذُو pour le nominatif, en ذَا pour l'accusatif, en ذِي pour le génitif et au féminin en ذَاتُ. L'UTC est alors formée d'une base substantivale simple S1, suivie d'une expansion d'identification complexe qui se compose de la base annective adjectivale AB2 et de son expansion annective S2. Par exemple, « مركبات ذات سلاسل مستقيمة = murakkabât *ḏât* salâsil mustaqîmat = composés possédant des chaînes

[186] La création lexicale en arabe, p. 180
[187] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation,

linéaires ». Voici son schéma morphosyntaxique (cf. Figure 11 : Schéma morphosyntaxique d'une UTC de type S1AB1S2A) :

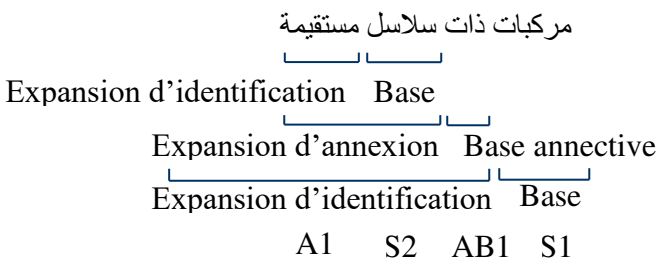


Figure 11 : Schéma morphosyntaxique *d'une* UTC de type S1AB1S2A1

- Base substantivale « غير = *ġ*ayr = autre que » : elle est comme une base annective qui est toujours suivie par une expansion adjectivale. L'UTC est alors formée d'une base substantivale simple S1, suivie d'une expansion d'identification complexe qui se compose de la base annective substantivale SB1 et de son expansion adjectivale A1. Par exemple, « مركب غير مشبع = murakkab *ġayr mušbaʿ* = composé insaturé ». Voici son schéma morphosyntaxique (cf. Figure 12 : Schéma morphosyntaxique d'une UTC de type S1SB1A1) :

Figure 12 : Schéma morphosyntaxique *d'une* UTC de type S1SB1A1

2.1.3.2.3.2 UTC de la nomenclature chimique

Dans la nomenclature chimique en arabe, les unités terminologiques complexes, composées de deux mots, correspondent aux composés chimiques, issus des noms triviaux comme « كبريتات الصوديوم = *kibrîtât alṣûdyûm* = sulfate de sodium », de type S1S2 (cf. 2.1.3.2.2.2 UTC substantivale à expansion annective), et aux composés chimiques, issus de la nomenclature systématique comme « ميتوكسي بنزالدهيد = mîtûksî banzâldahîd = méthoxybenzaldéhyde ». Ce terme, comme la majorité des composés chimiques issus de la nomenclature systématique, s'exprime en arabe par des UTC, en raison de la présence de formants qui sont des préfixes multiplicatifs et/ou des préfixes géométriques et/ou des préfixes fonctionnels (cf. 2.2.2.1.4 Dérivation). La construction syntaxique de ces UTC n'est pas répertoriée dans la grammaire arabe ; ces nouvelles constructions morphosyntaxiques sont des calques du français et/ou de l'anglais contenant parfois des signes brachygraphiques tels que des caractères latins comme H, des caractères grecs comme α, des caractères numériques comme 2 et des ponctuations comme les parenthèses () (cf. 2.2.2.1.3 Siglaison). Il s'agit

toujours d'une structure binaire, formée de deux éléments au moins, correspondant à des substantifs et/ou des formants et/ou d'affixes (Albeiriss, 2016)[188].

Par exemple, « هيدروكلوريد إيثيل إستر الغليسين » est une UTC constituée d'une base complexe, comportant plusieurs unités terminologiques simples, et d'une expansion annective, composée d'une unité terminologique simple. Pour cette UTC, nous avons identifié différents équivalents en anglais et en français :

- « glycine ethyl ester hydrochloride »
- « ethyl glycinate hydrochloride »
- « ethyl 2-aminoacetate hydrochloride »
- « ethyl aminoacetate hydrochloride »
- « amino-acetic acid ethyl ester hydrochloride »
- « glycine d'éthyle ester chlorhydrate »
- « chlorhydrate d'ester éthylique de glycine »

Pour cette UTCen arabe, nous observons l'emprunt des termes de l'anglais, notamment avec « hydrochloride » ; mais l'ordre des composants en arabe est différent des autres langues.

Voici son schéma morphosyntaxique (cf. Figure 13 : Schéma morphosyntaxique d'une UTC de type S1F1F2S2) :

هيدروكلوريد إيثيل إستر الغليسين
Base complexe Expansion annective
S1 F1 F2 S2

Figure 13 : Schéma morphosyntaxique *d'une* UTC de type S1F1F2S2

Les systèmes de nomination et de communication permettent d'identifier les unités terminologiques simples et complexes ; cependant, il existe d'autres procédés linguistiques pour la création lexicale en arabe.

2.2 Création lexicale en arabe

La création des termes s'effectue par différents procédés linguistiques, notamment la composition, la dérivation, l'emprunt et le transfert sémantique. Ces procédés

[188] Étude Terminologique de la Nomenclature de la Chimie en Arabe dans une Approche de Fouille de Textes

morphosyntaxique et morphosémantique sont relatifs à la forme et au sens des termes afin d'assurer le lien entre le terme et l'unité référentielle qu'il dénomme. Ils sont basés sur la formation d'un nouveau mot qui peut se faire grâce aux unités lexicales déjà existantes. Il s'agit du procédé linguistique plus général, la néologie.

2.2.1 Néologie et définition

La néologie et son principal objet d'étude, le néologisme, sont des concepts plutôt complexes à définir comme l'a souligné Rey (1976)[189]et l'explique Sablayrolles :

> *« La néologie n'est sans doute pas un concept discret, mais comporte plutôt différents degrés sur une échelle. Cette conception large et scalaire de la néologie explique la variabilité des jugements au sujet des néologismes et la présence dans le corpus d'éléments qui ne seraient pas spontanément et unanimement considérés comme des néologismes » (Sablayrolles 2000)[190].*

De ce fait, la définition de Guilbert (1975)[191] considérant la néologie lexicale comme étant « la possibilité de création de nouvelles unités lexicales, en vertu de règles de production incluses dans le système lexical » nous semble pertinente mais nous adoptons plutôt celle de Cabré :

> *« La néologie dans son sens général, est la discipline qui s'occupe des aspects relatifs aux nouveaux phénomènes qui apparaissent dans les langues. » (Cabré,1998)[192]*

La néologie, processus de formation de nouvelles unités lexicales, spécifique à un système donné, produit des unités lexicales nouvelles appelées néologismes, afin de répondre à un besoin de communication précis.

2.2.1.1 Néologie et néonymie

La néologie se présente dans toutes les situations de la communication et non seulement dans le discours spécialisé. Ainsi, deux sortes de néologies se distinguent : la néologie proprement dite et la néonymie :

> *« On utilise le terme de "néonymie" pour la néologie terminologique, le terme "néologie" ou "néologie lexicale" étant alors réservé à la langue commune. La néonymie est la mise au point d'un terme pour*

[189] Néologisme, un pseudo concept
[190] La néologie en français contemporain : examen du concept et analyse de productions néologiques récentes, p. 12
[191] La créativité lexicale, p. 31
[192] La terminologie - Théorie, méthode et applications, p. 252

une unité référentielle n'ayant pas encore de dénomination dans la langue concernée ou bien ayant déjà une, voire plusieurs dénominations jugées insatisfaisantes. » (Lelubre, 1992)[193]

Le néonyme est donc une création appartenant à la langue de spécialité qui répond toujours à un besoin de communication et qui doit respecter des critères précis.

2.2.1.2 Principaux critères de création néonymique

Les terminologues ont établi un certain nombre de critères auxquels doit satisfaire un terme, comme la brièveté, la simplicité, la dérivabilité…

Par conséquent, les néonymes, appartenant au discours spécialisé, favorisent la formation syntagmatique, et évitent la synonymie en respectant le principe de l'unité notionnelle : à une notion correspond une seule dénomination dans un domaine de spécialité donnée. Ils sont alors monoréférentiels.

Les langues de spécialité disposent de différents moyens pour la création lexicale, du recours aux formants gréco-latins, réservant une place de choix aux éponymes, à la métaphore et la métonymie, en passant par l'emprunt et le développement des termes syntagmatiques. Cette création lexicale permet soit d'obtenir une nouvelle forme lexicale dans la langue (néonymie de forme), soit d'établir une forme lexicale déjà existante avec laquelle elle n'avait auparavant aucune relation, c'est-à-dire, réutiliser une forme existante (néonymie de sens).

2.2.2 Types de néologies

La plupart des linguistes classent les néologies en trois types :

« la trichotomie classique : formel, sémantique, emprunt

[..] *la néologie formelle (i-e la création d'un signifiant non attesté dans un état immédiatement antérieur de la langue, quelle que soit la dénomination adoptée dans tel ou tel classement) et la néologie sémantique (i-e un nouveau sens pour une lexie dont le signifiant existait déjà avec un autre signifié). Les typologies qui sont trichotomiques (elles sont plus nombreuses) ajoutent l'emprunt à ces deux classes. » (Sablayrolles, 1996)* [194]

A chaque type de néologie correspond des procédés spécifiques :

- la néologie de forme est obtenue par les procédés suivants: la dérivation, la composition, la syntagmatique et par les procédés de troncation ou réduction,

[193] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation
[194] Néologismes : une typologie des typologies, p. 26

- la néologie de sens ou sémantiques est obtenue soit par l'extension du signifié de la forme de base, soit par sa restriction ou par le changement du signifié de la forme de base,
- la néologie d'emprunts qui comprend les emprunts proprement dits et les calques.

Mais certains linguistes, parmi lesquels Cabré (1998)[195], distinguent quatre types en ajoutant, pour sa part, la néologie de fonction obtenue par changement de la catégorie grammaticale ; d'autres proposent d'autres distinctions encore :

- *« M. Fuchs (1911) ajoute les archaïsmes,*
- *J.M. Gautier (1952) les mots rares, les termes provinciaux et les archaïsmes,*
- *M. Riffaterre (1953) les mots qui réapparaissent (et qui ne sont pas des archaïsmes),*
- *J. Dubois (1962) les sigles et les abréviations,*
- *J.C. Corbeil (1971) les créations ex nihilo et les changements de catégorie grammaticale,*
- *L. Guilbert les néologismes phoniques et graphiques,*
- *Goose (1975) les créations ex nihilo, les abrègements, et les emplois nouveaux (au sein d'une rubrique fourre-tout "autres procédés"),*
- *Rey (1976) les anciens noms propres et les sigles (ajoutés aux emprunts et néologismes de sens dans la sous-classe des mots morphèmes qui, avec les mots complexes et certains syntagmes, constituent les unités lexicales),*
- *J. Picoche l'abréviation d'un mot savant, la lexicalisation d'un sigle, la dérivation et composition savante,*
- *Hagège (1983) les mots factices,*
- *J. Tournier (1985) les néologismes purement morphologiques tels que les réductions de signifiant par aphérèse, apocope, siglaison ». (Sablayrolles, 1996)*[196]

Concernant l'arabe, une variante de cette typologie peut être proposée en prenant en compte son propre procédé de formation. Plusieurs terminologues, notamment Kettani-Idrissi (1987)[197], ont proposé des typologies des néologies adaptées à l'arabe en suivant l'approche traditionnelle qui s'articule autour des procédés qui sont :

- La dérivation (الاشتقاق) et la composition (نحت) : néologie de forme ;
- Le transfert de sens (مَجاز) : néologie de sens ;

[195] La terminologie - Théorie, méthode et applications, p. 256
[196] Néologismes : une typologie des typologies, p. 27
[197] La néologie arabe, problèmes et perspectives

- L'emprunt (تعريب) : néologie d'emprunt.

D'autres typologies, différente de l'approche traditionnelle, ont vu le jour, notamment celle de Ghazi (1987)[198] qui considère les procédés néologiques en deux types, les procédés morphosyntaxiques et les procédés morphosémantiques (cf. Figure 14 : Typologie des néologies en arabe selon Ghazi) :

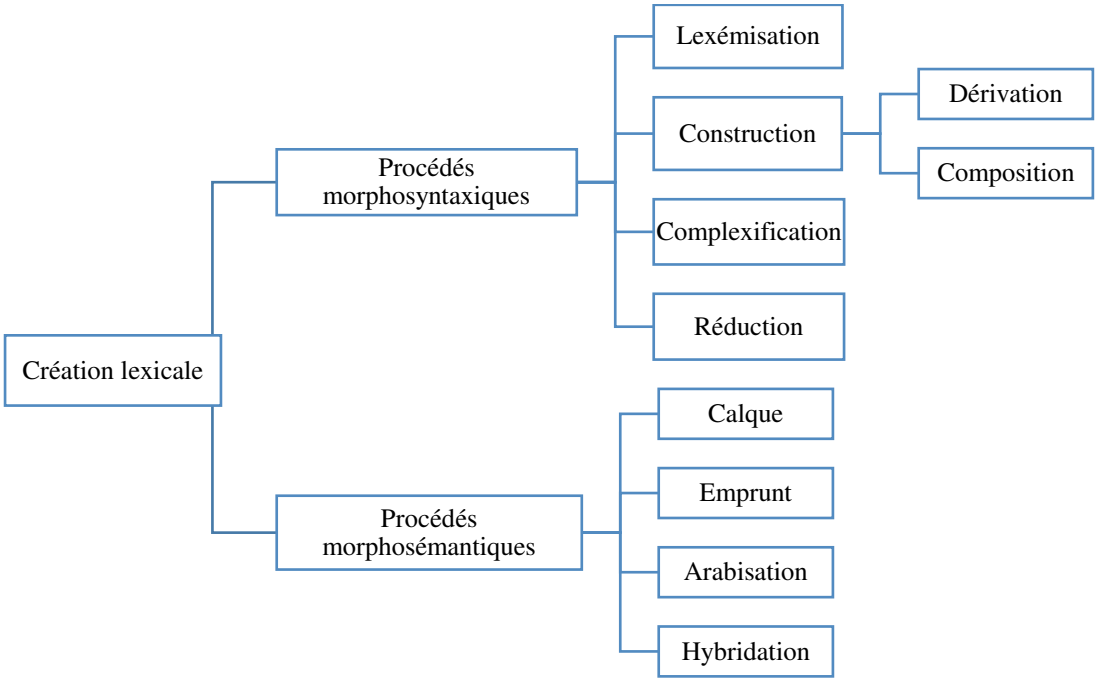


Figure 14 : Typologie des néologies en arabe selon Ghazi

Cette typologie a été appliquée pour le domaine de la chimie en arabe (Jaber, 2012)[199].

Notons qu'aujourd'hui, les linguistes proposent de redéfinir les procédés de création des termes arabes contemporains, notamment avec les travaux de Dichy sur la notion de néologie stratifiée :

> *« La stratification historique du vocabulaire arabe définit un mode particulier de néologie […] la néologie stratifiée. Cette dernière se distingue du processus néologique habituellement décrit, qui consiste en une reprise d'un vocable, souvent tombé en désuétude dans son acception originelle, en conservant certains sèmes, à l'exclusion de certains autres, et en en modifiant d'autres. La néologie stratifiée comporte deux traits définitoires fondamentaux :*
>
> *a) La conservation du sens ancien des vocables à côté de leur sens moderne : on pourrait dire que les entrées du lexique arabe admettent, […] une relation entre leurs acceptions médiévales et leurs*

[198] Propositions pour une typologie nouvelle de la création lexicale, p. 149
[199] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours, p. 179

acceptions contemporaines, que l'on pourrait qualifier de cumulative ou stratifiée. […] Par contraste, les processus néologiques généralement cités dans d'autres langues entraînent le plus souvent un effacement des acceptions anciennes. […] ;

b) une interpénétration potentielle des sens ancien et moderne, corollairement au trait précédent. […] Conservatrice des sens anciens des vocables dans un présent qui en propose simultanément une réinterprétation, la néologie stratifiée inscrit ainsi dans la langue arabe la fidélité contradictoire du lexique à la longue mémoire des textes. » (Dichy, 2001)[200]

Cette proposition concerne la langue générale.

Pour notre travail de recherche, nous adoptons l'approche traditionnelle de la typologie des néologies, à savoir, la néologie de forme et la néologie de sens.

2.2.2.1 Néologie de forme

La néologie de forme est obtenue généralement par les procédés de dérivation et de composition mais également par la siglaison et l'emprunt.

2.2.2.1.1 Dérivation

La dérivation est une création lexicale consistant à ajouter au radical soit un préfixe, soit un suffixe, soit les deux ; ceci effectuera le changement de forme, de catégorie grammaticale (catégoriseurs) ou de sens (modificateurs). Les éléments ajoutés ne sont pas autonomes : ils ne peuvent pas exister comme des unités sémantiques à part. La dérivation est aussi appelée affixation considérant que les morphèmes liés, préfixes et suffixes, portent le nom d'affixes :

« (…) il ne faut pas oublier que ce ne sont pas des abstractions, basées sur des parties de mots liées, antérieurs ou postérieurs. Les affixes n'ont une existence véritable que moyennant les mots dérivés, dont ils sont les éléments à occurrence répétée. » (Kocourek, 1991)[201]

En arabe, l'« اشتقاق = *'ištiqâq* = dérivation » permet la « création de nouveaux mots - donc n'existant pas dans le stock lexical - en recourant au qiyâs à partir des racines et des schèmes arabes connus ». (Lelubre, 1992)[202]. Ce procédé est considéré comme étant le plus grand et le plus important moyen d'enrichissement lexicographique et le terme arabe formé par « اشتقاق = *'ištiqâq* = dérivation » a l'avantage d'être court et dérivable. Il permet le développement de la terminologie arabe, en garantissant « une stabilité et une pérennité sans pareille à la langue,

[200] La néologie stratifiée de l'arabe, fidélité contradictoire du lexique à sa mémoire, à partir du champ sémantique peuple-patrie-nation, p. 13

[201] La langue française de la technique et de la science. Vers une linguistique de la langue savante

[202] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 502

mais ne favorise pas les apports extérieurs » (Lachheb, 2007)[203]. En effet, certains le considèrent comme « une technique peu efficace puisque la nature de la langue arabe impose des règles fixes et des formes syntaxiques qui ne peuvent pas être modifiées ou ignorées facilement pour couvrir le flot de synonymes » (Hudrisier et al., 2017)[204].

La dérivation se fait généralement de deux manières :

> *« En arabe, la dérivation se fait, généralement, par le jeu de la flexion interne, parfois au moyen de la flexion externe, comme c'est le cas dans le duel, le pluriel dit "sain" et le nom de relation ». (Hamzé, 1998)*[205]

Pour notre travail, nous avons abordé la dérivation interne (cf 2.1.2 Système de nomination de l'arabe) et à présent, nous analysons la dérivation externe (affixation) qui est composée de la suffixation et de la préfixation.

2.2.2.1.1.1 Préfixation

La préfixation est l'ajout d'un préfixe au radical. En chimie et plus particulièrement dans la nomenclature chimique, ce procédé est très employé notamment pour les fonctions des préfixes, en commençant par ceux qui désignent des atomes, des ligands ou des groupes substituants comme 'chloro' ou 'amino' et en finissant par ceux qui se rapportent, soit au nombre de groupes ajoutés comme 'méthyl' ou 'éthyl' (préfixe numérique), soit au nombre d'une même molécule ajoutée comme 'di' ou 'tri' (préfixe multiplicatif) et en passant par ceux qui indiquent l'addition d'atomes et de groupes comme 'diméthyl' (préfixe additif) ou l'absence d'atomes ou de groupes comme 'déhydro' (préfixe soustractif).

À partir de la normalisation de la nomenclature chimique (Leigh et al., 2001)[206], les préfixes peuvent être classés selon trois catégories :

- Les préfixes multiplicatifs, d'origine grecque majoritairement mais parmi eux quelques-uns latins, sont employés lorsque le composé chimique ou le groupe chimique possèdent plusieurs substituants identiques. Deux types de préfixes existent selon si le constituant est simple, préfixes de base (di, tri, tétra...) ou complexe, préfixes modifiés (bis, tris, tétrakis...). Ces préfixes multiplicatifs permettent leurs propres modifications évitant ainsi l'ambiguïté des dénominations des composés chimiques comme le 'bis(phosphate)'. Les préfixes multiplicatifs latins (bi, ter, quater …) sont employés

[203] Lexique du commerce électronique Anglais - Français – Arabe

[204] Normalisation de la langue et de l'écriture arabe : enjeux culturels régionaux et mondiaux

[205] De la racine au mot ou du mot à la racine : problématique de la création d'une nouvelle mémoire de l'emprunt en arabe », p. 62

[206] Principes de nomenclature de la chimie. Introduction aux recommandations de l'IUPAC, p. 73

pour les composés cycliques afin de décrire l'assemblage identique des cycles comme le 'biphényle'. Généralement dans la nomenclature chimique, le préfixe grec 'mono', indiquant l'unicité d'un atome ou d'un groupe, est omis considérant que, par défaut, un atome seul ou un groupe seul existe comme 'l'oxyde de fer' à la différence du 'monoxyde d'azote'.

- Les préfixes non-séparables, d'origine grecque ou latine, décrivent le changement de l'enchaînement des atomes de carbone d'un composé chimique comme la conversion d'une chaîne en un cycle (cyclo dans cyclohexane). Certains de ces préfixes non-séparables indiquent la place des groupes 'alkyles' dans un composé chimique permettant de décrire le changement de la chaîne simple des groupes 'alkyles' en une chaîne ramifiée tels que 'iso' (préfixe grec signifiant égal, désignant deux groupes méthyle substitués au carbone le plus éloigné de la chaîne principale) dans 'isobutyle' et 'tert' (préfixe anglais signifiant tertiaire, désignant deux groupes méthyle substitués au carbone le plus près de la chaîne principale) dans 'tert-pentyle'. D'autres préfixes non-séparables sont employés pour signaler soit le manque d'un élément dans la nomenclature soustractive (déhydro) soit l'adjonction d'un élément dans la nomenclature additive (diméthyl). Quelques préfixes non-séparables spécifiques sont obligatoires dans le cas où un des atomes de carbone est remplacé par un hétéroatome comme le 'phosphabenzène', les préfixes dépendant du degré d'oxydation. Pour finir, les préfixes multiplicatifs et les préfixes numériques (di, éthyl) sont également des préfixes non-séparables.
- Les préfixes séparables, issus des noms des substituants, sont aussi nombreux qu'il existe de groupes caractéristiques en chimie ; par exemple le préfixe 'hydroxy' correspondant à la présence de la fonction alcool dans un composé chimique, comme 'hydroxy-4'acétanilide', plus connu sous le nom de 'paracétamol'.

Dans la chimie en arabe, la nomenclature ressemble étroitement à celle que nous venons de présenter mais avec des caractères arabes : il s'agit d'un calque du français et/ou de l'anglais, comme « ثلاثي إيثيل أمين = *ṯulâṯî ʾîṯîl ʾ*amîn = triéthylamine ». En effet, l'arabe est une langue très pauvre en formants, au contraire du français et de l'anglais, langues indo-européennes, où les unités de nomination sont formées par agencement d'affixes ou de formants avec des radicaux syllabiques. Par conséquent, l'arabe emploie des préfixes ou des formants qui sont directement

empruntés au français et à l'anglais (généralement formants d'origine gréco-latine) ; on parle alors de préfixes et de formants antéposés exogènes (Lelubre, 1992)[207].

Pour notre travail de recherche, nous distinguons plusieurs types de préfixes arabes de la nomenclature chimique en analysant leurs traductions et en les modélisant. Nous présentons dans un tableau les préfixes principaux de la nomenclature chimique en français et en arabe.

- Préfixe numérique : en arabe, ils sont exprimés par leur emprunt grec et/ou latin comme « بنتان = bantân = pentane » où 'pent' est un préfixe grec, indiquant la présence du nombre d'atome de carbone dans la molécule, notre exemple contenant 5 atomes. Voici une liste non-exhaustive des préfixes numériques indiquant le nombre de carbones (cf. Tableau 2 : Préfixes numériques de la nomenclature chimique). Notons que pour désigner respectivement 'un', 'deux', 'trois' et 'quatre' atomes de carbone dans une molécule, les chimistes emploient respectivement 'méth', 'éth', 'prop' et 'but' qui ne sont pas des préfixes grecs et/ou latins mais des inventions (Leigh et al., 2001) :

Nombre d'atome de carbone	Préfixe arabe	Préfixe français
5	بنتا	Pent
6	هكسا	Hex
7	هبتا	Hept
8	أوكتا	Oct
9	نونا	Non
10	ديكا	Déc
16	هكساديكا	Hexadéc

Tableau 2 : Préfixes numériques de la nomenclature chimique

Ce préfixe numérique, indiquant le nombre de carbone dans une molécule, est la forme minimale de la dénomination d'un composé chimique, constitué d'au moins un atome de carbone, auquel est ajouté d'autres préfixes et/ou suffixes comme hexane (hex+ane) ; ces préfixes numériques jouent alors le rôle de radical (cas en arabe issu du français et/ou de l'anglais).

- Préfixe multiplicatif : en arabe, ils sont exprimés par leurs équivalents arabes comme « ثنائي كلوروميثان = *ṯunâʾî klûrûmîṯân* = dichlorométhane », où le préfixe 'di' indique le nombre de substituants identiques dans la molécule, notre exemple contenant deux atomes de chlore. Voici une liste non-exhaustive des préfixes multiplicatifs indiquant

[207] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation

le nombre substituant identique dans une molécule (cf. Tableau 3 : Préfixes multiplicatifs de la nomenclature chimique) :

Nombre de substituant identique	Préfixe arabe	Préfixe français
2	ثنائي / ثنائية	Di
3	ثلاثي	Tri
4	رباعي	Tétra

Tableau 3 : Préfixes multiplicatifs de la nomenclature chimique

Ce préfixe multiplicatif, syntaxiquement autonome, peut être considéré comme un formant indiquant le nombre de substituant identique dans une molécule.

- Préfixe géométrique : en arabe, ils sont exprimés par leur emprunt grec et/ou latin comme « سيكلوهكسان = sîklûhaksân = cyclohexane » où 'cyclo' est un préfixe grec, indiquant la position des atomes dans la molécule, notre exemple formant un cycle. D'autres préfixes géométriques sont traduits par leurs équivalents arabes comme « مفروق هكسادسنال = mafrûq haksâdasanâl = trans hexadécénal » où le préfixe 'trans' indique la position du substituant dans la molécule, notre exemple informant que le substituant est à l'opposé du composé. Voici une liste non-exhaustive des préfixes géométriques indiquant la position des atomes dans une molécule (cf. Tableau 4 : Préfixes géométriques de la nomenclature chimique) :

Préfixe arabe	Préfixe français
سيكلو	Cyclo
نيو	Néo
إيزو	Iso
ن-	N-
مفروق	Trans
مقرون	Cis
بارا	Para

Tableau 4 : Préfixes géométriques de la nomenclature chimique

Les préfixes géométriques cis et trans, syntaxiquement autonome, peuvent être considérés comme des formants indiquant la structure formée par les atomes dans le composé chimique.

- Préfixe fonctionnel : en arabe, ils sont exprimés par leur emprunt français et/ou anglais comme « هيدروكسي بوتانون = hîdrûksî bûtânûn = hydroxybutanone », où le préfixe 'hydroxy' indique la présence d'un groupe fonctionnel dans le composé chimique, notre exemple correspondant à la fonction alcool. Voici une liste non-exhaustive des préfixes fonctionnels indiquant les groupes fonctionnels présents dans une molécule (cf.

Tableau 5 : Préfixes fonctionnels de la nomenclature chimique). Notons qu'une molécule chimique peut contenir zéro, un ou plusieurs groupes fonctionnels :

Groupe fonctionnel	Préfixe arabe	Préfixe français
Alcool	هيدروكسي	Hydroxy
Acide carboxylique	كربوكسي	Carboxy
Amine	أمينو	Amino
Composé nitré	نيترو	Nitro
Composé halogénure	فلورو كلورو برومو يودو	Fluoro Chloro Bromo Iodo

Tableau 5 : Préfixes fonctionnels de la nomenclature chimique

Ce préfixe fonctionnel, syntaxiquement autonome, peut être considéré comme un formant indiquant la présence d'un groupe fonctionnel dans une molécule. Il y a autant de formants fonctionnels qu'il existe de groupes fonctionnels en chimie.

Pour les autres termes de la chimie en arabe, nous n'identifions ni des préfixes et des formants antéposés exogènes ni des préfixes et des formants antéposés endogènes dans notre modeste corpus. D'ailleurs, l'arabe ne dispose que d'un seul préfixe endogène ; il s'agit du préfixe de négation « لا = lâ » (Lelubre, 1992)[208] . Par exemple, « هدرجة لامتناظرة = hadrajat lâmutanâẓirat = hydrogénation asymétrique ».

2.2.2.1.1.2 Suffixation

La suffixation est l'ajout d'un suffixe au radical. En chimie, mais également dans la nomenclature chimique, ce procédé est très employé notamment pour la fonction attribuée aux suffixes, en commençant par ceux qui désignent la charge d'un élément comme 'ure' ou 'ium' et en finissant par ceux qui se rapportent au groupe fonctionnel comme 'ol' ou 'one' et en passant par ceux qui indiquent l'absence d'atomes ou de groupes comme 'yle' (suffixes soustractif) ou ceux qui renseignent le nombre de groupes ajoutés comme 'di' (suffixe numérique).

Dans la nomenclature chimique (Leigh et al., 2001)[209], ces suffixes permettent :

[208] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 189

[209] Principes de nomenclature de la chimie. Introduction aux recommandations de l'IUPAC, p. 79

- D’identifier le groupe fonctionnel auquel appartiennent les composés chimiques en distinguant les différentes fonctions chimiques comme la fonction chimique alcool pour ‘l’éthanol’, indiquée par le suffixe ‘ol’. Un composé chimique peut contenir, soit un seul groupe fonctionnel, composé monofonctionnel, c’est le cas de ‘l’éthanol’, soit plusieurs groupes fonctionnels, composé polyfonctionnel, comme ‘l’aminobutanol’, où intervient un préfixe fonctionnel en plus du suffixe indiquant la présence des fonctions chimiques ‘amine’ (amino) et ‘alcool’ (ol).
- De diversifier les composés chimiques en indiquant le degré d’oxydation de valence comme ‘l’hexacyanoferrate III de potassium’, en distinguant les anions des cations comme le ‘chlorure d’ammonium’ et en différenciant les composés saturés des insaturés comme ‘éthane’ de ‘éthylène’.

Grâce au procédé de la dérivation, procédé le plus productif du point de vue de la linguistique et de l’enrichissement lexical, les suffixes forment de nouveaux mots à partir de morphèmes catégoriseurs en fonction du nom (changement de la catégorie grammaticale de la base dérivationnelle) et de morphèmes modificateurs en fonction d’adjectif et de nom (changement du sens de la base dérivationnelle), constituant toute une famille de mots, par exemple le terme ‘acide’ comme nom et adjectif, ‘acidité’, ‘groupe acide’...

Voici les suffixes les plus fréquents en chimie :

- tion, age, lyse : suffixes catégoriseurs en fonction de nom permettant d’obtenir un nom d’action ; cela correspond en chimie à des processus et des réactions comme absorption ou craquage et alkylation ou hydrolyse ;
- eur, ant, ent, ique : suffixes catégorisateurs en fonction du nom permettant de former les noms d’agent ; cela correspond en chimie aux corps participant à une réaction ou à un processus chimique, par exemple, catalyseur et solvant ; mais les suffixes ique et ent peuvent spécifier le type d’agent comme antibiotique ou détergent et ce dernier suffixe ent peut aussi former les noms de processus chimiques ou des propriétés des substances comme blanchiment ou rayonnement ;
- on, ode, tope : suffixes catégoriseurs en fonction du nom formant les noms des particules liés à l’atome et ou au courant électrique, par exemple, anion et cation, cathode, isotope ;
- eté, ité : suffixes catégoriseurs en fonction du nom formant les noms des propriétés de la matière, par exemple, molarité et dureté ;

- ium : suffixe catégoriseur en fonction du nom que possède la majorité des noms des éléments du tableau périodique;
- ite, ate : suffixes modificateurs en fonction du nom permettant d'indiquer le degré d'oxydation le plus bas ou le plus haut de l'élément principal du composé chimique comme 'chlorite de sodium' ou 'chlorate de soude' ;
- eux, ique : suffixes modificateurs en fonction de l'adjectif permettant d'indiquer l'oxydation la moins élevée ou la plus élevée de l'élément chimique comme 'fer ferreux' ou 'fer ferrique'.

A cela, il faut ajouter les suffixes savants, d'origine grecque ou latine comme métrie (mesure) dans 'colorimétrie' et phobe (crainte) dans 'hydrophobe'

Dans la chimie en arabe, nous avons déjà vu (cf. 2.2.2.1.1.1 Préfixation) que la nomenclature chimique est un calque du français et/ou de l'anglais et que l'arabe est une langue très pauvre en formants. Par conséquent, l'arabe emploie des suffixes ou des formants qui sont directement empruntés au français et à l'anglais (généralement formants d'origine gréco-latine) ; on parle alors de suffixe et de formants postposés exogènes (Lelubre, 1992)[210].

Pour notre travail, nous distinguons plusieurs types de suffixes arabes de la nomenclature chimique en analysant leurs traductions et en les modélisant. Nous présentons dans un tableau les suffixes principaux de la nomenclature chimique en française et en arabe.

- Suffixe de saturation : en arabe, ils sont exprimés par leurs emprunts français et/ou anglais comme « بوتان = bûtân = butane », où le suffixe 'ane' indique la présence d'un groupe spécifique dans le composé chimique, notre exemple correspondant aux alcanes. Voici une liste non-exhaustive des suffixes de saturation (cf. Tableau 6 : Suffixes de saturation de la nomenclature chimique) :

Classe	Suffixe arabe	Suffixe français
alcane	ان	ane
alcène	اين	ène
alcyne	اين	yne

Tableau 6 : Suffixes de saturation de la nomenclature chimique

Ces suffixes de saturation, différenciant les composés saturés des insaturés, ne sont pas autonomes et sont accompagnés d'un radical désignant le nombre de carbone et/ou sont précédés d'un ou plusieurs préfixes. Par exemple, le suffixe des alcènes et des alcynes est

[210] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation

considéré le même en arabe, en raison de la difficulté de transcrire certains phonèmes en arabe (Jaber, 2012)[211]. Cela implique une ambiguïté supplémentaire dans la compréhension, puisque les alcènes et les alcynes sont deux catégories différentes des composés organiques.

- Suffixe fonctionnel : en arabe, ils sont traduits par leur emprunt français et/ou anglais comme « إيثانول = *ʾîṯânûl* = éthanol », où le suffixe 'ol' indique la présence d'un groupe fonctionnel dans le composé chimique, notre exemple correspondant à la fonction alcool. Voici une liste non-exhaustive des suffixes fonctionnels indiquant les groupes fonctionnels présents dans une molécule (cf. Tableau 7 : Suffixes fonctionnels en arabe et en français). Notons qu'une molécule chimique peut contenir zéro, un ou plusieurs groupes fonctionnels :

Classe	Suffixe arabe	Suffixe français
alcool	اول	ol
acide carboxylique	اويك	oïque
cétone	اون	one
aldéhyde	ال	al
amine	امين	amine
amide	اميد	amide

Tableau 7 : Suffixes fonctionnels en arabe et en français

Certains suffixes fonctionnels, syntaxiquement autonome, peuvent être considérés comme des formants indiquant la présence d'un groupe fonctionnel dans une molécule. Il y a autant de suffixes et de formants fonctionnels qu'il existe de groupes fonctionnels en chimie.

Par exemple, le terme « حمض –2 أمينو –3 هيدروكسي بوتان ثنائي أويك = acide 2-amino-3-hydroxybutanedioïque » :

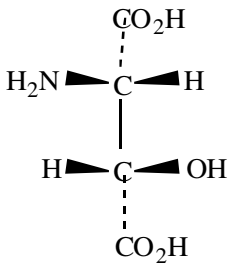


Figure 15 : Représentation écrite de la molécule d'acide 2-amino-3-hydroxybutanedioïque

[211] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours, p. 192

Cette molécule contient quatre atomes de carbone, un groupe amine sur le deuxième carbone, un groupe alcool sur le troisième carbone et deux groupes acides carboxyliques. La fonction principale est l'acide carboxylique ; le suffixe est 'oïque'. Mais, puisqu'il y a deux groupes acides carboxyliques, le suffixe est alors 'dioïque'. En préfixe, nous avons l'acide suivit par les préfixes des fonctions chimiques par ordre alphabétique du groupe amine 'amino' et alcool 'hydroxy' précédé de l'emplacement de ces derniers à l'aide de la numérotation des carbones qui sont respectivement '2-amino' et '3-hydroxy'. La structure de base est une chaîne de quatre atomes de carbone liés par des liaisons simples ; il s'agit d'un butane.

Pour les autres termes de la chimie en arabe, les suffixes lexicaux sont limités par rapport aux suffixes grammaticaux (modalités de genre, de nombre, de cas) ; ce sont des suffixes endogènes (Lelubre, 1992)[212].

- Suffixe î : il correspond au suffixe nominal de relation « ياء النسبة = yâ' an-nisba » ; le suffixe « يّ = î » est ajouté à un nom pour exprimer dans un adjectif l'origine, la matière ou la relation. Par exemple, « كيميائي = kîmya î = chimique ».

- Suffixe iyyat : il correspond au ma**s**dar **s**inâ`î de la grammaire arabe (Roman, 1999)[213] ; le suffixe « يّة = iyyat » est ajouté à un nom pour désigner généralement des noms abstraits (théories, courants d'idées ; pratiques ; phénomènes). Par exemple, « حمضية = *ḥ*am*ḍ*iyyat = acidité ».

2.2.2.1.2 Composition

La composition est le deuxième procédé de création lexicale, aussi important que la dérivation, consistant à créer une nouvelle unité lexicale, grâce à l'association de deux ou plusieurs lexèmes autonomes, comme le 'dioxyde de carbone', en les complétant encore par d'autres morphèmes lexicaux ou grammaticaux, comme 'l'alcool éthylique' dont le sens résultera de la fusion des significations de tous ses composants ; elle permet d'obtenir plus d'informations sous forme condensée et motivée, même si parfois cela n'est forcément pas explicite comme le 'tétrahydrofurane'. La composition se distingue de la dérivation par l'emploi de lexèmes autonomes, séparés ou dans un mot composé ; mais parfois elle est confondue, notamment dans certains cas de la dérivation par préfixation où les préfixes, d'origine grecque et/ou latine, ont le rôle de formant comme le 'trans hexadécénal'.

[212] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation
[213] La création lexicale en arabe, p. 58

Pour l'arabe, il y a différents avis (Diab-Duranton, 2015)[214] ; mais, nous référons à Lelubre (1992) qui a réalisé ses travaux sur la terminologie de la physique en arabe :

> *« Le نحت (composition à partir de mots tronqués) a ses défenseurs (les Anciens y ont eu recours ; les termes créés de cette manière sont "maniables" : ils sont relativement courts et ils sont dérivables puisqu'on peut facilement former à partir d'un tel terme, l'adjectif correspondant et un verbe, en en tirant une racine) mais de manière générale, il est rejeté (senti comme contraire au "génie de la langue arabe", il est difficilement soumis au قياس, les règles de formation de ces composés n'étant pas clairement définies ; il est de ce fait peu transparent.) On n'y recourt que contraint et forcé - c'est aussi une affaire de "bon goût" (الذوق السليم) - notion à vrai dire peu scientifique - et de fait, c'est l'usage qui tranche. » (Lelubre,1992)[215]*

En effet, la composition n'est pas propre à la langue arabe et un moyen de création lexicale peu utilisé ; mais, elle peut jouer un rôle intéressant dans l'affixation des termes étrangers et servir ainsi de mécanisme d'abréviation des termes arabes, comme c'est le cas du domaine de la chimie.

Ce procédé dépend et résulte des champs conceptuels du domaine qui le concerne ; de plus, le langage spécialisé est le plus souvent caractérisé par les modèles de formation du type composition syntagmatique. Pour notre corpus, il s'agit plutôt de composition savante que la composition syntagmatique qui est d'ailleurs confondu avec la dérivation ; par exemple, « مفروق هكسادسنال = mafrûq haksâdasanâl = trans hexadécénal ». Mais, cette composition syntagmatique est bien présente dans le domaine de la chimie en arabe, avec ou sans altération phonétique, comme dans respectivement « كهرجابية = kahrajâbiyya = électropositivité » et « أكسدة إرجاع = ʾaksada ʾirjâʿ = oxydo-réduction » (Jaber, 2012)[216]. Le premier exemple correspond en l'amalgame de deux formes existantes constituant une UTS et est appelé aussi mot-valise (apocope du premier élément et aphérèse du second) ; quant au second exemple, il s'agit d'une UTC avec une extension de coordination par le coordonnant "zéro" (cf. 2.1.3.2.1 Extension par coordination).

2.2.2.1.3 Siglaison

La siglaison est un procédé qui consiste à abréger des syntagmes, en gardant soit les lettres initiales des éléments constituants de ce syntagme, soit les premières syllabes (acronymes), soit encore d'autres syllabes. Ce sont des signes nouveaux, souvent très vite

[214] Substitution et créativité lexicales en arabe
[215] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation
[216] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours, p. 182

lexicalisés mais qui n'apportent pas plus de sens que leur forme développée. Ces abréviations sont très répandues dans les domaines de spécialité ; elles sont le produit de la tendance à l'économie linguistique au moment où le concept et le terme qui le désigne ont acquis une stabilité sûre facilitant leur emploi généralisé dans le discours.

Ce mode de création lexicale n'est pas très fréquent en arabe ; mais, puisque la terminologie scientifique et technique arabe est tributaire de celle établie en anglais et/ou en français, alors des abréviations sont employées. Elles sont majoritairement avec des caractères latins, mais parfois avec des caractères arabes. Par exemple, « LDA = diisopropylamidure de lithium » est un composé très utilisé en chimie organique pour favoriser les réactions d'élimination ; sa forme abrégée est largement employée, facilitant sa mémorisation et sa présentation.

À cela, il faut ajouter que le domaine de la chimie est caractérisé par l'emploi récurent de brachygraphies (Kaczmarek, 2011)[217] ; ce sont des expressions généralement non linguistiques, ou des unités qui ne sont pas totalement articulées, composées d'éléments qui peuvent être, notamment des combinatoires de lettres, de chiffres et de symboles ou encore se combiner avec des lexèmes pleins (Kocourek, 1982)[218]. Ces formes graphiques abrégées sont utilisées dans les textes de spécialité, notamment scientifiques et techniques, à côté des unités terminologiques. En chimie, elles correspondent aux symboles, aux formules internationaux et aux équations ; elles sont généralement normalisées comme « g pour gramme ». Mais, parfois, il s'agit de notation relevant de conventions implicites ou pas « R_f = rendement final ». Dans notre corpus, ces signes brachygraphiques sont soit repris tels quels, c'est-à-dire, les signes internationaux tels quels avec l'alphabet latin, et éventuellement grec comme 'ppm', soit ils sont arabisés comme « مل = ml ».

2.2.2.1.4 Emprunt

L'emprunt est un procédé néologique qui contribue à l'enrichissement lexical dans une langue par l'appropriation d'unités linguistiques d'une autre langue (Depecker, 2002)[219]. Son recourt est la dernière alternative pour la création lexicale, puisque le terme emprunté est souvent l'objet d'essais pour le remplacer par un terme de la langue grâce aux différents procédés néologiques que nous avons vu (Lelubre, 1992)[220] même s'il est très efficace :

[217] Nomenclatures française et polonaise de la chimie organique. Analyse comparative, p. 64
[218] La langue française de la technique et de la science, p. 72
[219] Entre signe et concept : éléments de terminologie générale, p. 115
[220] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 304

« L'emprunt constitue la solution la plus évidente, la plus paresseuse, mais aussi la plus efficace internationalement, car elle neutralise partiellement les différences interlinguistiques et respecte ainsi la notion originelle, plus aisément repérée. On peut dire que l'emprunt dénomme et connote son origine ce qui explique son succès malgré ses inconvénients ». (Rey, 1979)[221]

Son avènement est possible s'il présente entre autres, des 'affinités électives' entre la langue donneuse et la langue emprunteuse ; son utilité et son intégration au sein de la structure de la langue emprunteuse au niveau, notamment phonologique, graphique, morphosyntaxique et sémantique, sont confirmées dès lors qu'il est prêt à former des dérivés et des composés, c'est-à-dire, qu'il s'adapte aux caractéristiques morphosyntaxiques de la nouvelle langue. De plus, la proximité géographique et/ou linguistique mais également les relations politiques et/ou économiques et/ou cultures influencent ce recourt à l'emprunt et notamment sa fréquence (Guilbert, 1975)[222].

L'arabe ne fait pas exception et a recours à l'emprunt et cela n'est pas nouveau pour cette langue, notamment dans les domaines scientifiques ; en effet, l'arabe a souvent emprunté des termes dans le passé et jusqu'à récemment :

- *« Très anciennement au persan, à l'araméen, au grec, sanscrit, etc. ;*
- *À l'époque `abbasside, au grec : emprunts nombreux dans une première phase, remplacés pour la plupart ensuite par des termes arabes (philosophie, sciences) ; termes persans (militaires, techniques) ;*
- *Durant la période dite de Jumûd, emprunts à l'italien, au français, à l'ottoman (turc et persan) ;*
- *Depuis la Nahda : français, anglais ». (Lelubre, 1992)*[223]

C'est un des procédés qui a contribué à la modernisation rapide de la langue arabe et l'assimilation de vocabulaire d'origine étrangère, plus efficace que la dérivation ou la composition.

Pour la chimie en arabe, la grande majorité de ses termes sont empruntés (Jaber, 2012)[224]. En effet, si comme les autres domaines scientifiques, la chimie s'est développée en occident, sa particularité vient de l'élaboration d'un système international pour la dénomination des composés chimiques utilisé dans toutes les langues, notamment l'arabe. Par conséquent, nous

[221] La terminologie : noms et notions, p. 68
[222] La créativité lexicale, p. 68
[223] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 290
[224] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours, p. 243

retrouvons parmi les termes empruntés, des unités comme « مول = mûl = mol », des éléments et des composés chimiques comme « بيرمنغنات البوتاسيوم = bîrma*ngânât albûtâs*yûm = permanganate de potassium », des éponymes comme « قاعدة شيف = qâʿidat *šîf* = base de Schiff » et des affixes comme « ثنائي كلورو ميثان = *ṯunâʾî klûrûmîṯân* = dichlorométhane ».

Nous distinguons trois types d'emprunts :

- Emprunt intégral : le terme étranger est pris tel quel ; il est intégralement emprunté. Par exemple, « أكسجين = uksijîn = oxygène ». Cependant, il existe plusieurs variantes pour certains emprunts en l'absence de règles pour leur transcription en arabe ; en effet, son système phonétique ne possède pas toutes les consonnes et voyelles que possèdent l'anglais et/ou le français à partir desquelles les termes sont empruntés et sont prononcés en fonction de la connaissance de ces langues.
- Emprunt adapté ou intégré : seule une partie du terme étranger est empruntée ; en fait, il subit une troncation par aphérèse ou par apocope et est complété par un affixe de l'arabe (Lelubre, 1992)[225]. Par exemple, « فلورية = flûriyyat = fluorescence ».
- Emprunt partiel : le terme étranger est partiellement emprunté ; il s'agit généralement d'UTC auxquelles sont empruntées une partie des éléments les constituant (Lelubre, 1992)[226]. Par exemple, « كروماتوغرافيا العمود = krûmâtû*ġâfîyâ* alʿamûd = chromatographie sur colonne ».

À partir de là, le terme emprunté est intégré au système de l'arabe et devient le point de départ dans d'une nouvelle racine, souvent quadriconsonantique, qui servira à la création d'autres termes. Par exemple « أكسد = ʾaksada = oxydation ».

La néologie de forme permet la création d'une nouvelle forme lexicale dans la langue ; mais il est possible aussi à partir d'une forme lexicale déjà existante d'envisager cette création lexicale : la néologie de sens.

2.2.2.2 Néologie de sens

La néologie de sens ou sémantique est obtenue soit par l'extension du signifié de la forme de base, soit par sa restriction ou par le changement du signifié de la forme de base ; cela

[225] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 304
[226] Idem, p. 305

s'obtient par divers mécanismes qui donnent lieu aux différents types de figures, notamment la métaphore et la métonymie.

2.2.2.2.1 Métaphore

La métaphore est généralement définie comme « *une figure qui consiste dans l'emploi d'un mot concret pour exprimer une notion abstraite, en l'absen*ce de tout élément introduisant formellement une comparaison » (Dubois, 1994)[227]. Par conséquent, elle correspond à la nomination d'un objet par un autre, lié au premier, par un lieu de similitude et elle s'appuie sur les analogies existant dans le réel ou construites par le sujet parlant.

En langue de spécialité, la métaphore est en priorité considérée comme un procédé néologique qui permet de suppléer à des manques lexicaux à l'intérieur d'un domaine du savoir en faisant notamment appel à des ressources de la langue générale. En effet, ce procédé néologique est reconnu comme efficace pour l'enrichissement lexical, puisqu'il détient le pouvoir de dénommer et de conceptualiser des réalités nouvelles (Assal, 1994)[228]. Par conséquent, la métaphore terminologique est envisagée comme un 'laboratoire privilégié' pour observer les dynamiques à l'œuvre dans les langues spécialisées (Rossi, 2015)[229]. Elle est « un moyen possible de la nomination référentielle, par laquelle elle actualise une coïncidence, partielle, imaginée entre une entité et une unité référentielle et non pas une relation constatée entre elles » (Lelubre, 1992)[230].

Le domaine de la chimie en arabe ne fait pas exception et présente également des métaphores dans son vocabulaire (Jaber, 2012)[231] ; la plupart de ces métaphores sont calquées sur les termes français ou anglais, par exemple, « غاز نبيل= *ġâz* nabîl = gaz noble ». De plus, de nombreuses métaphores de la chimie appartiennent à d'autres disciplines, par exemple « سحابة إلكترونية = sâḥabat ʾiliktrûniyyat = nuage électronique » emprunté au domaine météorologique.

2.2.2.2.2 Métonymie

La métonymie est une figure de rhétorique par lequel un terme est substitué à un autre terme avec lequel il entretient une relation de contiguïté :

[227] Dictionnaire de linguistique et des sciences du langage
[228] La métaphorisation terminologique
[229] Métaphores et termes nomades dans les langues de spécialité
[230] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 268
[231] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours, p. 213

« Le glissement de sens opéré par la figure est déjà expliqué par un glissement de référence entre deux objets qui sont reliés par un rapport extralinguistique, révélé par une expérience commune qui n'est pas liée à l'organisation sémantique d'une langue particulière ». (Le Guern, 1973)[232]

En effet, la métonymie réalise généralement, l'extension au référent entier du nom d'une qualité, d'une forme, d'une fonction de l'une de ses composantes, en saisissant immédiatement le référent qu'elle nomme, par un nom qui, en quelque sorte lui revient en raison de son appartenance à un ensemble. Par conséquent, il s'agit d'une opération de nomination référentielle et non pas linguistique.

Le glissement métonymique de sens se déclenche à partir de plusieurs mécanismes : la partie pour le tout, l'espèce pour l'individu, l'auteur pour l'œuvre, la cause pour l'effet, le singulier pour le pluriel, le signe pour la chose, le physique pour le moral, l'objet propre pour la personne, la personne pour la propriété, le contenant pour le contenu, l'instrument pour l'agent, le lieu pour le produit, la matière pour l'objet, la cause pour l'effet, l'action et le résultat de l'action.

Dans la chimie en arabe, la métonymie est présente (Jaber, 2012)[233] et elle est largement employée comme 'l'action et le résultat de l'action'. Par exemple, « تركيب = tarkîb » qui désigne 'l'action de composer' ou bien 'le résultat de la composition' ; il s'agit du modus informis, le masdar de la Tradition grammaticale arabe (cf. 2.1.2.2 Modus impersonnel) qui contribue à un glissement de sens, du au phénomène de contiguïté conceptuelle (Lelubre, 1992)[234]. La métonymie est également employée comme 'l'auteur pour l'œuvre' appelé l'antonomase, éponyme ou encore la métonymie du nom propre, par exemple « قاعدة شيف = qâʿidat *šîf* = base de Schiff ». Mais, son emploi le plus fréquent est dans la nomenclature chimique désignant 'la partie pour tout' ; cette emploi involontaire de la métonymie rend parfois la communication scientifique ambiguë. En effet, l'élément chimique est souvent employé pour désigner la molécule chimique contenant un seul type d'élément ; généralement, l'élément chimique correspond au composé chimique que l'on trouve à l'état stable à température et pression normale. Par exemple, « أكسجين = *'uksij*în = oxygène » est un élément chimique ; mais, il peut désigner également un gaz et/ou la molécule contenant deux atomes d'oxygène.

[232] Sémantique de la métaphore et de la métonymie, p. 25
[233] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours, p. 203
[234] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation

Dans ce chapitre, nous avons déterminé les moyens dont la langue arabe dispose pour la création et la formation des termes de la chimie.

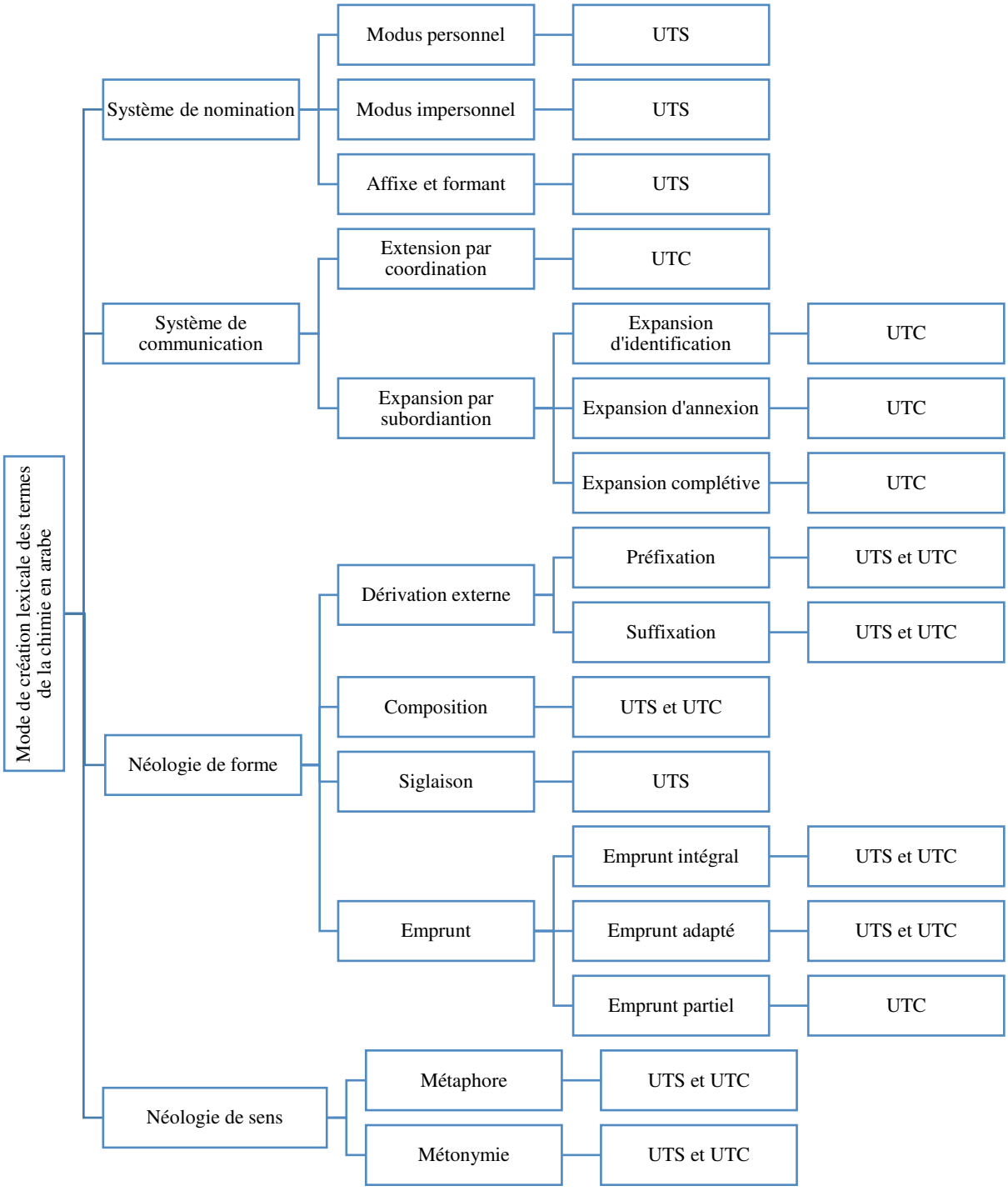


Figure 16 : Mode de création lexicale des termes de la chimie en arabe

Ainsi, nous avons présenté l'étude terminologique de la chimie en arabe qui nous a permis d'identifier les modes de création lexicale de ses termes. À présent, nous pouvons

déterminer leurs patrons morphosyntaxiques et construire les règles de grammaire pour la conception d'un extracteur morphosyntaxique des termes de la chimie en arabe.

Auparavant, nous présentons notre corpus de travail dans la partie suivante.

Partie II : Constitution et dépouillement du corpus de la chimie en arabe, avec sa classification

Qu'il s'agisse d'un travail de fouille de textes ou d'un travail terminologique, l'utilisation d'un corpus est centrale, tant pour l'élaboration d'hypothèses que leur validation, comme dans presque toutes les études de la langue. Les données textuelles de ce corpus définissent la nature des données terminologiques ainsi que le type des traitements informatiques. Ces données peuvent être organisées au sein d'une classification.

Concernant l'arabe, cette langue, complexe tant par sa graphie agglutinante que son absence de voyellation, souffre encore d'un manque d'outils informatiques.

Cette partie est constituée de trois chapitres ; dans le premier, nous présentons la constitution de notre corpus de travail, en expliquant la démarche suivie afin d'obtenir le corpus recueilli. Puis, dans le second chapitre, nous exposons la norme de son dépouillement, en recourant à des outils informatiques afin d'analyser les données et les informations à extraire. Enfin, dans un troisième chapitre, ces données et ces informations sont définies et classées en fonction de leur place dans le domaine de la chimie.

Chapitre 3 : Constitution du corpus

Le point de départ de notre travail de recherche porte sur la constitution d'un corpus. Pour mettre au point ce corpus, la démarche suivie soulève un certain nombre de problématiques, tant sur la méthodologie de collecte de textes que sur les caractéristiques du corpus recueilli. Dans cette perspective, nous nous intéressons à la notion de corpus et nous étudions les éléments essentiels qui font d'un recueil de données un corpus.

3.1 Définition du corpus

Un corpus est la base du travail de fouille de textes. Il est défini comme étant 'un ensemble déterminé de textes sur lesquels on applique une méthode définie' (Dubois, 1969)[235]. En effet, un corpus permet de repérer les caractéristiques d'une langue. Il est considéré aussi bien comme une source fournissant des preuves lexicales que comme un moyen de connaissance permettant de distinguer les utilisateurs des communautés de locuteurs. Ainsi, le corpus est utilisé comme une référence de travail :

> *« Une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage. » (Sinclair, 1996)*[236]

Un corpus est appelé aussi 'documentation', 'données' ou 'sources'. Par conséquent, nous pourrions dire qu'un corpus serait un assemblage homogène et significatif de données linguistiques permettant d'élaborer la description et la formalisation des faits linguistiques. Parmi les nombreuses définitions existantes du corpus en linguistique, celle que Sinclair propose dans son rapport est abondamment citée ; mais, Habert suggère quelques précisions :

> *« Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extralinguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue ». (Habert, 2000)*[237]

Cette définition du corpus nous parait complète, mais celle de Rastier (2005)[238], qui dans le cadre de la sémantique textuelle interprétative, parle « *d'*un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de

[235] Lexicologie et analyse d'énoncé, p. 115
[236] Preliminary Recommendations on Corpus Typology, p. 4
[237] Des corpus représentatifs : de quoi, pour quoi, comment ?, p. 1
[238] Enjeux épistémologiques de la linguistique de corpus, p. 31

manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière *pratique en vue d'une gamme d'applications* » est tout aussi plausible. En fait, il existe une multitude de définitions du corpus dont nous ne livrerons pas ici la liste ; elles divergent plus ou moins d'un chercheur à un autre, en fonction de son ancrage théorique ou disciplinaire.

En ce qui nous concerne, nous considérons le 'corpus' comme une grande collection de textes authentiques, mémorisés sous forme électronique, rassemblés selon un ensemble de critères spécifiques avec un objectif d'étude précis, permettant d'attester les différentes dénominations d'usage de notre domaine de spécialité.

3.1.1 Historique de la notion de corpus

Les premières études menées sur le 'corpus' remontent à la fin du 19ème siècle, notamment avec des travaux en acquisition du langage (Mc Enery et al., 2001)[239]. Vers 1920, en Angleterre, Palmer et Hornby travaillent sur la phraséologie orale et écrite dans le cadre de l'apprentissage du langage ; ces travaux aboutissent à la rédaction du fameux dictionnaire Learner's Dictionary of Current English. En parallèle, aux Etats-Unis, en 1921, Thornbike a compilé un corpus de 4,5 millions de mots provenant de quarante et une sources distinctes afin de constituer une liste de fréquences. En 1970, Sinclair publie le résultat de ces études sur corpus électronique ; mais ce rapport est peu diffusé et n'est publié que récemment. Et en 1980, le projet COBUILD a pour objectif de créer un corpus d'anglais contemporain pour la composition d'un dictionnaire et d'une grammaire entièrement nouveaux. Cela a permis de rassembler 320 millions de mots formant désormais la Bank of English.

Avec le développement de l'informatique, les corpus électroniques apparaissent, explosent et deviennent le support textuel des chercheurs. Aujourd'hui, pratiquement tous les linguistes ont recours au corpus et les plus farouches iront jusqu'à dire « hors des corpus point de salut » (Cappeau et al, 2007)[240]. La question de l'utilité des corpus est désormais obsolète et fait l'objet d'un accord.

[239] Corpus linguistics : an introduction

[240] L'exploitation sociolinguistique des grands corpus, p. 100

3.1.2 Typologie et caractéristique du corpus

Pour construire un corpus de textes, plusieurs caractéristiques sont à définir comme la taille du corpus ou le type de corpus ; mais, c'est surtout l'objectif de l'étude qui permet de choisir les textes du corpus :

> *« Tout corpus suppose en effet une préconception des applications, fussent-elles simplement documentaires, en vue desquelles il est rassemblé : elle détermine le choix des textes, mais aussi leur mode de « nettoyage », leur codage, leur étiquetage ; enfin, la structuration même du corpus. (Rastier, 2005)*[241]

Pour notre travail de recherche, notre type de textes relève clairement de l'analyse linguistique et non des registres ou 'genres' que distingue Biber (1994)[242], les types de textes correspondant à des corrélations de caractéristiques linguistiques qui participent d'une même fonction globale. De ce fait, les types de textes ne se confondent ni avec les typologies fonctionnelles ni avec les 'genres' qui sont les catégories intuitives qu'utilisent les locuteurs pour répartir les productions langagières.

De plus, Biber fournit un certain nombre de paramètres situationnels permettant de décrire les documents intégrés dans un corpus :

« Canal : écrit / parlé / écrit lu

Format : publié / non publié

Cadre : institutionnel / autre cadre public / privé-interpersonnel

Destinataire :

- *pluralité : non compté / pluriel / individuel / soi-même*
- *présence : présent / absent*
- *interaction : aucune / peu / beaucoup*

Connaissances partagées : générales / spécialisées / personnelles

Destinateur :

- *variation démographique : sexe, âge, profession etc.*
- *statut : individu / institution dont l'identité est connue*
- *Factualité : informatif-factuel / intermédiaire / imaginaire*

[241] Enjeux épistémologiques de la linguistique de corpus, p. 31

[242] Representativeness in corpus design, p. 380

Objectifs : persuader, amuser, édifier, informer, expliquer, donner des consignes, raconter, décrire, enregistrer, se révéler, améliorer les relations interpersonnelles, ...

Thèmes : ... » (Biber, 1994)[243]

Ces paramètres permettent d'examiner le lien entre cet ancrage situationnel et la caractérisation proprement linguistique du corpus, mais il s'agit là des paramètres d'un corpus de référence représentant la langue générale. Pour un domaine technique ou scientifique tel que la chimie, employant des usages spécialisés, dans une langue de spécialité, représentée dans un corpus spécialisé, les analyses et les visées sont différentes générant des nouveaux paramètres.

3.1.2.1 Texte spécialisé

Pour construire un corpus textuel de spécialité, il faut « sa*voir ce qu'on entend par texte* spécialisé ou comment nous identifions les textes spécialisés » (Cabré, 2008)[244]. Ainsi, Cabré définit la notion de textes spécialisés :

« Les textes spécialisés sont les productions linguistiques, orales ou écrites, qui se manifestent dans le cadre des communications professionnelles et dont la finalité est exclusivement professionnelle. » (Cabré, 2008)[245]

Pour reconnaitre ces situations professionnelles, nous définissons des critères en nous basant sur les conditions discursives, cognitives et linguistiques de Cabré (2008)[246] et dans le cadre de ce travail nous considérons que nous pouvons représenter l'expert du domaine. Ceci nous amène à prendre en compte le domaine des textes (thèmes et sujets), la source des textes (auteurs et supports) et le genre des textes (langue et forme, spécialisation et vulgarisation, unités lexicales et constructions syntaxiques). Ces critères permettent d'identifier les textes produits par des spécialistes mais aussi de déterminer quels types de textes appartiendront au corpus :

« Il convient de déterminer quels types de texte nous devons retenir pour que le corpus qui en résulte soit suffisamment équilibré ». (Cabré, 2008)[247]

En effet, les textes de spécialité, comme les textes de la langue générale, présentent une grande diversité tant par le choix de leur domaine que par celui de leur source mais également celui de

[243] Representativeness in corpus design, p. 380
[244] Constituer un corpus de textes de spécialité, p. 38
[245] Idem
[246] Idem, p. 39
[247] Idem, p. 38

leur genre. Cela implique différents types de textes aussi bien sur leur représentation physique que sur leur représentation logique des documents.

De ce fait, notre corpus se voit être constitué de plusieurs sujets et de plusieurs thèmes, de différents auteurs et de différentes sources mais également de niveau de langue, de spécialisation et de vulgarisation distincts. Cela entraîne l'emploi d'unités lexicales et de constructions syntaxiques très riches et nous permettra de répondre à notre objectif.

3.1.2.2 Taille de corpus et représentativité

Dans la constitution d'un corpus, il faut « déterminer la quantité de productions qui feront partie de ce *corpus, pour savoir s'il sera suffisamment représentatif de chaque spécialité* ou bien seulement pour analyser un thème préalablement choisi » (Cabré, 2008)[248]. La question est de savoir si le corpus constitué, considéré comme un échantillon de données langagières, évitera les erreurs statistiques classiques que présente Biber (1994)[249], notamment l'incertitude, survenant quand un échantillon est trop petit pour représenter avec précision la population réelle, et la déformation, se produisant quand les caractéristiques d'un échantillon sont systématiquement différents de ceux de la population que cet échantillon a pour objectif de refléter. En effet, un échantillon trop petit ne peut pas bien représenter la population ; il est systématiquement biaisé et s'écarte significativement des caractéristiques de la population. Pour ces raisons, certains linguistes dont Sinclair préconisent un grand corpus ; à titre d'exemple, le nombre minimum d'un million de mots par cellule :

> *« Un corpus est supposé contenir un grand nombre de mots. L'objectif fondamental de la constitution d'un corpus est le rassemblement de données en grandes quantités » (Sinclair, 1996)*[250]

L'objectif de la constitution d'un corpus est de produire un échantillon représentatif de la langue traitée ; puisqu'il est difficile de déterminer précisément les caractéristiques de ces données langagières, alors le linguiste recense le maximum de données avec l'idée implicite que plus il y en a mieux c'est. Habert (2000)[251] le résume par la formule « gros, c'est beau » impliquant que « la constitution de corpus toujours augmentables et jamais finis au sein desquels, la représentativité serait proportionnelle au volume des données » (Arbach et al., 2013)[252]. De plus, l'évolution de l'informatique facilite la constitution de ces méga-corpus, au

[248] Constituer un corpus de textes de spécialité, p. 38
[249] Representativeness in corpus design, p. 219
[250] Preliminary recommendations on Corpus Typology, p. 9
[251] Des corpus représentatifs : de quoi, pour quoi, comment ?, p. 3
[252] Aspects théoriques et méthodologiques de la représentativité des corpus, p. 8

risque de perdre le peu qui a été trouvé, en raison de l'absence de leur connaissance en profondeur. Cela débouche de fait sur une vision fragmentée et partielle, mettant en doute la fiabilité, l'utilisabilité et donc la pertinence des méga-corpus.

Pour cela, la constitution de corpus doit répondre à des conditions de production et de réception plus nettement définies et corrélées à leurs caractéristiques langagières (Cabré, 2008)[253] : ce sont là les corpus de référence, conçus pour fournir une information en profondeur sur une langue afin de représenter toutes les variétés pertinentes du langage et son vocabulaire caractéristique, de manière à pouvoir servir de base à des grammaires, des dictionnaires et d'autres usuels fiables.

Pour la constitution des corpus spécialisés, la problématique de la taille des corpus dépend, comme les corpus de référence, de la finalité du corpus, puisque selon la nature de la recherche, l'étude peut réclamer des données très vastes ou au contraire se satisfaire d'un corpus restreint :

> *« […] ceci dépend de la finalité du corpus […] s'il s'agit de constituer un corpus pour étudier un problème particulier, la taille du corpus doit être en adéquation avec les finalités proposées. Par exemple, le corpus que nous devons constituer pour analyser l'usage d'un pronom en position enclytique sera de taille moins importante que celui qu'il faudrait pour extraire la terminologie d'un domaine de spécialité. L'extraction de collocations nécessitera un corpus encore plus important. » (Cabré, 2008)*[254]

Autrement dit, afin d'obtenir un corpus représentatif et des conclusions généralisables tirées à partir de ce corpus, il faut que la taille de l'échantillon (le corpus) soit conforme aux conditions de validité de l'objectif d'exploitation du corpus, c'est-à-dire aux « visées linguistiques du corpus » (Arbach et al., 2013)[255]. Il doit également prendre en compte, en amont de la constitution d'un corpus, « l*es limitations d'ordre matériel* » (Arbach et al., 2013)[256] afin de déterminer le temps qu'il faudra pour recueillir les textes et les traiter suivant les étapes d'informatisation, d'annotation, d'étiquetage et d'analyse (Meyer, 2002)[257]. En effet, la constitution de corpus nécessite un temps considérable et ce travail laborieux et fastidieux n'est pas du goût de tout le monde, puisqu'il faut y consacrer son temps et ses forces. De ce fait, certains réfléchissent à la possibilité de récupérer des corpus déjà constitués (Cappeau et al.,

[253] Constituer un corpus de textes de spécialité, p. 40
[254] Idem
[255] Aspects théoriques et méthodologiques de la représentativité des corpus, p. 9
[256] Idem
[257] English corpus linguistics : An introduction, p. 32

2007)[258]. Aussi, cette situation d'exploiter des corpus constitués par d'autres est de plus en plus en vigueur :

> *« La constitution d'une « bibliothèque de données de corpus » riche, ouverte et partageable, est une préoccupation dans l'air du temps. Le développement du net a pu laisser croire, un temps, que les chercheurs allaient enfin disposer – qui plus est facilement ! - de toutes les données dont ils avaient besoin. Mais il s'est avéré que récupérer du corpus sur le net soulève aussi des écueils, et qu'il ne s'agit en aucun cas d'une voie royale. » (Cappeau et al., 2007)[259]*

Dans la notion même de corpus, il y a l'idée de mise en commun de données, même si cela parait paradoxal puisqu'un corpus est construit en fonction d'un objectif précis. Par conséquent, l'échange des corpus et leur réutilisation pour d'autres objectifs n'ont pas abouti jusque récemment. Pour certains domaines et certains objectifs de travail, cet échange peut faire sens ; mais cela suppose que « *la relation du chercheur à son terrain et son corpus n'est pas fondatrice de l'interprétation, et qu'il* est *de peu d'incidence qu'il ait ou non participé à l'élaboration du protocole, la récolte, l'enregistrement, l'ordonnancement, la transcription* » (Cappeau et al., 2007)[260]. A cela, ils ajoutent :

> *« Dans la procédure de recueil, les chercheurs qui sont à l'origine de la conception d'un corpus savent sur quels objectifs interprétatifs ils s'engagent, et il est rare qu'ils se contentent d'un regroupement guidé par le hasard ou les opportunités. Il convient alors de se demander dans quelle mesure il est possible de « détourner » ces corpus de leur objectif premier, de les exploiter au-delà de ce pour quoi ils ont été constitués : corpus pour tous usages vs corpus pour objectifs interprétatifs. » (Cappeau et al., 2007)[261]*

En d'autres termes, le corpus collecté peut s'inscrire dans une perspective descriptive, formelle, ou interprétative, impliquant des sens différents du mot 'corpus' ; cela peut influencer les caractéristiques et les objectifs du corpus mais aussi sa sollicitation.

La sollicitation des corpus, notamment des méga-corpus, a suscité le développement des outils informatiques destinés à leur analyse, tels que les concordanciers, les analyseurs ou les extracteurs, atteignant un nombre conséquent d'outils mis à disposition pour les langues comme l'anglais, le français, l'espagnol… Mais l'arabe voit ce foisonnement d'outils s'effondrer tant par sa graphie agglutinante que par la complexité de la reconnaissance optique

[258] L'exploitation sociolinguistique des grands corpus, p. 108
[259] Idem
[260] Idem, p. 102
[261] Idem

de ces caractères (Atwell et al., 2004)[262]. De plus, il faut ajouter le manque de coordination de méthodologies et de standards entre équipes, cela étant valable pour toutes les langues. Cette contrainte d'ordre matériel n'est pas anodine et doit être prise en compte, notamment dans le traitement des corpus.

Au vu des contraintes matérielles, nous ne fixons pas de limite de taille à notre corpus. D'une part, nous choisissons de couvrir tout le domaine de la chimie quels que soient les thèmes et les sujets, en vue de prétendre à la représentativité du domaine et non à son exhaustivité. D'autre part, nous restreignons notre recherche au monde universitaire, englobant les travaux d'enseignants et d'étudiants, afin de garantir un niveau de spécialisation et de vulgarisation des textes, rédigés en arabe, mais aussi leur fiabilité ainsi que leur authenticité.

3.2 Démarche de la recherche du corpus

Les critères de la constitution du corpus étant déterminés, nous pouvons commencer à collecter les données langagières, en fournissant à partir des choix des domaines, des sources et des genres des textes, les types de données et d'informations à extraire.

Notre connaissance du domaine nous permet de cibler ces populations langagières, en sélectionnant les textes que nous avons considérés comme pertinents et en les classant par thème dans une structuration du domaine que nous avons préalablement conçue (Cabré, 2008)[263]. Cela évite d'enrichir notre corpus de données incorrectes et inexactes en favorisant l'implémentation d'informations pertinentes afin d'affiner au maximum la conception de notre outil informatique pour obtenir une meilleure robustesse et diminuer le bruit.

À cela, nous ajoutons certaines phases de la constitution du corpus que propose Cabré, une fois les critères établis :

> *« a. la sélection des sources, b. les critères de sélection des textes et la décision de savoir s'il faut prendre le texte complet ou des fragments du même texte, c. les décisions quant à l'architecture de base, d. les décisions quant à l'infrastructure logicielle et matérielle (système de gestion de corpus textuels), e. la sélection des conventions pour la représentation des textes, f. les critères, langage et système de balisage structurel (Cabré, 2008)*[264]

[262] Review of Arabic Corpus Analysis Tools Un Examen d'Outils pour l'Analyse de Corpus Arabes
[263] Constituer un corpus de textes de spécialité, p. 44
[264] Idem, p. 41

Ces phases distinctes de la constitution effective du corpus nous permettent de déterminer les sources choisies.

Nous avons décidé de chercher notre corpus sur la toile, puisque l'application finale est destinée à la fouille de textes et le contexte de son application est la recherche d'information. De plus, la toile constitue un réservoir textuel et lexical fabuleux ; d'une part, elle favorise la constitution de corpus de grande taille, composés d'une masse conséquente de données lexicales (Namer, 2003)[265], provenant de sources multiples, variées et complètes, pour couvrir tout le domaine étudié du point de vue lexical, et d'autre part, elle fournit des documents sous format électronique, facilement téléchargeables et exploitables par des outils de traitement automatique des langues (Berland et al., 2002)[266]. De ce fait, la toile offre aujourd'hui aux linguistes une nouvelle ressource d'information ; ces dernières années témoignent de l'émergence de travaux tentant d'exploiter ce type de données afin de :

> *« […] multiplier les études lexicométriques, les découvertes de termes obéissant à des constructions lexicales imprévues ; confronter des conditions d'utilisation et de construction lexicales à partir de domaines spécialisés différents ; étudier l'impact du registre de langue sur la créativité lexicale… » (Namer, 2003)*[267]

Cependant, cette ressource requiert la connaissance, la maitrise et l'utilisation d'outils informatiques, pour traiter et exploiter les données textuelles :

> *« En effet, même quand il a identifié avec certitude le type et la localisation de l'arborescence de documents à récupérer, il reste encore au morphologue à résoudre les problèmes de récupération, de traitement, de filtrage, de classement et d'exploitation des données contenues dans ces documents Web ». (Namer, 2003)*[268]

Cette juxtaposition de textes issus de la toile ne fait pas un corpus électronique, en considérant que ce corpus électronique est un corpus encodé de manière standardisée et homogène pour permettre des extractions non limitées à l'avance (Sinclair, 1996)[269]. Il doit respecter des conventions de représentation et de codage répandues pour permettre la transmission et la réutilisation des données textuelles. Par la suite, ce sont ces codages et ces outils qui limiteront et orienteront notre recherche, aboutissant à la constitution d'un modeste corpus, issu des travaux des étudiants du département de chimie des universités du monde arabe.

---

[265] Le modèle Lstat : ou comment se constituer une base de données morphologique à partir du Web, p. 86
[266] Assistance automatique pour l'homogénéisation d'un corpus Web de spécialité, p. 2
[267] Le modèle Lstat : ou comment se constituer une base de données morphologique à partir du Web, p. 86
[268] Idem
[269] Preliminary Recommendations on Corpus Typology, p. 5

3.2.1 Exploration de la toile

La toile recense une quantité infinie de données, difficilement imaginable, permettant un large choix dans la constitution de notre corpus. Le domaine choisi étant une science dure, la qualité et la fiabilité du corpus occupent une place cruciale dans la constitution de notre corpus afin de valider les résultats obtenus, impliquant un choix restreint des documents tels que les articles, les brevets, les périodiques et les travaux de recherche.

3.2.1.1 Recherche de documents

Ces documents sont recherchés sur la toile en exploitant différents sites ; mais, ils sont peu abondants (Lachheb, 2007)[270], voir peu librement disponibles (Neifar et al., 2016)[271]. En effet, un nombre très limité de documents scientifiques en arabe sont disponibles sur la toile, puisque la plupart sont écrit en anglais et/ou en français, contrairement aux documents littéraires qui fleurissent de plus en plus et que nous pouvons consulter dans des sites comme Al-Waraq, un des meilleurs sites de littérature arabe, contenant des milliers de pages d'œuvres originales, ou encore Arabic Literature, un point d'accès aux meilleurs sites de tous les genres de littérature arabe, tenu à jour par la bibliothèque de Cornell University.

Quant aux sciences en général, et la chimie en particulier, nous ne pouvons que déplorer le retard du monde arabe, offrant à ses scientifiques arabisants très peu ou pas de ressources scientifiques en arabe (base de données, sites, périodiques), impliquant que certains utilisent des forums sur la toile pour publier leurs travaux. Ces sources, riches en information, ne peuvent pas être considérées comme fiables et authentiques, par la nature de la source, le forum étant un espace de discussion publique non soumis aux règles de publication scientifique, même si l'un des mobiles de publication est de transmettre les résultats d'une recherche ou d'un travail professionnel aboutissant à une idée confirmée par une expérimentation, une technique, une observation nouvelle.

De ce fait, une recherche sur les ouvrages scientifiques parait une perspective adéquate, permettant de nous affranchir de ce problème, afin de constituer notre corpus, puisque le support des textes d'un corpus écrit et publié peut être des périodiques et/ou des livres, avec des proportions plus ou moins proches selon les objectifs de travail.

[270] Lexique du Commerce Electronique. Anglais – Français – Arabe, p. 7
[271] Acquisition terminologique en arabe : État de l'art, p. 5

Notons que les livres choisis devront avoir une version numérisée afin de respecter le cadre de travail fixé ainsi que l'objectif, qui est la conception d'un extracteur morphosyntaxique.

3.2.1.2 Recherche de livres

Dans la constitution d'un corpus de spécialité exploitable, nous recherchons des livres traitant de la chimie en arabe sur la toile ; mais vu le nombre limité d'ouvrages scientifiques arabes numérisés, nous élargissons les critères de la constitution du corpus, notamment le thème des ouvrages et leur datation sans limite.

Pour cela, nous définissons dans une première liste des mots-clés appartenant de manière non-ambiguë au domaine de la chimie. La présence d'un de ces mots dans le titre du livre et/ou dans le livre est l'indice que le texte aborde le domaine, par exemple les mots-clés « كيمياء = kîmy*â'* = chimie » ainsi que ses variantes et équivalents comme 'chimie organique', 'chimie du carbone'… Nous rappelons qu'en interrogeant le Web, l'information porte sur le sujet de la requête et non la réponse à la requête. Ces mots-clés sont appelés 'généraux' en opposition à la seconde liste constituée de mots-clés appelés 'spécifiques' (Berland et al., 2002)[272], dans le sens où ils servent à préciser, parmi les textes abordant le domaine étudié, des points de vue différents tels que la règlementation et la technique. Nous supposons que ces mots-clés sont spécifiques, comme « كروماتوغرافيا = krûmât*ûġâfiyâ* = chromatographie », et permettent de délimiter et d'identifier le domaine de la chimie dans toute sa diversité, afin de sélectionner les livres les plus pertinents.

Plusieurs livres sont retenus ; mais, nous n'en citerons qu'un seul, « أسس الكيمياء العضوية = 'usus alkîmy*â' al'uḍwiyat* = les bases de la chimie organique », ouvrage universitaire libyen de Wael Ghaleb Mohamed et Walid Mohamed Alsayti de 2008. En effet, nous n'utiliserons pas d'ouvrages universitaires pour la simple et bonne raison que suite aux résultats obtenus lors de leur conversion et de leur nettoyage, nous n'avons pas pu mettre en œuvre les outils de traitement automatique de la langue. Néanmoins, nous présentons ce travail laborieux et fastidieux, pourtant si prometteur, que nous avons mis en place, lors du nettoyage de notre corpus.

[272] Assistance automatique pour l'homogénéisation d'un corpus Web de spécialité, p. 3

3.2.2 Prétraitement

Après avoir constitué notre corpus d'ouvrages traitant de la chimie, notre curiosité et notre enthousiasme nous ont poussés à tester une partie de ce corpus avant d'achever le dépouillement des données. Cela nous nous permis de réagir rapidement pour la question du format des textes constituant notre corpus, afin de les traiter avec des outils informatiques.

Nous avons choisi d'étudier le second chapitre de l'ouvrage « الهيدروكربونات = alhîdûkarbûnat = les hydrocarbures » ; ce choix est motivé par la volonté d'explorer différentes variétés pertinentes du langage et son vocabulaire caractéristique, de manière à servir de base à la conception de notre outil informatique.

Un traitement préalable est mis en place, comportant les étapes de conversion et de correction des erreurs récurrentes, correspondant à notre procédure de nettoyage, afin d'obtenir un corpus analysable ; cette phase est considérée « *comme l'étape fondament*ale de constitution de la chaîne de traitement » (Namer, 2003)[273].

3.2.2.1 Conversion du texte

La conversion est le changement de format des documents sélectionnés. En ce qui nous concerne, nous souhaitons traiter notre corpus par des outils informatiques. Nous avons enregistré nos ouvrages au format PDF ; nous allons donc les convertir en format TXT, sachant que lors de cette opération, une perte d'information est possible et sera à préciser.

Ce passage des formats des textes retenus en texte brut constitue pour nous une étape nécessaire à la constitution de corpus, puisque notre objectif est l'analyse de données lexicales, plus précisément, l'analyse de données terminologiques, le but final étant la conception d'un extracteur de morphosyntaxique des termes de la chimie en arabe.

A première vue, la conversion est correcte ; mais de plus près, certains caractères ne sont pas identifiés dans le texte (cf. **Erreur ! Source du renvoi introuvable.**). En effet, nous observons principalement la présence d'une espace dans un mot en arabe et le remplacement de certains caractères arabes par d'autres.

[273] Le modèle Lstat : ou comment se constituer une base de données morphologique à partir du Web, p. 88

الهيدروكربونات

Hydrocarbons

يطلق اسم الهيدروكربونات على المركبات التي تتكون من ذرات الكربون والهيدروجين فقط وتنقسم إلى هيدروكربونات أليفاتية وهيدروكربونات أروماتية .

الهيدروكربونات الأليفاتية Aliphatic hydrocarbons : هي عبارة عن مركبات ذات سلاسل مستقيمة أو متفرعة أو حلقية وقد تكون مشبعة أو غير مشبعة ولقد اشتق اسم أليفاتية من الكلمة اليونانية aleiphas وتعني " الدهن fat "

التشبع Saturated : يقصد بالتشبع هو أن تكون جميع روابط C-C أحادية بمعنى أن عدد ذرات الهيدروجين هو الحد الأقصى الذي يمكن للهيدروكربون أن يحتويه سواء كان المركب حلقي أو غير حلقي .

عدم التشبع Unsaturated : المركب غير المشبع هو الذي تحتوي جزيئاته على روابط ثنائية أو ثلاثية ويكون عدد ذرات الهيدروجين أقل من العدد الأقصى الذي يمكن للهيدروكربون أن يحتويه .

الهيدروكربونات الأروماتية Aromatic hydrocarbons : هي هيدروكربونات تحتوي على حلقة بنزين (ص[129].)

الهيدآوربونارت

Hydrocarbons

يطلق اسم الهيدآوربونات على المركبات التي تتكون من ذرات الكربون والهيدروجين فقط وتنقسم إلى هيدآوربونات أليفاتية وهيدآوربونات وأروماتية .

الهيدآوربونات الأليفاتية Aliphatic hydrocarbons : هيدروكربونات عبارة عن مركبات ذات سلاسل مستقيمة أو متفرعة أو حلقية وقد تكون مشبعة أو غير مشبعة ولقد اشتق اسم أليفاتية من الكلمة اليونانية aleiphas وتعني " الدهن fat "

التشبع Saturated : بالتشبع يقصد بأن تكون جميع اواربط C-C أحادية بمعنى بأن عدد ذرات الهيدروجين هو العدد الأقصى الذي يمكن للهيدآوربون بأن يحتويه سواء كان المركب حلقي أو غير حلقي .

عدم التشبع Unsaturated : المركب غير المشبع هو الذي تحتوي جزيئاته على اواربط ثنائية أو ثلاثية وقد تكون عدد ذرات الهيدروجين أقل من العدد الأقصى الذي يمكن للهيدآوربون بأن يحتويه .

الهيدآوربونات الأوروماتية Aromatic hydrocarbons : هي هيدآوربونات تحتوي على حلقة بنزين (ص[129].)

Figure 17 : Image avant et après conversion, avec identification des erreurs selon Word

Par conséquent, une procédure de nettoyage est mise en place.

3.2.2.2 Procédure de nettoyage

Le nettoyage consiste à préparer le texte, en isolant et en structurant les éléments d'information, mais aussi en définissant et en segmentant le flux de données en unités de sens, afin de rendre analysable le corpus étudié par un traitement informatique.

Nous présentons les corrections et les traitements des différents caractères textuels, notamment le traitement de la ponctuation en général, des espaces en particulier, le traitement des unités fréquentes et le traitement des caractères arabes. Pour finir, nous tentons un nettoyage semi-automatique à l'aide de deux outils principalement, le premier étant les 'expressions régulières' :

> *« Les expressions régulières sont des chaînes de caractères permettant de définir des règles sophistiquées de recherche et de remplacement de motifs. Elles constituent un puissant outil pour la manipulation de textes et de données. Les expressions régulières permettent d'exécuter en quelques secondes des tâches habituellement longues et fastidieuses. » (Friedl ,2003)*[274]

Quant au second outil, il s'agit de la 'macro' :

> *« C'est une action ou ensemble d'actions utilisées pour automatiser des tâches [...] les macros sont enregistrées dans le langage de programmation Visual Basic pour Applications »*[275]

3.2.2.2.1 Traitement des différents caractères textuels

En informatique, l'écriture de la langue arabe est restée fidèle à sa structure manuscrite ; mais, sa représentation possède les caractéristiques de l'écriture arabe telles que les ligatures, ce qui engendre des difficultés d'identification :

> *« L'écriture arabe connaît trois types de ligatures : les ligatures contextuelles, les ligatures linguistiques et les ligatures esthétiques. Une ligature contextuelle est une chaîne de caractères prenant des formes spéciales suivant leur position dans le mot en obéissant à des règles grammaticales strictes et liées uniquement à l'écriture. Les ligatures linguistiques sont indispensables pour l'écriture d'une langue donnée et obéissant à des règles grammaticales. Souvent elles ont un statut de lettre et parfois même une place à part dans le dictionnaire, ce qui les rapproche des digraphes. Les ligatures esthétiques sont des graphies optionnelles qui existent pour des raisons*

[274] Maîtrise des expressions régulières, p. 5

[275] https://support.office.com/fr-fr/article/automatiser-des-t%C3%A2ches-avec-l-enregistreur-de-macro-974ef220-f716-4e01-b015-3ea70e64937b

esthétiques, de lisibilité et/ou de tradition. On peut les remplacer par leurs composantes sans changer la validité grammaticale, ou le sens du texte. » (Zghibi, 2002][276].

À cela, il faut ajouter une multitude de graphismes des lettres arabes, qui lors de leur écriture, se lient les unes aux autres, entraînant « *quatre morphologies différentes d'une même lettre en* fonction de son emplacement dans le mot : initiale, m*édiane, finale et isolée à l'exception des* six lettres (ا, د, ذ, ر, ز, و) qui ne possèdent que deux formes seulement » (Zghibi, 2002).[277]

De ce fait, nous remarquons que, dans certains mots arabes, les lettres d'un même mot sont séparées par une espace. Nous observons aussi le remplacement de caractères dans certains mots arabes ne permettant plus leur compréhension. Enfin, nous repérons dans certains mots la présence de lettres arabes sous la forme isolée.

- Traitement des espaces

La présence d'une espace dans un même mot est due aux ligatures situées entre les caractères arabes ; pour certains caractères, la taille de la ligature est proche de celle d'une espace entre deux mots et se traduit en informatique comme une espace classique, impliquant une séparation entre les caractères d'un même mot tel qu'illustré dans le tableau suivant (cf. Tableau 8 : Mots arabes séparés par une espace) :

Mot erroné	Mot corrigé	Équivalent
كيمي اء	كيمياء	Chimie
كرب ون	كربون	Carbone
عذ صر	عنصر	Élément

Tableau 8 : Mots arabes séparés par une espace

Pour certains mots, la séparation par une espace s'effectue lorsqu'ils contiennent une lettre finale comme « ا » pour « كيمياء » ; pour d'autres, il s'agit d'une lette médiane, par exemple « ص » pour « عنصر ». Ce phénomène est observé tout au long du corpus pour la plupart des lettres arabes, en position médiane et/ou en position finale. Cependant, la recherche automatique de cette erreur et sa correction ne sont pas réalisables par les expressions régulières, qui ne peuvent pas prendre en compte ce phénomène, puisque l'espace étant le caractère utilisé pour séparer deux mots, il n'est pas possible de le distinguer avec notre erreur. Pourtant, nous avons pu corriger cette erreur, en tenant en compte de la multitude de

[276] Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646, p. 163
[277] Idem, p. 162

graphismes de l'écriture arabe : il suffit pour notre recherche et notre correction de cibler les lettres médianes et/ou finales d'un mot contenant une espace, ce qui implique l'écriture d'une expression régulière pour chaque lettre de l'alphabet arabe et pour chaque forme, médiane et finale comme les exemples suivants (cf. Tableau 9 : Expressions régulières de quelques mots arabes) :

Mot erroné	Rechercher expression régulière	Remplacer expression régulière	Mot corrigé
كيمي اء	Espaceا	ا	كيمياء
كرب ون	Espaceو	و	كربون
عذ صر	Espaceص	ص	عنصر

Tableau 9 : Expressions régulières de quelques mots arabes

- Traitement des remplacements de caractères dans certains mots arabes ne permettant plus leur compréhension

Le remplacement de caractères tels que 'ك = k' pour 'المركبات' par 'آ = â' obtenant 'المرآبات' ne permet plus la compréhension. En effet, lors de la conversion, certains caractères arabes sont mal interprétés par le système informatique, en raison de la taille de leurs ligatures, confondant les caractères entre eux. La correction de cette erreur par une expression régulière n'est pas envisageable, puisqu'il s'agit de caractères arabes ; son identification et son remplacement ne peut que fausser les mots qui la contiennent, impliquant une correction manuelle pour le traitement de ces caractères.

- Traitement des lettres arabes sous la forme isolée

La présence de lettres arabes sous la forme isolée dans certains mots arabes tels que 'والكربوﻫيدرات' au lieu de 'والكربوهيدرات' ne permet pas la reconnaissance du mot, du fait que le caractère 'ه', pour notre exemple, situé en début de mot, est représenté par une lettre isolée. En effet, lors de la conversion, certains caractères arabes se voient attribuer une autre forme, généralement, la forme isolée, en raison de leurs ligatures, confondant les caractères entre eux. La correction de ces formes par une expression régulière n'est pas envisageable, puisqu'elles ne se différencient pas, impliquant une correction manuelle pour son traitement.

3.2.2.2.2 Finition du nettoyage

D'autres erreurs sont présentes dans le corpus, mais elles ne sont pas significatives et correspondent plutôt à des erreurs de frappe ou à des fautes d'orthographe et/ou de grammaire ; elles sont corrigées manuellement si cela est nécessaire, puisque pour un travail de fouille de textes, notre champ de travail est la langue telle qu'elle est écrite, comprenant ces erreurs

d'orthographe et de grammaire. Mais, pour l'étude terminologique de la chimie, nous corrigeons les erreurs d'orthographe et/ou de grammaire afin d'indexer les termes de chimie identifiés pour les intégrer dans notre système. De plus, notre domaine présente une abondance de termes empruntés du français et/ou de l'anglais, impliquant une variation d'orthographe pour une même dénomination (cf. 2.2.2.1.4 Emprunt). Par conséquent, nous les repérons dans notre corpus et nous les précisons dans les fiches terminologues.

3.2.2.2.3 Nettoyage semi-automatique du corpus

Le corpus constitué comporte un nombre important de pages à analyser, rendant le nettoyage long et fastidieux ; la réalisation répétitive des différentes étapes de la procédure de nettoyage présuppose la possibilité de l'automatiser ou du moins automatiser une partie de la procédure.

Afin de ne pas oublier une expression régulière parmi les nombreuses que nous avons établies et en vue de faciliter la procédure de nettoyage, nous proposons de mettre en place une macro, en prenant en compte les expressions régulières des mots contenant une espace. Voici un extrait de la macro (cf. Figure 18 : Extrait de la macro pour le nettoyage semi-automatique du corpus) :

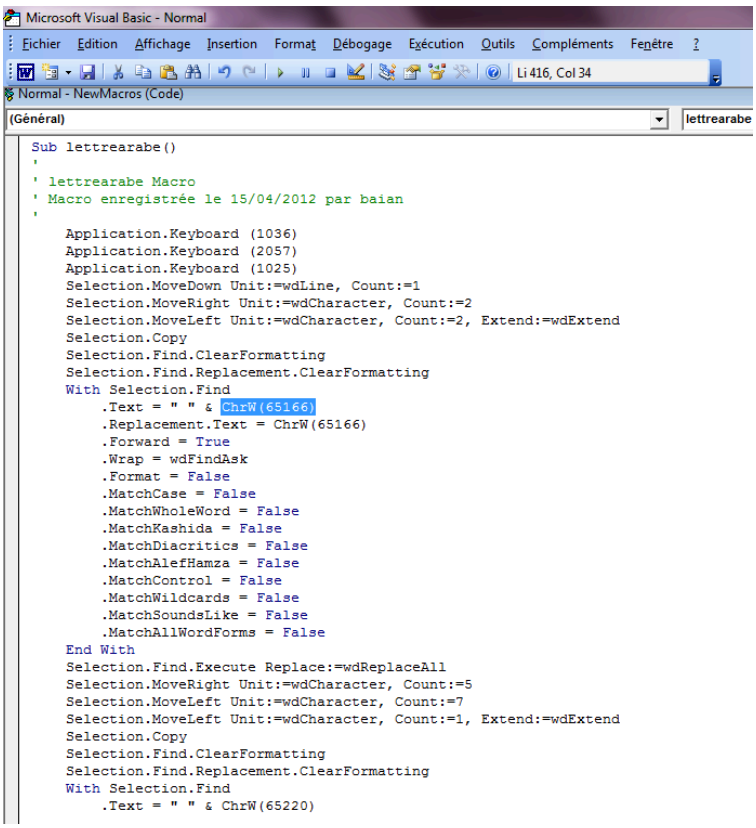


Figure 18 : Extrait de la macro pour le nettoyage semi-automatique du corpus

Il s'agit du codage dans la macro pour la lettre finale « ا = â », précédée d'une espace ; cependant, lors de la vérification de cette macro, nous remarquons que les caractères arabes n'apparaissent pas dans le codage, mais ils sont représentés par une référence, « ChrW (65166) » pour notre exemple. Cette macro prend en compte la liste exhaustive des expressions régulières ; nous avons fait ce travail rigoureusement et avec une telle minutie que nous considérons notre procédure de nettoyage valide. Voici un extrait du corpus obtenu avant la procédure de nettoyage (cf. Figure 19 : Extrait du corpus avant nettoyage) et après la procédure de nettoyage (cf. Figure 20 : Extrait du corpus après nettoyage), englobant l'exécution de la macro et les corrections manuelles :

الوزن الذريA) Atomic weight) : (هو مجموع عدد البروتونات والنيوترونات في النواة ويساوي

تقريبا رقم الكتلة الذي يكتب أعلى رمز العنصرMass numberA = p + n

Figure 19 : Extrait du corpus avant nettoyage

الوزن الذري A (Atomic weight) : هو مجموع عدد البروتونات والنيوترونات في النواة ويساوي تقريبا رقم الكتلة الذي يكتب أعلى رمز العنصر Mass number A = p + n

Figure 20 : Extrait du corpus après nettoyage

À présent, notre corpus est prêt à être analysé mais nous avons conscience que le contrôle complet des résultats de cette procédure de nettoyage est nécessaire, présentant un des inconvénients majeurs à la constitution de corpus ; des pistes doivent être envisagées afin de diminuer l'ampleur de la tâche dévolue à l'opérateur humain.

3.2.2.3 Limite du corpus

Suite au prétraitement de notre corpus, nous avons testé une ébauche de notre système sur une partie du corpus ; mais, nous n'avons obtenu aucun résultat exploitable. Cela nous oblige donc à repenser la constitution notre corpus.

Nous avons pensé que notre procédure de nettoyage était valable, considérant qu'au vu des résultats, pour l'œil humain, le corpus était analysable puisque les caractères apparaissaient correctement. Cependant, la machine n'a pas la même vision des choses, puisqu'elle analyse les codes et les références des caractères et non leur graphie et leur forme.

La difficulté principale de ce prétraitement est liée à la Reconnaissance Optique des Caractères (ROC), plus connue sous les initiales en anglais OCR (Optical Character Recognition) ; il s'agit du codage des diacritiques. Il en résulte que les problèmes liés à la conversion des textes (contenu rédactionnel) d'un document PDF sont similaires à ceux rencontrés lors du scan des documents. C'est justement cette traduction en texte depuis le format des documents

sélectionnés qui constitue la tâche la plus complexe à réaliser pour le traitement informatique de l'écriture arabe (Zghibi, 2002)[278], en raison de l'utilisation, notamment des ligatures contextuelles, linguistiques et esthétiques mais aussi des signes diacritiques, rendant difficile la reconnaissance des caractères optiques, impliquant que la conversion au format texte est très faible.

Bien qu'il soit parlé par presque 300 millions de personnes, l'arabe ne dispose pas d'assez de ressources (Meftouh et al., 2007)[279], notamment des ressources gratuites, et le manque de coordination de méthodologies et de standards entre équipes ne favorise pas la recherche qui n'a commencé que récemment pour cette langue complexe, tant par sa graphie agglutinante que par la multitude de graphismes des lettres arabes.

A cela s'ajoutent certaines particularités des textes de chimie en arabe, notamment la notation des formules mathématiques et/ou chimiques et l'utilisation du français et/ou de l'anglais juxtaposé à l'arabe comme dans l'extrait corrigé (cf. Figure 20 : Extrait du corpus) ; les formules sont écrites en caractères latins, des chiffres… Et quant au caractère multilingue, il s'agit d'une pratique courante due à l'absence de traduction et/ou pour informer le lecteur de l'équivalent du terme dans la langue source, l'anglais et/ou le français. Ceci implique l'utilisation d'un mélange de scripts (suite de commandes, d'instructions permettant d'automatiser une tache), à laquelle la multitude de niveaux de texte et de ses langues rajoute une complexité supplémentaire. Cela suppose l'utilisation d'un mélange de codage, permettant de reconnaitre et de rédiger un texte combinant les caractères arabes et latins. Parmi les différents systèmes prenant en compte les caractères arabes (Zghibi, 2002)[280], nous retenons l'Unicode UTF-8 pour la conformité de son codage, contrairement à la plupart des textes présents sur le web, notre source pour la constitution de corpus numériques, et encore moins à ceux des librairies, notre source pour la constitution de ressources textuelles, qui n'utilise pas ce codage. Cela ne facilite pas le prétraitement du corpus et explique les résultats obtenus.

Étant consciente de ces particularités des données de notre domaine d'étude, nous avons sous-estimé la problématique de la reconnaissance des caractères arabes et nous avons pensé que nous pourrions exploiter les corpus issus de la toile. Face à cette situation, nous avons essayé de chercher des solutions pour enregistrer un texte directement sur la machine :

[278] Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646, p. 162
[279] Constitution d'un corpus de la langue Arabe à partir du Web, p. 2
[280] Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646, p. 166

« Il y a au moins deux façons d'enregistrer un texte en machine : par saisie directe sur clavier ou par numérisation, c'est-à-dire par scannage de l'édition de référence, en papier, suivi d'une opération de Reconnaissance Optique des Caractères » (Mouelhi, 2008)[281]

Dans son travail de recherche, l'auteur a essayé les deux méthodes pour s'affranchir de la problématique de la reconnaissance des caractères arabes et a fini par opter pour la première, suite à l'échec de la seconde.

Pour cette seconde méthode, nous avons vu ses contraintes, même si cela permet un gain de temps non-négligeable dans l'enregistrement de notre corpus en machine par rapport à la première méthode. La numérisation, plus précisément, la reconnaissance optique de caractères pour l'arabe est une opération difficile voire impossible, en raison des causes habituelles telles que la qualité du papier de l'édition de référence, la police de caractères utilisée ou encore les caractères cassés ou tordus. Mais les caractéristiques de cette langue que nous avons déjà citées notamment ses ligatures et la multiplicité de ses formes mais encore la qualité des OCR arabes et l'état d'avancement des recherches dans ce domaine au moment où nous avons entrepris ce travail sont à la base des limites de l'exploitation de notre corpus.

Quant à la première, la saisie directe sur clavier, il nous a été proposé d'écrire le texte pour nous affranchir de ces contraintes ; mais cette démarche de longue haleine représente un frein important à l'entrée massive de données textuelles et une cause fréquente d'erreurs. A cela, elle ajoute une certaine contradiction à notre étude puisque notre travail s'inscrit dans une fouille de textes et a pour objectif l'extraction d'information avec comme finalité la recherche d'information.

Par conséquent, nous réorientons notre recherche pour la constitution du corpus en nous adressant, cette fois-ci, directement à la source.

3.2.3 Retour aux sources

Pour nous affranchir de ces contraintes, pour un souci de fiabilité et de qualité mais surtout pour rattraper le temps perdu, nous nous sommes adressés directement à la source, suite aux conseils de nos enseignants. Mais contrairement à nos attentes, nous avons reçu peu de réponses, si bien que nous obtenons très peu de documents et cela pour plusieurs raisons.

La principale raison, comme nous l'avons dit précédemment et confirmée par nos confrères chimistes du monde arabe, est que la plupart des chercheurs rédigent leurs travaux en anglais,

[281] Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî, p. 55

langue des sciences dominant la scène internationale. A cela s'ajoute le fait que peu de personnes nous ont répondus par manque de temps ou ne disposant pas de corpus à fournir. Enfin, une inquiétude subsiste quant aux droits d'auteur pour les textes du corpus et leur utilisation à des recherches scientifiques. C'est un argument qui parait compréhensible et justifiable pour les chercheurs en science, à qui cette question de corpus n'aspire pas grand-chose ; nous le disons en connaissance de cause.

Cela dit, nous ne perdons pas espoir et nous nous basons sur les modestes textes fournis pour constituer notre corpus, en considérant qu'il peut être complété au fur et à mesure de la collecte des données, sans affecter notre travail de recherche, puisqu'il suffit d'ajouter les nouveaux termes identifiés ainsi que les nouvelles relations ; au contraire, cela permettra de confirmer nos intuitions, nos observations et nos hypothèses.

De ce fait, nous avons recherché les départements de chimie dans les universités arabes et nous leur avons demandé par courrier électronique de nous fournir un corpus spécialisé en arabe en leur expliquant que nous recherchons des textes en arabe dans le domaine de la chimie sous format texte (UTF8) ou MS Word, pour une étude expérimentale dans le cadre de notre thèse.

Nous avons envoyé plus d'une trentaine de mails, du Maghreb à l'Orient, balayant tous les pays arabes afin d'enrichir le corpus au niveau morphosyntaxique ; nous avons obtenu 10% de réponses et parmi elles, nous n'avons obtenu qu'une réponse positive : il s'agit du professeur Abderrahmane Teniou de l'université de Constantine 1 en Algérie, que je remercie sincèrement et sans qui ce travail n'aurait pas abouti, qui nous a fourni principalement des travaux de recherche d'étudiants du département de chimie (cf. Annexe 1 : corpus).

A cela, nous avons ajouté quelques cours magistraux que nous avons trouvé sur les sites des départements de chimie du monde arabe, afin d'enrichir cette collection de textes. Ce choix est tout-à-fait légitime, puisque ces sources de données sont aussi fiables, leurs documents sont tout aussi spécialisés, et leur format de texte est en .doc (cf. Annexe 1 : corpus).

D'autre part, il nous a été proposé de rechercher des corpus dans les organisations et/ou les institutions internationales telles que l'ONU ; cependant, l'accès à ces documents est réservé à ses membres et/ou à ses adhérents. Malgré cela, nous avons récupéré quelques extraits : il s'agit de données fiables et spécialisées, et leur format de texte est en .doc (cf. Annexe 1 : corpus).

Par conséquent, le corpus recueilli est modeste et n'est constitué que d'une petite centaine de milliers de mots, impliquant que nous avons sélectionné le texte complet et non des fragments du même texte et que nous avons pris tous les documents fournis et non les plus représentatifs du domaine. Cependant, nous supposons que notre corpus est homogène, puisque nous

disposons de documents caractéristiques du domaine afin de constituer un corpus pour une étude terminologique, comportant un nombre important de candidats termes et fournissant un bon indicateur quant à leur pertinence et à leur centralité par rapport au domaine. De ce fait, nous considérons que cette collection de documents est suffisamment représentative et variée pour être utilisée comme une source pour notre travail. Nos chances de documenter certains phénomènes et/ou d'identifier des contraintes sur les données recueillies ne sont pas minimes ; au contraire, la qualité et la fiabilité de nos données offrent une variété et une diversité à notre corpus sans équivoque.

Chaque document sélectionné est indexé et son information de type documentaire est enregistrée (Cabré, 2008)[282] précisant son auteur, son titre, son année de parution et son pays d'édition ; il s'agit là des informations qu'un document unique peut comporter dans un corpus.

Dans ce chapitre, nous avons constitué un corpus analysable, composé de textes issus de l'université de Constantine 1 en Algérie, contenant une thèse et un mémoire, quelques cours magistraux et des extraits d'organisations et/ou d'institutions internationales. L'ensemble de ce corpus spécialisé, vulgarisé, fiable et authentique est formé d'environ 100 000 mots non voyellés. Ce corpus répond aux exigences de notre travail de recherche ; à présent, il est possible de réaliser son dépouillement.

[282] Constituer un corpus de textes de spécialité, p. 44

Chapitre 4 : Dépouillement du corpus

Le dépouillement du corpus est l'analyse textuelle visant le repérage de termes et l'extraction d'information sur les concepts sous-jacents mais également les particularités d'emplois des termes relevés. Cela permet ainsi de recueillir les connaissances sur le domaine en question et sur la langue de spécialité qui le décrit. Ce travail de dépouillement du corpus, assisté par ordinateur, nécessite la mise en place d'une norme de dépouillement, indiquant sur quelles bases s'est effectuée la recherche. Nous allons donc décrire les éléments typographiques des textes arabes de la chimie, en mettant en évidence les caractéristiques de la langue étudiée et la particularité de notre domaine de spécialité, juxtaposant et employant la ponctuation, les chiffres et les lettres latines et/ou arabes tout au long de notre corpus. À partir de là, nous nous attaquons à la phase fondamentale de constitution de la chaîne de traitement des données textuelles (segmentation, étiquetage, lemmatisation) avec la spécificité de la langue arabe.

4.1 Analyse des éléments typographiques

Dans un corpus arabe, les éléments typographiques sont la ponctuation, les chiffres et les lettres.

4.1.1 Ponctuation

Auparavant, les textes défilaient comme d'interminables suites de caractères, sans ponctuation, obligeant le lecteur à comprendre tout d'un bloc, en relisant sans doute plusieurs fois le document afin d'en saisir le sens. Ce n'est que plus tard avec les bibliothécaires d'Alexandrie que les signes de ponctuation sont apparus rendant au lecteur une langue plus aérée et plus intelligible.

Système de renfort de l'écriture avec les chiffres, les symboles ..., la ponctuation appartient aujourd'hui à la grammaire d'une langue, ce qui permet d'étudier le style d'un auteur et les caractéristiques d'un genre ou d'une époque, puisque les signes de ponctuation, au même titre

que les mots ou les constructions qui agrémentent une œuvre, ne sont pas choisis arbitrairement :

> *« Le choix des ponctuations dépend, dans le passé comme aujourd'hui, des situations, des genres (...), de l'auteur et des styles ». (Catach, 1994)*[283]

La ponctuation, « ensemble des signes visu*els d'organisation et de présentation accompagnant* le texte écrit, intérieurs au texte » (Catach, 1994)[284], fait partie intégrante du texte, puisque « les signes de ponctuation ont à la fois une signification et un sens dans un contexte donné ; ils jouent un *rôle dans l'élaboration du texte, donc dans la construction du sens* ; ceci implique *qu'ils infléchissent le sens par leur présence, de même que par leur absence* » (Mourad, 2001)[285]. Ainsi, elle doit être prise en compte au même titre que les mots dans une étude linguistique ; même si au final la fonction des signes de ponctuation se limite à faciliter la lecture et la compréhension du texte et ne pourra en aucun cas nous renseigner sur le style de l'auteur.

En arabe, la ponctuation est aussi partie intégrante de l'acte d'écriture, même si celle-ci n'est pas très utilisée (Awad, 2013)[286] ; les signes de base tels que le point «.», les deux points «:» et le point d'interrogation «؟» ont été introduits plus ou moins avec succès, mais certains signes de ponctuation, en revanche, ont été placés d'une façon un peu capricieuse et fantasque tel que l'utilisation des guillemets et/ou des parenthèses pour encadrer les citations et les crochets qui marquent un mot ou un groupe de mots (Mouelhi, 2008)[287]

À cela s'ajoute les signes propres au domaine étudié : dans la chimie en général, dans la nomenclature en particulier, la ponctuation est souvent requise pour nommer un composé chimique comme « 1,3,5,7-tétraazatricyclo[3.3.1.1^{3,7}]décane », impliquant que les signes de ponctuation considérés comme séparateur de chaînes de caractères dans les compilateurs informatiques font partie intégrante du texte (Kaczmarek, 2011)[288]. De ce fait, la ponctuation d'un texte écrit, devant suivre les règles normales de la typographie des signes () [] {} « » - , ; : ? ! ... se retrouve biaisée, d'autant plus que pour l'arabe, celles-ci ne sont pas toujours uniformes, puisqu'elles utilisent aussi bien la ponctuation latine que sa ponctuation propre.

[283] La ponctuation. Histoire et système, p. 113
[284] La ponctuation. Histoire et système, p. 9
[285] Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des Applications informatiques : SegATex et CitaRE, p. 12
[286] La ponctuation arabe : Histoire et règles — Étude contrastive avec le français et l'anglais, p. 20
[287] Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî, p. 162
[288] Nomenclatures française et polonaise de la chimie organique. Analyse comparative, p. 55

Ainsi, dans les textes, on retrouve aussi bien la virgule ‘,’ des caractères latins qu’une autre inversée et orientée à droite ‘،’. Il est à noter que l’apostrophe n’existe pas en arabe.

4.1.2 Chiffre

Contrairement à ce qui est répandu, les chiffres arabes « 1, 2, 3, ... » sont ceux utilisés en occident et dans les pays arabes, même si certains d’entre eux emploient également les chiffres indiens « ١, ٢,٣, ... ». Il est à noter que les chiffres contrairement aux lettres s’écrivent de gauche à droite, comme en français (Habash, 2010)[289] :

Chiffre arabe	0	1	2	3	4	5	6	7	8	9
Chiffre indien	٠	١	٢	٣	٤	٥	٦	٧	٨	٩

Tableau 10 : Chiffres arabes avec leurs équivalents indiens

Comme les signes de ponctuation, les chiffres sont souvent requis pour nommer un composé chimique, par exemple « حمض 2– أمينو 3– هيدروكسي بوتان ثنائي أويك = acide 2-amino-3-hydroxybutanedioïque », et sont même parfois combinés aux signes de ponctuation, impliquant qu’ils font partie intégrante du texte (Kaczmarek, 2011)[290]. De ce fait, la conversion des chiffres, arabes et/ou indiens, en lettres se retrouve aussi biaisée.

Quant aux nombres écrits en lettre, qu’ils soient cardinaux ou ordinaux, ils sont utilisés pour nommer un composé chimique par exemple « ثلاثي إيثيل أمين = *ṯulâṯî ’îṯîl ’*amîn = triéthylamine » et sont même parfois juxtaposés aux signes de ponctuation et/ou au chiffres. Ils sont considérés comme partie intégrante du texte et sont analysés comme des déclencheurs des termes du domaine (2.1.3.2.3.2 UTC de la nomenclature chimique).

4.1.3 Lettre

Contrairement à l’alphabet latin qui s’écrit de gauche à droite avec des lettres allant de la lettre « a » à la lettre « z », l’écriture arabe s’écrit de droite à gauche et se compose de consonnes, de voyelles et de signes diacritiques.

4.1.3.1 Consonne

La langue arabe compte 28 ou 29 consonnes, selon si le *’*alif (ا) et la hamza (ء) sont considérés comme une seule lettre (le *’alif* n’étant qu’un simple support de la hamza) ; les

[289] Introduction to Arabic Natural Language Processing, p. 13

[290] Nomenclatures française et polonaise de la chimie organique. Analyse comparative, p. 63

grammairiens arabes classiques s'étaient divisés sur le nombre exact des lettres de l'alphabet arabe, certains disaient qu'elles sont au nombre de 28 et d'autres, au contraire, considéraient que le *'alif* et la hamza sont deux lettres différentes augmentant ainsi la taille de l'alphabet à 29 lettres (Habash, 2010)[291] :

ا ء ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي

Figure 21 : Consonnes de la langue arabe

Aujourd'hui, cette question n'a pas plus lieu d'être avec l'avènement de l'informatique (norme ISO 233, 1993 – Unicode, 2000 à 2002)[292], puisqu'à chaque caractère graphique un code numérique est attribué, d'autant plus que la hamza, ayant le *'alif* comme support, possède deux autres formes, correspondant à ses deux variantes, le « 'alif hamza en chef أ » et le « 'alif hamza souscrit إ », soit deux autres codes en plus de ce qu'elle possède déjà. À cela s'ajoute une variante du 'alif, le « 'alif bref ou tordu ى (maqsûra) » et le « 'alif madda en chef آ », mais également d'autres formes de la hamza, le « wâw hamza en chef ؤ » et le « yâ ' hamza en chef ئ » et celle du ta ', le « ta ' lié ة (marbûta) » (Dichy, 1990)[293].

Chaque consonne arabe possède entre deux et quatre morphologies différentes, lors de son écriture, selon son emplacement dans le mot : initiale, médiane, finale et isolée ; par exemple, la consonne ha possède les formes :

ﻫ ﻬ ﻪ ه ھ

Figure 22 : Consonne « ha » avec ses différentes variantes

À ces caractères se juxtaposent souvent dans les textes arabes les lettres latines, consonnes et voyelles, qui, comme la ponctuation et les chiffres, sont souvent requises pour nommer un composé chimique, par exemple « يجفف الإيثر فوق $CaCl_2$ مدة 24 ساعة = Yujaffafu al*'îṯir* fawqa $CaCl_2$ muddat *24 sa'a*t = l'éther est séché sur $CaCl_2$ pendant 24 heures », impliquant que les caractères latins font partie intégrante du texte (Kaczmarek, 2011)[294].

4.1.3.2 Voyelle

En arabe, nous avons les voyelles brèves et les voyelles longues.

4.1.3.2.1 Voyelle brève

[291] Introduction to Arabic Natural Language Processing, p. 5
[292] L'ISO 233 est la norme internationale de la translittération des caractères arabes en caractères latins
[293] L'écriture dans la représentation de la langue : la lettre et le mot en arabe
[294] Nomenclatures française et polonaise de la chimie organique. Analyse comparative, p. 62

Apparues après les consonnes, les voyelles brèves « ◌َ ◌ُ ◌ِ » précisent la prononciation du texte purement consonantique ; elles sont ajoutées au-dessus et au-dessous des consonnes, impliquant que chaque forme de consonne est susceptible de porter ces signes.

Mais les textes arabes, à l'exception du texte coranique, du cadre scolaire et/ou jeunesse, ne sont jamais totalement voyellés ; au contraire, ils ne le sont que partiellement et dans la plupart du temps d'une façon très arbitraire (Dichy, 1990)[295]. Pourtant, la voyellation (ou la vocalisation) intégrale d'un texte permet de garantir son interprétation, impliquant une analyse directe, rapide et fiable avec un minimum d'ambiguïtés morpholexicales et morphosyntaxiques (Mouelhi, 2008)[296]. Mais cela peut être justifié par le « côté énigmatique » de l'arabe afin d'intriguer le lecteur (Monteil, 1960)[297], mais également par « la loi du moindre effort » (Zipf, 1949)[298], puisque « *l'usage de la vocalisation réduit co*nsidérablement la rapidité de la *composition typographique qui ne dépasse pas les 60 mots à la minute, alors qu'avec les* machines à caractères latins on dépasse les 100 mots à la minute » (Zghibi, 2002)[299]. Cela implique que la saisie de toutes les voyelles brèves devient rapidement une opération pénible, fastidieuse et coûteuse en terme de temps et d'investissement. À cela s'ajoutent d'autres signes orthographiques, les diacritiques ; les consonnes portent souvent deux signes superposés, impliquant que « les ligatures esthétiques rendent pratiquement impossible la position exacte *de signes de vocalisation surtout lorsqu'il s'agit d'une superposition de trois ou quatre* lettres » (Zghibi, 2002)[300], en rappelant que l'écriture arabe connaît trois types de ligatures : les ligatures contextuelles, les ligatures linguistiques et les ligatures esthétiques.

Notre corpus ne fait pas exception à cette règle et est entièrement non-voyellé ; cela ne facilite pas notre travail d'analyse et de traitement des données. Pour nous affranchir de ce problème, l'utilisation d'un outil de voyellation automatique des textes arabes nous parait adéquate, d'autant plus qu'il permettra de rectifier le tir sur ce choix nuisible à l'alphabétisation et désastreux pour le traitement automatique de l'arabe. À notre connaissance, le développement de cet outil fiable et rapide, proposé par Ghénima (1998)[301] n'est pas encore disponible.

De ce fait, nous considérons que la voyellation du texte, plus précisément, la voyellation partielle du texte, sera fournie par l'extracteur morphosyntaxique, puisque sa conception repose

[295] L'écriture dans la représentation de la langue : la lettre et le mot en arabe
[296] Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî, p. 163
[297] L'arabe moderne, p. 42
[298] Human Behavior and the Principle of the Least Effort
[299] Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646, p. 164
[300] Idem, p. 164
[301] Un système de voyellation de textes arabes

sur la construction d'une grammaire d'identification des termes, en définissant exactement quelles voyelles devront être saisies, à quels endroits et pourquoi, sachant qu'il existe trois types de voyelles brèves : les voyelles morphologiques, les voyelles casuelles et les voyelles d'appui, dites aussi les voyelles de liaison (Mouelhi, 2008)[302].

4.1.3.2.2 Voyelle longue

Quant aux voyelles longues, elles correspondent à l'allongement phonétique des voyelles brèves, « fat*ḥ*a a ◌َ », « kasra i◌ِ » et « damma u◌ُ » ; mais, il est à noter ici que ces voyelles longues correspondent aux consonnes « *'*alif ا », « wâw و » et « yâ *'* ي ».

4.1.3.3 Signe diacritique

Les signes diacritiques complètent les consonnes et les voyelles de la langue arabe ; parmi ces signes, nous recensons la *šadda*, le sukûn, le tanwîn, la madda et la wasla.

5.1.1.1.1 Šadda ◌ّ

La « *šadda* ◌ّ» est la marque de la gémination consonantique et ressemble à un petit 3 couché ; elle se combine aux voyelles, brèves ou longues, en se plaçant au-dessus d'une consonne, permettant le remplacement d'un redoublement de consonnes dont la première est quiescente, c'est-à-dire, sans voyelle.

5.1.1.1.2 Sukûn ◌ْ

Le « sukûn ◌ْ», appelé aussi voyelle zéro ou signe de quiescence, est l'absence de voyelle ; en forme de zéro, il est toujours placé au-dessus de la consonne qui n'est pas vocalisée.

5.1.1.1.3 Tanwîn ◌ٌ◌ٍ◌ً

Le « tanwîn (fathatan ◌ً, dammatan ◌ٌ, kasratan ◌ٍ) », appelé aussi « nûnation », est le redoublement de la graphie des voyelles brèves « fatha ◌َ, damma ◌ُ, kasra ◌ِ» et est ajouté à la lettre finale d'un nom, d'un adjectif et d'un adverbe, comme la marque casuelle indéterminée.

5.1.1.1.4 Madda ~

La madda ressemble à un petit *'alif* couché et est placée au-dessus du *'alif* en formant le « *'*alif madda en chef آ», correspondant à une hamza suivie d'une voyelle longue « â », en rappelant que ce caractère n'est en fait qu'une variante de la consonne « *'*alif ا» comme indiqué supra (cf. 4.1.3.14.1.3.1 Consonne).

[302] Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî, p. 159

5.1.1.1.5 Wasla

La wasla ressemble à une damma et est placée au-dessus du *'alif* en formant le « *'alif* wasla en chef » ; elle est le signe de la « hamza de liaison », en rappelant que ce caractère n'est en fait qu'une variante de la consonne *'alif* comme indiqué supra (cf. 4.1.3.1 Consonne).

4.1.4 Mot

Le mot peut être considéré comme une suite de caractères entre deux blancs (Dubois, 1994)[303] et correspond à une suite de formes collées les unes aux autres en arabe (Dichy, 1990)[304]. Nous rappelons la graphie arabe et les différentes graphies du mot arabe.

4.1.4.1 Mot graphique arabe

Le mot graphique en arabe se compose de trois groupes d'éléments : d'une base, d'affixes (préfixe et suffixe) et de clitiques (proclitique et enclitique), supposant « une structure *d'objet complexe ap*pelait mot maximal » (Cohen, 1970)[305] : Cela implique, théoriquement, une étape préliminaire de séparation des formes (cf. 4.2.1 Segmentation), mais concrètement, cela n'est pas nécessaire puisque le dépouillement n'exige qu'une identification des unités terminologiques.

Cependant, une liste des proclitiques et des enclitiques, simples et/ou composés qui peuvent être agglutinés à des mots minimaux (base + affixe), ainsi qu'une liste des préfixes et des suffixes seront intégrées à la conception de l'extracteur morphosyntaxique pour des recherches ultérieures.

Quant à la base, généralement un radical triconsonantique combiné à des signes diacritiques (redoublement de consonnes, voyelles brèves et/ou longues, etc.), elle forme le noyau lexical et est éventuellement entourée d'extensions (Dichy, 1997)[306].

À partir d'une racine, d'une combinaison de voyelles, d'affixes et d'un schème verbal, nous obtenons un mot graphique arabe, soumis aux phénomènes de flexion, variation de la forme des mots en fonction de facteurs grammaticaux, et de dérivation, formation de nouveaux mots à partir de mots existants (Belguith et al., 2006)[307].

---

[303] Dictionnaire de linguistique et des sciences du langage, p. 312
[304] L'écriture dans la représentation de la langue : la lettre et le mot en arabe
[305] Études de linguistique sémitique et arabe
[306] Pour une lexicomatique de l'arabe, p. 295
[307] Analyse et désambiguïsation morphologiques de textes arabes non voyellés, p. 495

4.1.4.2 Graphies multiples du mot arabe

Dans la langue, les mots peuvent présenter plusieurs graphies, notamment les sigles, les abréviations et les homographies.

4.1.4.2.1 Sigle

Le sigle est considéré comme une réduction de nature morphologique, qui consiste à abréger de façon maximale un segment linguistique en ne retenant que les initiales de la totalité ou d'une partie des unités lexicales qui le composent (cf. 2.2.2.1.3 Siglaison) ; en face de ces signes, la première question qui se pose est de savoir s'il faut rétablir leur signification intégrale, mais sachant que cet emploi est motivé, porteur de sens et répond à des besoins particuliers, cela implique que le respect de la forme s'impose ici.

En arabe, ces processus sont très peu utilisés ; quant au domaine de la chimie, il s'agit d'un emploi récurrent, répondant aux besoins de communication afin de nommer un composé chimique, une méthode d'analyse chimique, … « *à tel point qu'ils* finissent souvent par être terminologisés » (Jaber, 2012)[308]. Par exemple, « كروماتوغرافيا الطبقة الرقيقة (ك ، ط ، ر) = *krûmâtûġâfiyâ* al*ṭabaqat* alraqîqat (k, *ṭ*, r) = chromatographie sur couche mince (CCM) ».

4.1.4.2.2 Abréviation

Très répandue dans les langues spécialisées, l'abréviation reproduit la forme initiale d'une unité lexicale ; comme les sigles, la forme est respectée.

En arabe, l'abréviation est plus employée que les sigles mais reste une pratique peu courante, sachant qu'« *un grand nombre d'abréviations représentent des translittérations d'autres* abréviations latines » (Mouelhi, 2008)[309]. Dans le domaine de la chimie, les abréviations sont d'un emploi fréquent pour les composés chimiques, par exemple 'Ph' pour 'phényl' (Kaczmarek, 2011)[310].

4.1.4.2.3 Homographie

L'homographie est la conformité, au niveau de la manifestation écrite, de deux ou plusieurs unités linguistiques distinctes, impliquant des ambiguïtés morpholexicales dans les

[308] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours, p. 169
[309] Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî, p. 230
[310] Nomenclatures française et polonaise de la chimie organique. Analyse comparative, p. 64

langues en général, en arabe en particulier, notamment en raison de sa graphie agglutinante et non-voyellée.

4.1.4.2.3.1 Ambiguïté et voyellation

En arabe, un mot non-voyellé est une forme, qui se voit associer plusieurs significations disjointes et mutuellement exclusives jusqu'à ce qu'elle soit voyellée ; sachant qu'en moyenne six voyellations sont possibles par mot (Debili, 2001)[311], cela implique une ambiguïté notoire dans un texte. Mais au-delà de ces ambiguïtés, nous considérons ces différentes formes comme des homographes en adoptant la notion d'homographie consonantique de Mouelhi :

> *« Nous appelons homographie consonantique toute conformité, au niveau de la manifestation écrite, entre deux ou plusieurs structures consonantiques de mots entièrement non vocalisés. Cette homographie peut disparaître dès lors qu'on vocalise partiellement (en plaçant les voyelles distinctives) ou totalement le mot homographe. Exemples : كتب ». (Mouelhi, 2008)*[312]

L'exemple cité (cf. Tableau 11 : Voyellations potentielles du mot « كتب » (Mesfar, 2008)) présente 15 voyellations possibles, correspondant à 16 homographes du mot « كتب » (Ouersighni, 2002)[313] :

[311] Traitement automatique de l'arabe voyellé ou non

[312] Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî, p. 230

[313] La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe : utilisation pour la détection et le diagnostic des fautes d'accord, p. 19

Voyellation	Translittération	Traduction	Catégorie grammaticale
كَتَبَ	*Kataba*	a écrit	Verbe, Accompli, Voix active, 3ème personne, masculin, singulier
كُتِبَ	*Kutiba*	a été écrit	Verbe, Accompli, Voix passive, 3ème personne, masculin, singulier
كَتَّبَ	*Kattaba*	a fait écrire	Verbe, Accompli, Voix active, 3ème personne, masculin, singulier
كُتِّبَ	*Kuttiba*	a été fait écrire	Verbe, Accompli, Voix active, 3ème personne, masculin, singulier
كَتِّبْ	*Kattib*	fais écrire	Verbe, Impératif, 2ème personne, masculin, singulier
كُتُبُ	*Kutubu*	des livres	Substantif, masculin, pluriel, nominatif, déterminé
كُتُبَ	*Kutuba*		Substantif, masculin, pluriel, accusatif, déterminé
كُتُبِ	*Kutubi*		Substantif, masculin, pluriel, génitif, déterminé
كُتُبٌ	*Kutubũ*		Substantif, masculin, pluriel, nominatif, indéterminé
كُتُبٍ	*Kutubĩ*		Substantif, masculin, pluriel, génitif, indéterminé
كَتْبُ	*Katbu*	un écrit	Substantif, masculin, singulier, nominatif, déterminé
كَتْبَ	*Katba*		Substantif, masculin, singulier, accusatif, déterminé
كَتْبِ	*Katbi*		Substantif, masculin, singulier, génitif, déterminé
كَتْبٌ	*Katbũ*		Substantif, masculin, singulier, nominatif, indéterminé
كَتْبٍ	*Katbĩ*		Substantif, masculin, singulier, génitif, indéterminé

Tableau 11 : Voyellations potentielles du mot « كتب» (Mesfar, 2008)[314]

Pour le mot voyellé, Mouelhi parle d'« homographie globale », considérant que « les mots homographes ont le même agencement de consonnes et le même vocalisme » (Mouelhi, 2008)[315] ; mais, ce cas n'est pas présent dans notre corpus, puisque nos textes ne sont pas voyellés.

Cette ambiguïté liée à la voyellation des mots est d'autant plus perceptible dans les emprunts comme 'oxygène' ; en l'absence de voyelles brèves, les emprunts se trouvent attribuer plusieurs graphies, parfois fantaisistes, en cumulant les voyelles longues, afin de se rapprocher le plus précisément possible de la langue source : « أكسجين = 'uksijîn » ou « أوكسجين = 'ûksijîn » ou « أكسجن = 'uksijin » ou « أوكسيجين = 'ûksîjîn » ou encore « أوكسجن = 'ûksijin » (Jaber, 2012)[316]. Nous considérons ces différentes formes comme des homographes, d'autant plus que certains emprunts ne sont pas codifiés ou se trouvent contestés.

4.1.4.2.3.2 Ambiguïté et agglutination

En arabe, le mot maximal (base + affixe + clitiques) est segmenté jusqu'à obtenir une base, en passant par un mot minimal (base + affixe) ; selon les différentes formes,

[314] Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard, p. 56
[315] Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî, p. 230
[316] Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours, p. 248

l'interprétation peut varier et générer une ambigüité aussi bien sur son sens que sur sa fonction dans la phrase, impliquant une ambiguïté notoire dans un texte. Mais au-delà de ces ambiguïtés, nous considérons ces différentes formes comme des homographes en adoptant la notion d'ambiguïté agglutinante de Mouelhi (2008)[317].

4.1.4.2.3.3 Graphie multiple

Par graphie multiple, nous entendons les ambiguïtés morpholexicales qui ne sont dues ni à la voyellation ni à l'agglutination, mais au mot graphique, notamment son identification et sa catégorisation.

4.2 Analyse des formes

Pour analyser les formes, il faut procéder préalablement au découpage du texte en mots, la segmentation. Puis, chaque terme est identifié et se voit attribuer son étiquette grammaticale et des listes d'index et de concordances sont éditées. Enfin, nous exposons les éléments essentiels de la lemmatisation des formes de notre corpus, avec la spécificité de langue arabe constituée de noms, de verbes et de particules.

4.2.1 Segmentation

La segmentation est une procédure qui permet de préparer le corpus à l'analyse lexicale, en délimitant les segments de chaînes de caractères en éléments constituant différents niveaux structurels : paragraphe, phrase, syntagme, mot graphique, forme, morphème.

4.2.1.1 Segmentation lexicale

La segmentation lexicale, appelée aussi découpage des mots ou itémisation (en anglais word segmentation ou tokenization), est basée habituellement sur la reconnaissance d'une suite de caractères entre deux blancs et/ou la présence d'un signe de ponctuation, permettant de découper un texte en mots (tokens).

En arabe, différents types de mots existent, puisque la plupart des textes arabes intègrent aux caractères arabes des caractères latins. De ce fait, certains linguistes identifient neuf types de mots, « séparateurs, mot à caractères arabes, mot à caractères latins, nombre entier, caractère(s) arabe(s) + chiffre(s), caractère(s) latin(s) + chiffre(s), caractère(s) arabe(s) +

[317] Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî, p. 237

caractère(s) latin(s), caractère(s) arabe(s) + caractère(s) latin(s) + chiffre(s), nombre décimal, signe(s) de ponctuation, signe(s) de ponctuation forte, abréviation (lettre suivie *d'un* point *suivie d'une lettre suivie d'un point* » (Ouersighni, 2002)[318], tandis que d'autres se limitent à deux types :

- *« I- Le mot arabe : C'est une chaîne formée uniquement de caractères arabes, appartenant à l'alphabet arabe. […] Les lettres arabes sont :* ا أ إ ؤ ء ئ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ى ي
- *ii- Le mot non arabe : toute chaîne formée de caractères non arabes, des caractères latins, des chiffres, des ponctuations (les textes arabes contiennent un mixte de ponctuations arabes et latines : , ? ; . : / % * ؟ . , ' / " } { « # ' { ([- | ` _ \ ^)] = + etc). » (Abbes, 2004)*[319]

Notre corpus ne fait pas exception en contenant aussi bien des caractères arabes que des caractères latins ; mais, à cela s'ajoute la spécificité de notre domaine, qui est de combiner les caractères latins à la ponctuation et/ou aux chiffres, notamment pour nommer un composé chimique par exemple « 2–أمينو -1- (4-نيتروفينيل)-1، -3بروبان ثنائي أول 2 (R2, R1) = (1R,2R)-(−)-2-Amino-1-(4-nitrophenyl)-1,3-propanediol ». Par conséquent, il faut également prendre en compte dans la typologie des mots les séquences « caractère(s) grec(s) + caractère(s) arabe(s) + chiffre(s) + ponctuation(s) » et « caractère(s) grec(s) + caractère(s) latin(s) + chiffre(s) + ponctuation(s) ».

Même si nous prétendons avoir certaines facilités avec les outils informatiques, nous ne possédons ni les compétences ni les programmes nécessaires pour adapter le segmenteur à notre typologie de mots ; cette proposition peut être prise en compte dans un travail ultérieur.

En attendant que ce travail voie le jour, nous procédons au découpage du texte manuellement pour la simple et bonne raison qu'il nous parait plus rapide que de requérir à un segmenteur automatique ou semi-automatique pour notre travail, d'autant plus qu'il faut, dans tous les cas, au préalable, détecter manuellement toutes ces séquences constituées de caractère(s) grec(s) et/ou de caractère(s) arabe(s) et/ou de caractère(s) latin(s) et/ou de chiffre(s) et/ou de ponctuation(s). En effet, même s'il existe un système pour nommer les composés chimiques (UICPA), ces dénominations, en plus de ne pas être respectées et d'être parfois fantaisistes, restent difficiles à traduire en informatique par la présence aléatoire d'une ou plusieurs espaces

[318] La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe : utilisation pour la détection et le diagnostic des fautes d'accord

[319] La Conception et la réalisation d'un concordancier électronique pour l'arabe, p. 148

dans la dénomination d'un composé chimique, contrairement au français ; par exemple « حمض 2– أمينو 3– هيدروكسي بوتان ثنائي أويك = acide 2-amino-3-hydroxybutanedioïque »

De plus, les travaux sur la segmentation pour l'arabe ne sont pas nombreux ; mais citons tout de même, le système STAr (Belguith et al., 2005)[320], basé sur l'exploration contextuelle des signes de ponctuation et de certaines particules. Le segmenteur AraSeg nous parait également un bon outil pour le découpage des textes arabes avec un rendement de plus de 90% (Mouelhi, 2008)[321]. Cependant, le défaut de ces segmenteurs est qu'ils ne sont pas libres d'accès.

4.2.1.2 Segmentation syntaxique

La segmentation syntaxique (chunking) consiste à isoler les différents constituants du texte en unités indépendantes et supérieures aux mots, comme les propositions, syntagmes … Il s'agit là du travail que nous nous proposons de réaliser dans cette thèse (6.2.26.2.2 Patron morphosyntaxique).

Il existe une troisième segmentation, la segmentation morphologique ; elle permet quant à elle d'isoler les différents constituants des mots en unités distinctes et plus petites, qui sont les morphèmes. En arabe, cette segmentation morphologique correspond au découpage en pré-base, base et post-base (Abbes, 2004)[322].

4.2.2 Étiquetage

L'étiquetage correspond à la production de l'ensemble des étiquettes morphosyntaxique candidates pour chacun des mots identifiés à partir desquelles une liste d'index et de concordances sont éditées.

Habituellement, les linguistes en général, les terminologues en particulier, recourent à un étiqueteur ; généralement, son fonctionnement est probabiliste en se fondant sur l'utilisation d'arbres de décision mais également en se servant d'un dictionnaire de petite taille et le système comprend parfois un module de segmentation (Namer, 2003)[323].

[320] Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules

[321] AraSeg* : un segmenteur semi-automatique des textes arabes

[322] La Conception et la réalisation d'un concordancier électronique pour l'arabe, p. 116

[323] Le modèle Lstat : ou comment se constituer une base de données morphologique à partir du Web, p. 91

Nous distinguons deux étapes d'étiquetage, la première pour l'identification des termes du domaine et la seconde pour l'attribution des catégories grammaticales.

4.2.2.1 Identification des termes de la chimie

Le dépouillement du corpus a pour objectif d'établir une liste de candidats termes, permettant de déterminer les termes susceptibles de figurer comme vedette dans les fiches terminologiques. Le repérage de ces termes dans les textes dépouillés consiste à reconnaître et à sélectionner des désignations.

Généralement, les terminologues disposent de dépouilleurs terminologiques leurs fournissant une liste de candidats termes accompagnés de leurs contextes (Tellier, 2008)[324] ; ces outils déterminent la spécificité des mots apparaissant dans un corpus spécialisé en comparant leur fréquence dans ce corpus à un corpus de référence : plus le mot est spécifique, plus il est susceptible d'être un terme du domaine. Bien qu'ils procèdent à une estimation statistique, souvent juste, mais non infaillible, ces outils offrent un premier calcul de l'importance des unités dans le corpus en se basant sur la forme des termes (termes simple ou termes complexes) et sur leur catégorie grammaticale (nom, verbe, adjectif). Cependant, le dépouillement de notre corpus a été réalisé manuellement, en l'absence et/ou en raison de la difficulté d'accès aux outils de traitement pour l'arabe, impliquant que la sélection des termes est manuelle et s'annonce assez fastidieuse.

Au fur et à mesure de la lecture des textes de notre corpus, nous repérons les séquences de mots constituant des unités de signification dans notre domaine, correspondant à des termes, simples et complexes. Nous suivons une approche bottom-up, en partant des termes existants dans les textes du corpus, plus adaptée à l'élaboration d'une ressource terminologique, étant donné que le nombre de termes dans un vocabulaire est plus petit que le nombre de mots dans le lexique général. Cette démarche s'avère pertinente pour notre objectif de fouille de textes permettant ainsi d'identifier les informations à extraire.

Nous recensons une liste de ces candidats termes qui nous paraissent spécifiques au domaine ; mais la question est de savoir quels termes retenir et quels critères de sélection pour ces candidats termes.

Le premier critère est de savoir si les mots appartenant à la langue générale mais possédant au moins un sens spécialisé sont des candidats termes. Prenons l'exemple du terme « ماء = ma '=

[324] Verbes spécialisés en corpus médical : une méthode de description pour la rédaction d'articles terminographiques

eau », appartenant à la langue générale : ce terme, appelé aussi monoxyde de dihydrogène, oxyde d'hydrogène, hydrogénol, hydroxyde d'hydrogène, oxyde dihydrogéné, est un composé chimique, constitué de molécules de formule chimique H_2O, donc composé de deux atomes d'hydrogène et d'un atome d'oxygène, et peut être une base ou un acide ; il s'agit là d'une signification particulière de l'eau en chimie. De ce fait, nous considérons terme de chimie tout mot possédant un sens spécialisé dans le domaine, s'agissant d'un objet directement lié à la chimie et/ou qui fait référence à une activité et/ou à un fait du domaine.

Cependant, l'identification des termes du domaine, notamment le découpage du terme, n'est pas anodine :

> *« Il y a là, en fait, un double problème : celui du repérage d'un terme - parmi, par exemple, les syntagmes d'un texte scientifique, quels sont ceux qui représentent des termes, et quels sont ceux qui ne constituent pas des termes ? - et celui du découpage du terme - quand le terme est une unité terminologique complexe, quels en sont les éléments qui peuvent constituer, isoler, eux et eux seuls, des termes ? » (Lelubre, 1992)*[325]

Prenons l'exemple du terme « بيرمنغنات البوتاسيوم = bîrma*nġanât albûtâs*yûm = permanganate de potassium » : cette unité terminologique complexe est composée des termes 'permanganate' et 'potassium'. De ce fait, l'identification des termes du domaine peut dépendre de l'autorité de celui qui a constitué la terminographie, de la qualité de la documentation exploitée et citée, de la collaboration d'experts et de lexicographes, de la complétude des données portées sur les fiches, qui varie d'un domaine à l'autre, de la mise à jour des données et de l'existence d'un mécanisme de validation interne et/ou externe (Humbley, 1996)[326].

D'autre part, certains candidats termes sont transparents, indiquant clairement de quoi il s'agit par leur dénomination, comme « oxygénation », tandis que d'autres sont opaques comme « alkylation » (Frérot, 2000)[327]. Cela dit, même si notre connaissance du domaine permet de les identifier sans recourir à leur définition, par prudence, tant au niveau de connaissance des auteurs et que celui des lecteurs, nous avons vérifié les définitions des termes, transparents et opaques, ainsi que les concepts auxquels ils renvoient.

Quant à la fréquence des occurrences d'un candidat terme, critère important, ce n'est cependant pas une condition sine qua non pour la validation d'un terme, d'autant plus que le choix de ces candidats termes est lié à une double contrainte de pertinence :

[325] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation
[326] La légitimation en terminologie, p. 125
[327] Vitamines, Carotène et Polyphénols

- *Pertinence vis-à-vis du corpus. Il s'agit de retenir et de décrire des structures lexicales qui présentent des caractéristiques à la fois spécifiques et stables ; c'est à ce stade qu'intervient la validation par l'expert.*
- *Pertinence vis-à-vis de l'application. Les unités finalement retenues doivent l'être en fonction de l'utilisation dans l'application visée, qui s'exprime en termes d'économie et d'efficacité. La validation est à chercher du côté des utilisateurs de l'application. (Bourigault et al. 1999)* [328]..

Mais en adoptant la proposition de Chukwu (1998)[329], nous considérant comme « terminochimiste » en opposition au « terminolinguiste » et nous désignant comme détenteurs des termes de notre domaine, nous validons notre liste de candidats termes. De ce fait, nous n'effectuons pas de seconde sélection, qui permet habituellement de retenir seulement les termes pertinents pour les recherches ou pour les produits terminologiques. De plus, nous ne passons pas à l'attestation de l'usage ; nous ne vérifions pas dans d'autres sources que ces termes se retrouvent ailleurs dans la spécialité, qu'ils désignent toujours les mêmes concepts et que ces concepts appartiennent bel et bien à la spécialité dont on étudie la terminologie. Enfin, nous ne collaborons pas avec un expert afin de recourir à son approbation pour l'établissement de la liste de candidats termes.

Une fois la liste de termes validée (cf. Annexe 2 : liste des termes du corpus), nous l'encodons, dans un premier temps, dans un éditeur, sous forme de tableau constitué de cinq colonnes dans lesquelles nous indiquons le terme, sa référence, son équivalent, son contexte et éventuellement des remarques et/ou des observations, nous facilitant le repérage des termes mais également le va-et-vient entre le corpus et la liste :

[328] Pour une terminologie textuelle, p. 31

[329] Dépouillement de corpus à des fins terminologiques dans un univers dépendant du temps, p. 4

Terme	Référence	Équivalent	Contexte	Remarque
تكـاثف ألدولي	Réf. 1.a	réaction aldolique	لقد تم إصطناع عدة مركبات كيميائية بالإعتماد على إستراتيجيـة التكـاثف الألدولي	UTC, identification
صوديوم	Réf. 1.b	sodium	و تجفف فوق كبريتات الصوديوم Na_2SO_4	UTS, emprunt
جو من الأزوت	Réf. 1.c	Atmosphère d'azote	يتم هذا التفاعل في جو من الأزوت	UTC, complétive
قطر	Réf. 1.b	Distiller	يقطر ثلاثي إيثيل أمين فوق هيدروكسيد البوتاسيوم	UTS

Tableau 12 : Extrait de la liste des termes du corpus

Nous avons privilégié les contextes contenant les attestations simples et claires sur les termes, les cooccurrents des termes et leur comportement linguistique, les dérivés morphologiques et sémantiques, les relations paradigmatiques avec d'autres termes et les renseignements sur le domaine de spécialité (Pimentel, 2012)[330]. Ces contextes illustrent l'usage concret du terme dans les textes du corpus et servent également à repérer les structures argumentales des verbes.

Dans un second temps, nous rédigeons les fiches terminologiques de chaque terme du corpus (cf. Annexe 3 : Fiche terminologique). À l'aide d'un concordancier, nous avons identifié les termes en nous fondant sur les index et les concordances des formes du corpus, considérant que l'index est une simple liste énumérant les mots utilisés dans un texte suivis des références permettant de les retrouver et que la concordance fournit toutes les attestations d'un mot dans un texte, soit tous les contextes dans lesquels le mot est identifié (Muller, 1969)[331].

Cet index permet de détecter des fautes de frappe, des graphies erronées ou douteuses mais aussi de classer par ordre alphabétique et/ou hiérarchique les termes, afin d'éviter notamment les doublons dus au genre et/ou au nombre et/ou, pour l'arabe, au cas.

L'index alphabétique est plus fréquent mais l'index hiérarchique, basé sur le nombre d'occurrences et les formes, présenté par ordre décroissant de fréquence, permet la comparaison des fréquences en mettant en valeur certaines formes, notamment les cas d'homographie et facilite le repérage du vocabulaire spécifique (Labbé, 1990)[332].

[330] Description de verbes juridiques au moyen de la sémantique des cadres
[331] La statistique lexicale, p. 33
[332] Normes de saisie et de dépouillement des textes politiques, p. 36

Pour notre corpus, nous avons utilisé le concordancier « AntConc, version 3.4.4.w » qui permet d'obtenir une liste des mots, l'index lexical du texte, soit en fonction de leur fréquence (index hiérarchique), soit par ordre alphabétique (index alphabétique). Cette liste est un peu farfelue et n'est pas un index alphabétique, en raison de l'emploi de caractères latins et arabes simultanément ainsi que de la nomenclature chimique, elle valide certains de nos mots-clés ainsi que nos termes :

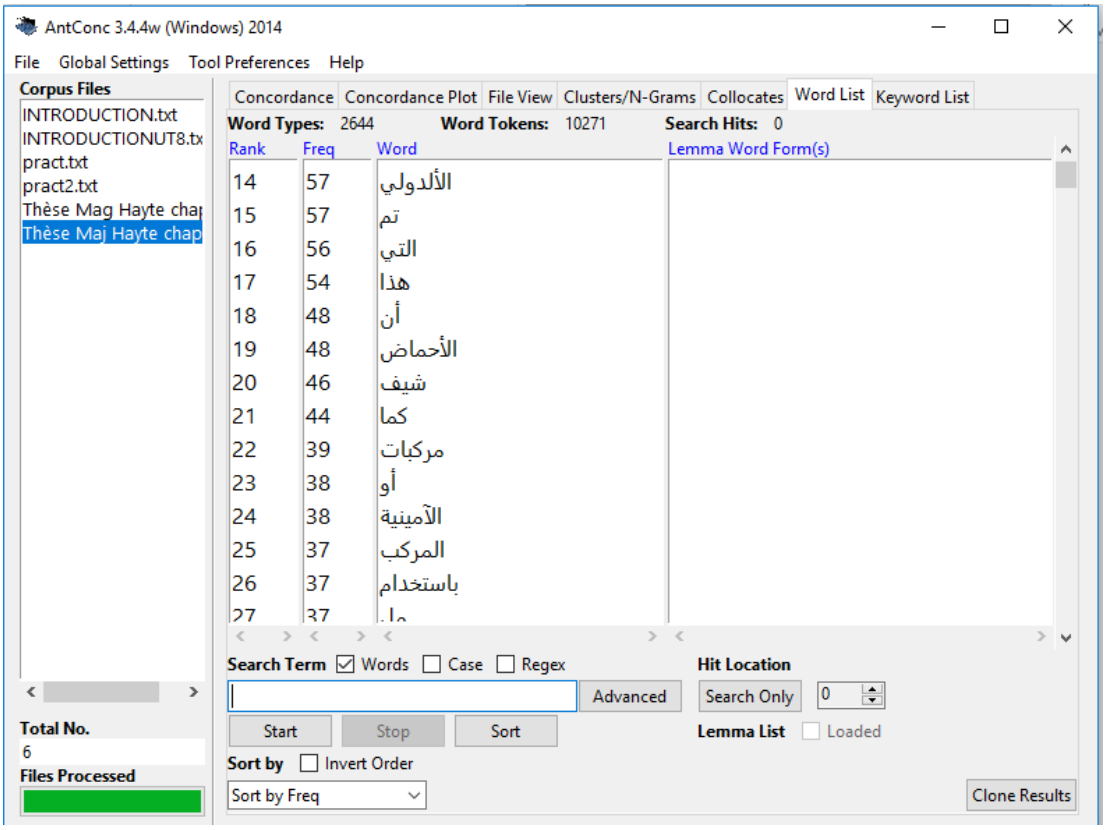


Figure 23 : Extrait de *l'index lexical à partir d'AntConc*

En cas de synonymie et/ou de variantes terminologiques, notamment des variantes orthographiques, le terme choisi correspond au terme le plus fréquemment rencontré dans le corpus et nous indiquons dans sa fiche ses différentes variantes, comme le cas du terme « نتروجين = nitrûjîn » et « أزوت = âzût », issu respectivement de l'anglais « nitrogen » et du français « azote ». Dès le début de la lecture de notre corpus, nous avons remarqué l'existence de plusieurs formes pour exprimer une seule notion, notamment pour les emprunts et les mots composés, dont les langues sources sont le français et l'anglais, puisque « les terminologies *scientifiques et techniques en arabe sont toutes tributaires, d'une façon ou d'une autre, des*

terminologies établies en anglais ou en français, qui constituent les terminologies de référence » (Lelubre, 1992)[333].

D'autre part, AntConc propose la fonction « collocates » permettant d'identifier les cooccurrents des mots ; mais, comme l'index lexical, elle est plutôt farfelue :

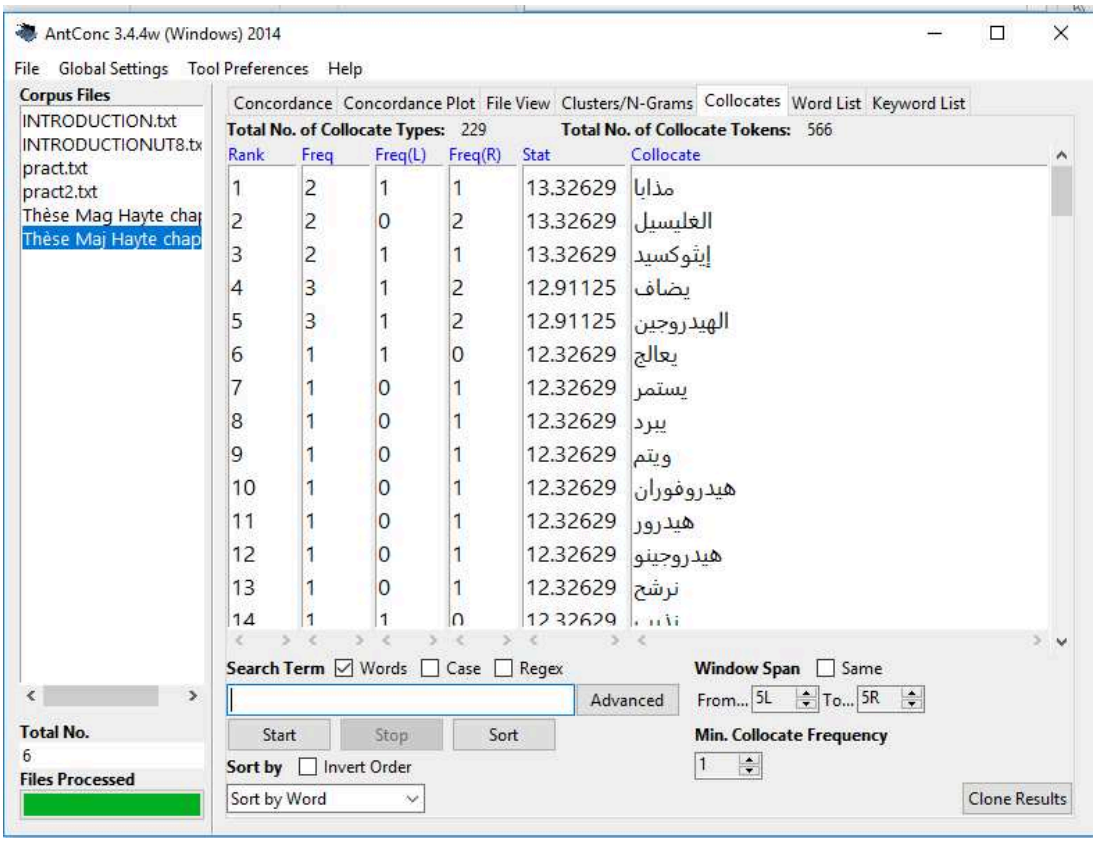


Figure 24 : *Extrait de la liste des mots à partir d'AntConc*

En plus de la possibilité d'erreur dans l'analyse textuelle, ce logiciel n'est pas spécifique à l'arabe. Pourtant, nous avons utilisé des expressions régulières, permettant ainsi de rechercher toutes les formes des termes retenus, mais en réponse à ces requêtes, nous obtenons beaucoup de bruit, en raison des ambiguïtés de la non-voyellation et de l'agglutination. Malheureusement, il n'y a pas de large choix concernant les outils informatiques pour le traitement de l'arabe, tant par sa complexité que par le manque de communication de ses chercheurs, d'autant plus que les outils en ligne sont rares. Mais, citons tout de même le concordancier AraConc (Abbes, 2004)[334], basé sur une reconnaissance morphosyntaxique des mots, conformément aux exigences d'une concordance spécifique à l'arabe. À notre connaissance, ce logiciel n'est pas en accès libre.

[333] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation
[334] La Conception et la réalisation d'un concordancier électronique pour l'arabe

De ce fait, nous proposons le logiciel « Kawâkib, version bêta publique » (Gaubert, 2010)[335], disponible sur le site http://mail.ifao.egnet.net:8080/kawakib/, permettant de rechercher les 25 racines les plus fréquentes, d'identifier une racine trilitère et sélectionner les mots outils du texte. L'utilisation de ce logiciel avec notre corpus ne présente que peu de résultats pertinents ; mais, pour la concordance, cet outil s'avère intéressant.

Quant à la concordance, elle consiste à identifier le nombre d'occurrence d'un mot, à l'aide d'un concordancier, avec un contexte significatif d'au minimum 3 mots devant et derrière, sauf si la forme recherchée se trouve au début ou à la fin d'un texte ou sur toute autre césure ; dans ces cas, seul le contexte significatif sera pris en compte. Voici par exemple un extrait des concordances du « كلور = chlore » dans notre corpus :

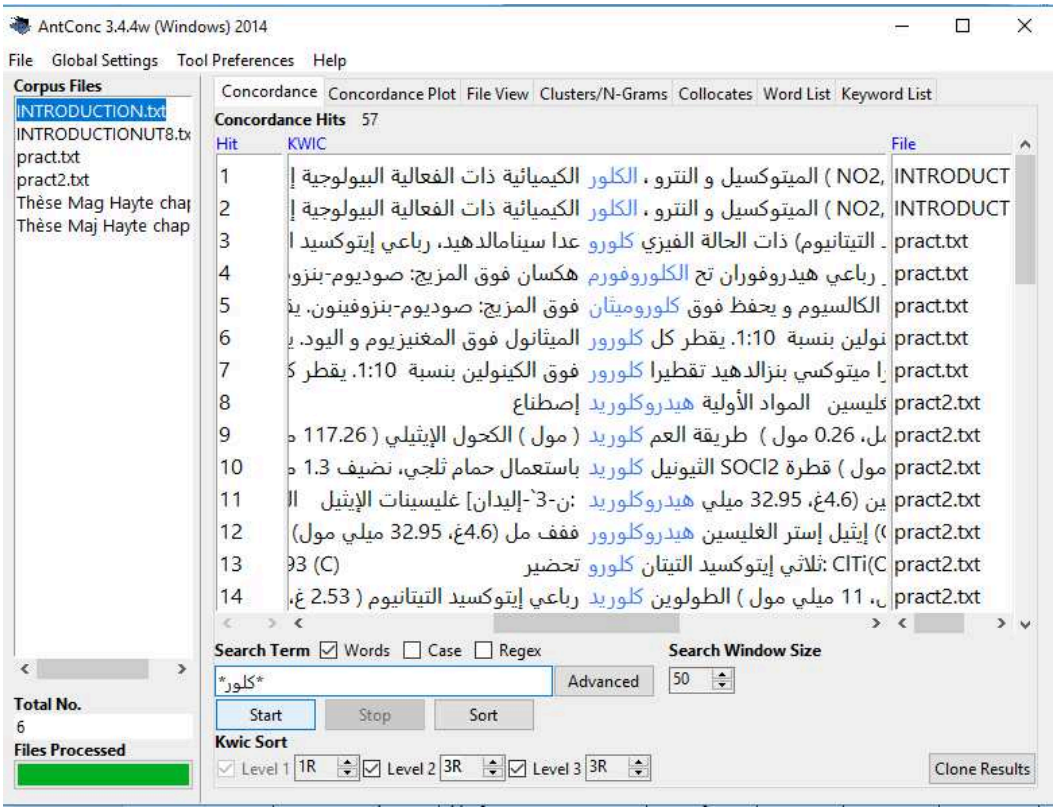


Figure 25 : Extrait de la concordance du terme « كلور= klûr = chlore » à partir d'AntConc

Le logiciel « AntConc » adopte une présentation classique où le terme est centré dans un micro-contexte d'une ligne ; les références sont fournies à droite du listing dans des colonnes prédéfinies « File » et le nombre de formes différentes (vocabulaire) est indiqué en haut à gauche du listing « Concordance Hits ».

[335] Kawâkib, une application web pour le traitement automatique de textes arabes

Grace à un concordancier, il est possible aussi d'éditer les concordances des formes d'homographie ainsi que la concordance inverse, c'est-à-dire, le mot précédant la forme concernée dans un classement alphabétique des attestations en contexte, permettant d'identifier nos termes qui sont souvent des syntagmes, qu'ils soient des syntagmes avec plusieurs éléments ou avec une préposition. Cependant, le logiciel « AntConc » n'étant pas spécifique à l'arabe, seule la concordance des formes, en général, des formes d'homographie en particulier est possible. Quant à l'étude des termes eux-mêmes, notamment des couples nom/verbe, le logiciel « Kawâkib » est plus pertinent et permet de repérer les noms morphologiquement reliés aux verbes. Notons que l'emploi des noms et des adjectifs est prédominant dans les domaines de spécialité, impliquant que les nominalisations sont beaucoup plus fréquentes que le verbe correspondant. Voici par exemple un extrait de l'analyse de la racine "ق ط ر" dans notre corpus :

Figure 26 : Extrait de l'analyse de la racine "ق ط ر" à partir de Kawâkib

Nous avons ainsi étudié plusieurs dizaines de couples nom/verbe (Albeiriss, 2017)[336] en recherchant la racine à partir du logiciel « Kawâkib » ; nous disposons ici d'un puissant outil d'aide à la lemmatisation, opération à laquelle nous consacrons la dernière partie de notre chapitre (cf. 4.2.3 Lemmatisation).

4.2.2.2 Étiquetage grammatical

[336] Terminologie basée sur corpus : les verbes de la chimie en arabe

L'étiquetage grammatical consiste à attribuer à chaque terme une catégorie grammaticale parmi les différentes catégories possibles, en fonction du contexte, sachant qu'il faudra parfois les désambiguïser (Cabré, 2008) [337].

En arabe, la langue est constituée de trois catégories : les noms, les verbes et les particules ; chacune de ces catégories se décline en plusieurs sous-catégories : adjectif, préposition, pronom… (Kouloughli, 1994)[338].

La qualité de cet étiquetage est étroitement liée à la qualité du texte écrit et particulièrement à sa vocalisation ; mais notre corpus étant non voyellé, avant d'arriver au choix du jeu d'étiquettes (verbe, nom, adjectif...), il faudra estimer la vocalisation possible de chaque terme d'autant plus qu'elle n'est pas forcément unique (Abbes, 2004) [339].

Les travaux sur l'étiquetage grammatical de l'arabe voyellé en général, de l'arabe non voyellé en particulier, ne sont pas très nombreux (Mourad et al, 2008)[340] en raison de la grande complexité morphologique de l'arabe et de son système alphabétique consonantique ; mais citons tout de même :

- Les travaux de Beesley (1996)[341] en collaboration avec Lauri Karttunen, basés sur « finite state technologie : FST » développée à Xerox et regroupant 4930 racines et 400 modèles qui permettent de produire 90000 lexèmes.
- Les travaux de Debili et Souissi (1998)[342], basés non pas sur les algorithmes préconisés pour le français et/ou pour l'anglais mais sur un nouveau jeu d'étiquettes grammaticales, permettant une diminution de l'ambiguïté de départ et un élargissement de la portée des règles de succession en associant des étiquettes aux formes non minimales de l'arabe ; les résultats atteignent des seuils de résolution de 91% mais utilisent un jeu de plus de 1700 étiquettes grammaticales.
- Les travaux de Khoja (2001)[343], dérivés du système d'étiquetage BNC de l'anglais, qui est une combinaison de données statistiques et des règles techniques obtenant des résultats encourageants et mettant un jeu de 131 d'étiquettes.

[337] Constituer un corpus de textes de spécialité
[338] Grammaire de l'arabe d'aujourd'hui, p. 3
[339] La Conception et la réalisation d'un concordancier électronique pour l'arabe, p. 131
[340] Nouvelles ressources et nouvelles pratiques pédagogiques avec les outils TAL, p. 3
[341] Arabic Finite-State Morphological Analysis and Generation
[342] Etiquetage grammatical de l'arabe voyellé ou non
[343] APT: Arabic Part-of-speech Tagger

- Les travaux de Diab et al. (2004)[344], basés sur des outils pour l'anglais, mettant en évidence un système qui permet la segmentation, notamment la segmentation des clitiques, l'étiquetage grammaticale (POS) et l'annotation de phrases de base en texte arabe ; les résultats atteignent une précision de 95,49% pour le POS.
- Les travaux de El Jihad et Yousfi (2005)[345], basés sur les modèles de Markov cachés, présentant un système qui permet un étiquetage morphosyntaxique, en utilisant un corpus d'apprentissage étiqueté manuellement ainsi qu'un jeu de 52 étiquettes de nature morphosyntaxique et subissant une amélioration par la réestimation de ces paramètres.
- Les travaux de Ghoul (2011)[346], basés sur l'adaptation de l'outil « TreeTagger » pour la langue arabe.

L'étiquetage est réalisé sur Xerox ; ce choix est influencé en grande partie par la facilité d'accès du logiciel, sur l'adresse www.xrce.xerox.com/research/mltt/arabicm. Mais l'analyse aurait pu aussi être réalisée avec n'importe quel autre logiciel d'étiquetage grammatical, puisque les étiqueteurs présentés sont plus au moins équivalents. De ce fait, nous n'avons pas de préférence pour un en particulier, l'essentiel étant d'affecter à chaque terme la bonne catégorie grammaticale en prenant en compte les spécificités de notre corpus. Cependant, même si ce logiciel utilise des règles à large couverture, il génère un taux assez élevé d'ambiguïtés lexicales ; par exemple, l'analyse du mot « تشكّل = taššakkala = se former » propose 61 solutions, prenant en compte la voyellation et l'agglutination :

[344] Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks

[345] Étiquetage morpho-syntaxique des textes arabes par modèle de Markov caché.

[346]Outils génériques pour l'étiquetage morphosyntaxique de la langue arabe : segmentation et corpus d'entrainement

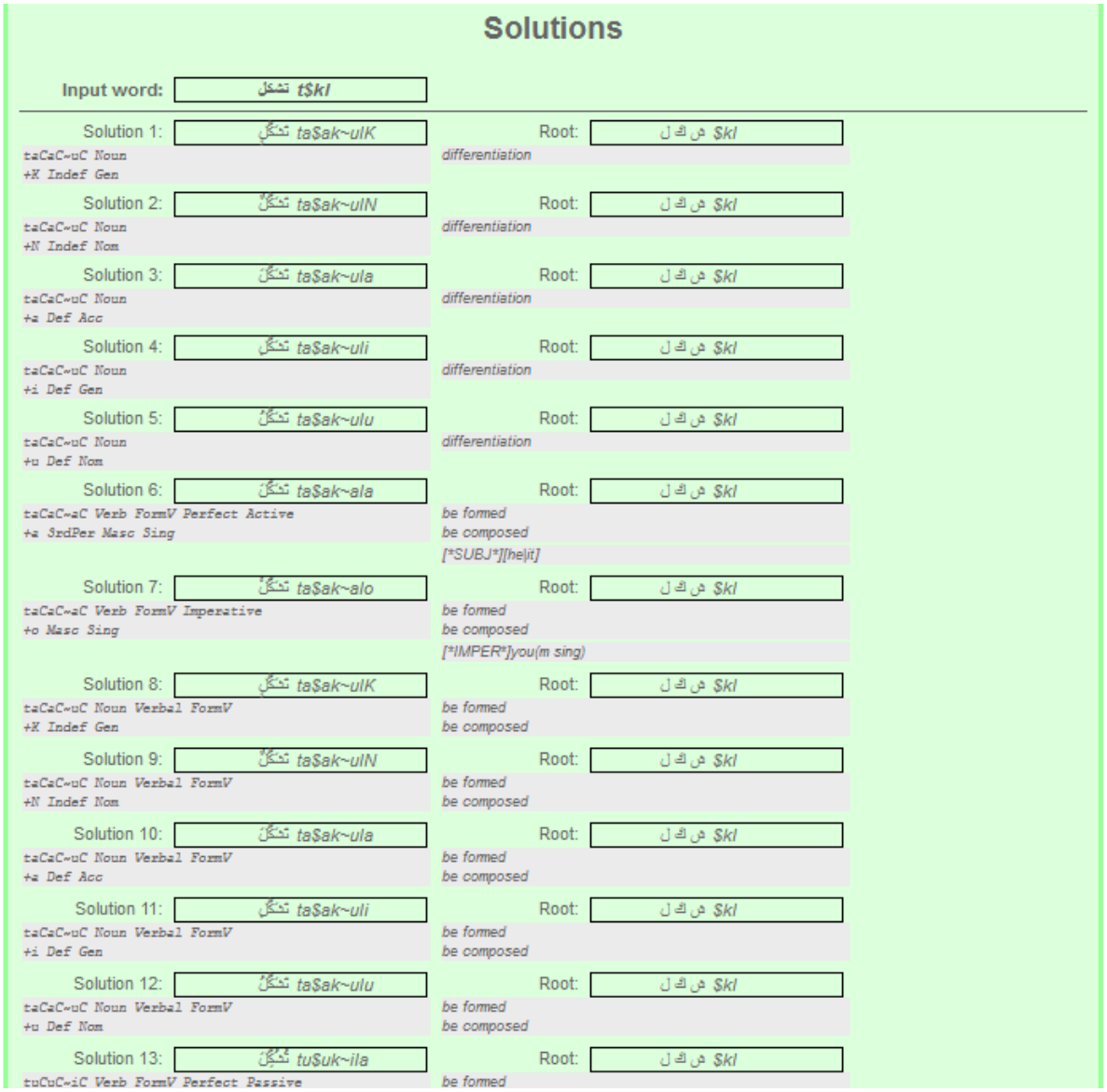


Figure 27 : Extrait de l'analyse du mot « تشكّل= *tašakkala* = se former » à partir de l'étiqueteur de XE-ROX

Quant aux emprunts, ils ne sont pas identifiés par l'étiqueteur :

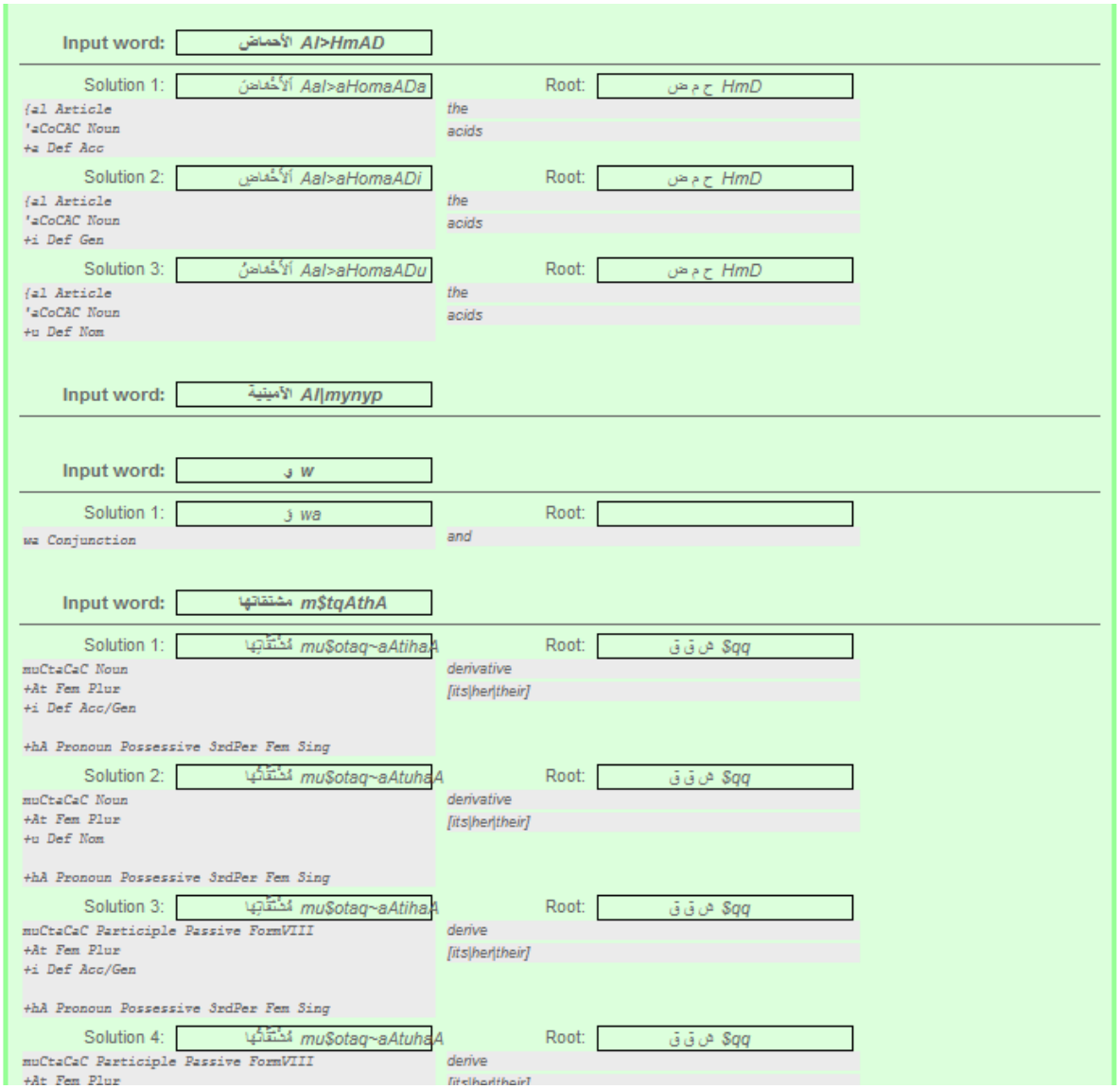


Figure 28 : Extrait de l'analyse du mot « *امينية = 'amî*niyat = amines » à partir de l'étiqueteur de XE-ROX

De ce fait, une vérification des résultats de cet étiqueteur, avant la lemmatisation, est nécessaire afin d'identifier la séquence graphique obtenue par l'affectation de la catégorie grammaticale correspondante et d'évaluer la validité de la catégorie affectée par l'étiqueteur en fonction de la forme du mot en présence. En cas d'ambiguïté et de présence de différente(s) solution(s), la catégorie est choisie par l'opérateur, sachant que la catégorie choisie par l'étiqueteur est à la fois la plus vraisemblable du point de vue de la graphie du mot et la plus fonctionnellement proche de la catégorie rejetée (Namer, 2003)[347]

Avec la grammaire d'identification que nous construisons, nous proposons quelques pistes d'améliorations pour l'étiquetage grammatical (6.2.36.2.3 Règle de grammaire).

[347] Le modèle Lstat : ou comment se constituer une base de données morphologique à partir du Web, p. 92

4.2.3 Lemmatisation

La lemmatisation consiste à attribuer une forme générique, un lemme, à chacun des mots découpés du texte lors de la segmentation, en regroupant sous une entrée unique les formes correspondant à un même signifiant. Cela permet, notamment de regrouper toutes les flexions d'un même verbe ainsi que les variations de genre et de nombre des noms, et ciblant de fait le sens de l'information recherchée ; les lemmes correspondent aux entrées des dictionnaires.

Ce traitement n'est pas du goût de tous les chercheurs et s'est vu vivement critiqué dans le passé (Labbé, 1990)[348]. Mais, en plus de paraître nécessaire et de comporter de nombreux intérêts en travaillant sur la reconnaissance des formes dans le texte, il offre à l'arabe, langue à morphologie complexe (Dichy, 1990)[349]), de lever les ambiguïtés pesant sur certaines de ses formes, en raison de sa non-voyellation, de son agglutination et de la richesse de son mot graphique. Cependant, nous déplorons l'absence d'un équivalent attesté en arabe pour la lemmatisation et adoptons la proposition de Mouelhi (2008)[350] « أصلمة = ʾaṣlama », comme son équivalent arabe.

Cependant, la lemmatisation est plus complexe, voir plus délicate en arabe par sa spécificité ; l'analyse morphologique s'intéresse, en arabe comme pour les autres langues, aux formes des mots et consiste à réduire le lexique aux seules informations non calculables (racines) et à utiliser des règles pour connaître le reste des informations. Elle doit prendre en compte les propriétés fondamentalement sémantiques des règles morphologiques, s'organisant selon une hiérarchie linguistiquement motivée, subordonnée à des exceptions que le système doit être à même de prendre en compte. Malheureusement, l'arabe ne dispose pas d'assez de ressources, notamment des ressources gratuites, pour diverses raisons. Néanmoins, nous pouvons citer :

- L'analyseur morphologique à états finis de Xerox (Beeseley, 2001)[351], utilisant les outils de Xerox de modélisation de langage à états finis, permettant d'obtenir pour chaque mot toutes ses listes de caractéristiques morphologiques possibles.
- Le système Sebawai (Darwish, 2002)[352] : basé sur une liste de paires du mot-racine arabe pour établir une liste des préfixes et suffixes, afin de construire des modèles de

[348] Normes de saisie et de dépouillement des textes politiques, p. 43
[349] L'écriture dans la représentation de la langue : la lettre et le mot en arabe
[350] Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî, p. 265
[351] Finate-State Morphological Analysis and Generation of Arabic at Xerox Research : Status and Plans in 2001
[352] Building a Shallow Arabic Morphological Analyzer in One Day

stemmes, permettant de calculer la probabilité d'apparition d'un préfixe, un suffixe, ou un modèle et de trouver la racine avec un taux de réussite de 84%.

- Le système AraParse (Ouersighni, 2002)[353], basé sur des ressources linguistiques, notamment sur une fonction d'analyse morpholexicale
- L'analyseur morphologique arabe basé-Web (Atwel et al., 2004), basé sur une méthode d'exploration contextuelle qui permet d'identifier le mot et ses caractéristiques contextuelles et donc de rechercher les affixes qui peuvent lui être associés.
- L'analyseur morphologique MORPH (Morphological analyser of non vowelled Arabic texts/ Analyseur morphologique de textes arabes non voyellés), basé sur une approche computationnelle, permettant de déterminer pour chaque mot sa base ou sa racine ainsi que la liste de toutes ses caractéristiques morphologiques possibles (Belguith et al., 2006)[354]

Pour notre travail, l'analyse morphologique est réalisée sur Xerox ; ce choix est influencé en grande partie par la facilité d'accès du logiciel, sur l'adresse www.xrce.xerox.com/research/mltt/arabicm, mais également, par ses listes de caractéristiques morphologiques possibles pour chaque mot (cf. Figure 27 : Extrait de l'analyse du mot « تشكّل = tašakkala = se former » à partir de l'étiqueteur de XE-ROX).

Comme nous l'avons dit précédemment (cf. 4.2.2.2 Étiquetage grammatical), l'arabe distingue trois catégories principales de mots : les verbes, les noms et les particules ; nous présentons les critères de la lemmatisation de chacune de ces catégories.

4.2.3.1 Lemmatisation du verbe

Le verbe, entité exprimant un sens dépendant du temps, présente une multitude de flexions, variant en personne, en temps, en mode et en voix ; ces différentes graphies sont regroupées sous une même racine, correspondant au lemme du verbe (Labbé ; 1990)[355].

Dans les dictionnaires arabes, l'entrée du verbe correspond à la troisième personne du singulier masculin de l'accompli actif ; nous considérons cette entrée, le lemme du verbe arabe. Il est entièrement vocalisé afin d'éviter toute ambiguïté, tant au niveau du verbe lui-même qu'au

[353] La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe : utilisation pour la détection et le diagnostic des fautes d'accord

[354] Analyse et désambiguïsation morphologiques de textes arabes non voyellés, p. 495

[355] Normes de saisie et de dépouillement des textes politiques, p. 51

niveau des catégories des mots, cette forme produisant une grande partie des homographies. Citons l'exemple « « شَرَّحَ = *šarraḥa* = filtrer ».

Pour cela, une analyse morphologique est nécessaire ; comme dans les autres langues, le verbe en arabe se décline en personne, en temps, en mode et en voix (Dichy et al., 1999)[356]. Voici les critères principalement utilisés pour désigner le verbe arabe :

- aspect (temps) : accompli (Acc.) ou inaccompli (Ina.)
- voix : active (Act.) ou passive (Pas.)

À l'accompli, les verbes se distinguent par des suffixes qui s'accordent en genre et en nombre ; quant à l'inaccompli, ils se distinguent par leurs préfixes qui s'accordent également en genre et en nombre. Il convient d'ajouter l'impératif, même si ce cas n'est pas répertorié dans notre corpus ; il sera utile dans des études ultérieures.

De plus, nous précisons la transitivité des verbes, en indiquant dans ce cas la particule de transitivité ; certains verbes admettent une double transitivité, comportant deux compléments avec la possibilité d'emploi d'une particule de transitivité qui sera indiquée (Abbes, 2004)[357].

Par ailleurs, la majorité des verbes arabes sont formés d'une racine triconsonantique et éventuellement tétraconsonantique, associée à différents schèmes, généralement définis comme un modèle décrivant un groupe de mots partageant certaines propriétés phonologiques, morphologiques, syntaxiques et sémantiques (Larcher, 2003)[358]. Il est alors possible de classer les verbes selon leurs radicaux, en renseignant le schème verbal associé, en indiquant s'il s'agit d'un verbe simple ou augmenté et s'il est triconsonantique ou tétraconsonantique (Mouelhi, 2008)[359]. Citons l'exemple du terme « جفف = jaffafa = sécher » :

Terme	Référence	Équivalent	Contexte	Remarque
جفف	Réf 1.b	Sécher	يجفف الإيثر فوق CaCl2 مدة 24 ساعة. Yujaffafu al'îṯir fawqa CaCl2 muddat 24 *sa'a*t. L'éther est séché sur CaCl2 pendant 24 heures.	UTS, verbe passif

Tableau 13 : Terme « جفف = jaffafa = sécher »

[356] Les verbes arabes, p. 11
[357] La Conception et la réalisation d'un concordancier électronique pour l'arabe, p. 130
[358] Le système verbal de l'arabe classique.
[359] Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî

Son analyse morphologique indique qu'il s'agit du lemme « جَفَّفَ = jaffafa = sécher », verbe triconsonantique augmenté, à la forme II, verbe à l'inaccompli et à la voix passive ; son analyse sur Xerox donne 8 solutions (cf. Figure 29 : Analyse morphologique sur Xerox du terme « جفف = jaffafa = sécher ») :

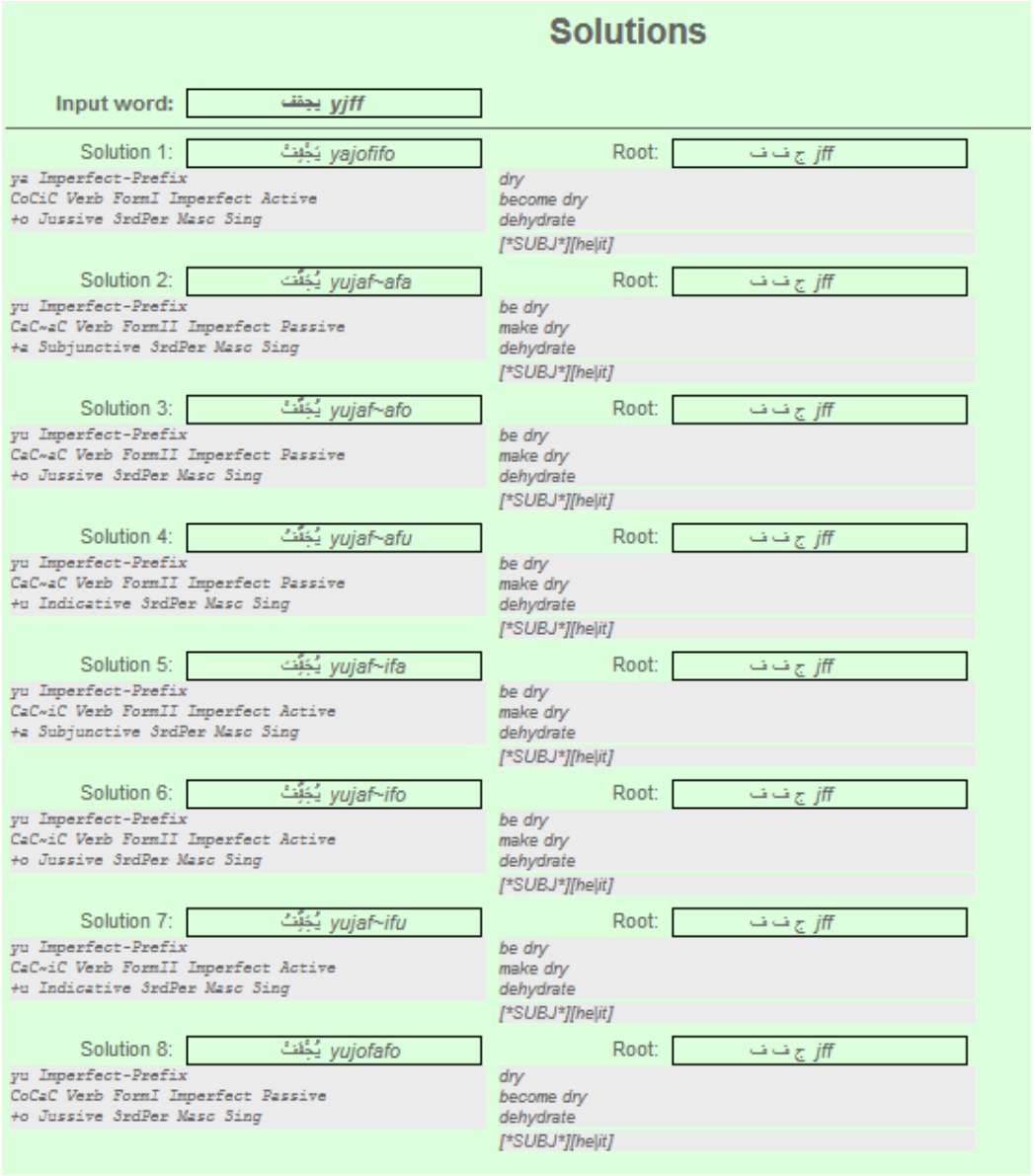


Figure 29 : Analyse morphologique sur Xerox du terme « جفف = jaffafa = sécher »

Notre analyse du terme « جفف » fait partie des propositions de l'analyseur sur Xerox (solution 4 de la figure 31) ; cependant, cela indique que certaines formes verbales demeurent ambiguës et requièrent notre intervention, puisque qu'a priori, aucune codification par défaut n'est admise car elle pourrait conduire à des erreurs.

4.2.3.2 Lemmatisation du nom

Le nom, élément exprimant un sens indépendant du temps, assume des fonctions diverses et variées comme celui qui fait l'action (agent), celui qui subit l'action (objet), celui de signifier l'instrument de l'action (instrument), celui de désigner, en général, un endroit (lieu), celui qui désigne l'action (nom d'action), etc.

Comme dans les autres langues, le nom en arabe ne fait pas exception et son lemme est le singulier, exception faite des pluriels qui n'ont pas de singulier et des pluriels dont le singulier n'a pas la même racine ; mais, nous indiquons égalemant le genre, le nombre et le cas (Kouloughli, 1994)[360] :

- Genre : masculin (Masc.) et féminin (Fem.)
- Nombre : singulier (Sing.), duel (Du.) et pluriel (Pl.)
- Cas : nominatif (Nom.), accusatif (Acc.) et génitif (Gen.)

D'autre part, le nom en arabe est divisé en deux familles : les noms qui ne peuvent pas être rattachés à une racine verbale, appelé noms primitifs, et les noms qui sont dérivés à partir d'une racine verbale (noms dérivés ou déverbaux) ; ces derniers sont généralement des racines triconsonantiques, associées à des schèmes verbaux et sont constituées principalement du « مصدر = *maṣd*ar = nom verbal ou nom d'action », de « اسم الفاعل = ism al-fâ ʿil = participe actif », de « اسم المفعول = ism al-mafʿûl = participe passif » et de « اسم المكان والزمان = ism al-makân wa al-zamân = nom de lieu et de temps ».

De ce fait, le système nominal de l'arabe admet différents systèmes de déclinaison, suivant la nature de la forme (simple, diptotes, etc.) et son nombre (singulier, duel ou pluriel). Notons que « le phénomène du pluriel irrégulier en arabe pose un défi à la morphologie, non seulement à cause de sa nature non concaténative, mais aussi parce que son analyse dépend fortement de la structure comme les verbes irréguliers. » (Kiraz, 1996)[361].

Citons l'exemple du terme « مركب = murakkab = composé » (cf. Tableau 14 : Analyse du terme « مركب = murakkab = composé ») :

[360] Grammaire de l'arabe d'aujourd'hui, p. 72
[361] Analysis of the Arabic Broken Plural and Diminutive

Terme	Référence	Équivalent	Contexte	Remarque
مركب	Réf. 1.e	Composé	تحضير المركب الكيرالي (+) (R5,R2,R1)-2هيدروكسي بينان3- أون Préparation du composé chiral (1R,2R,5R)-(+)-2-Hydroxy-3-pinanone	UTS

Tableau 14 : Analyse du terme « مركب = murakkab = composé »

Son analyse morphologique indique qu'il s'agit du lemme « مُرَكَّبِ= murakkab = composé », de la racine triconsonantique « ر ك ب », d'un nom singulier déterminé et au génitif. Son analyse sur Xerox donne 9 solutions (cf. Figure 30 : Analyse morphologique sur Xerox du terme « مركب = murakkab = composé ») :

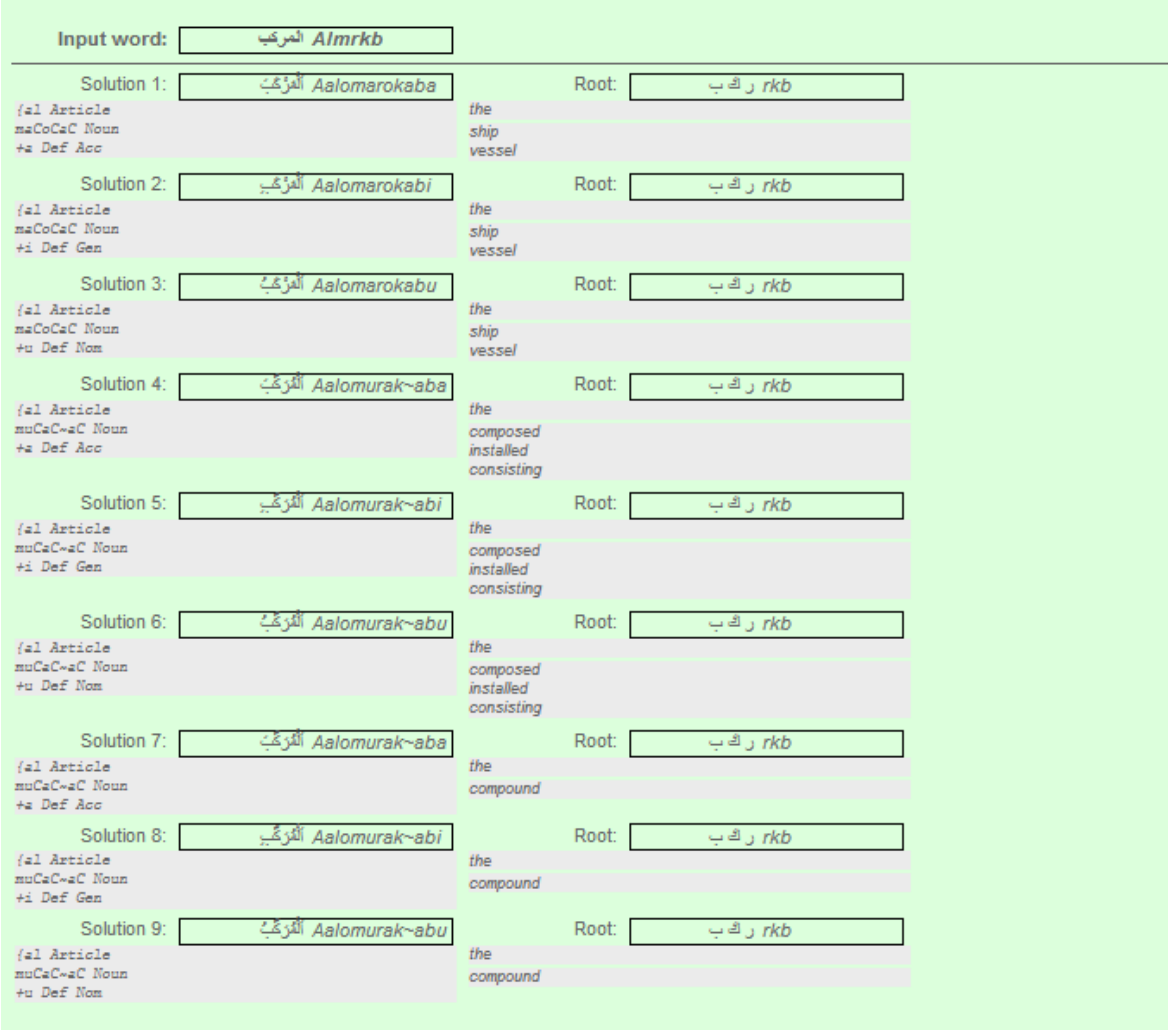


Figure 30 : Analyse morphologique sur Xerox du terme « مركب = murakkab = composé »

Notre analyse du terme « مركب = murakkab = composé » fait partie des propositions de l'analyseur sur Xerox (solution 5 de la figure 32). Cependant, comme pour les verbes, cela indique que certaines formes nominales demeurent ambiguës et requièrent notre intervention,

puisque qu'a priori, aucune codification par défaut n'est admise car elle pourrait conduire à des erreurs.

Concernant les unités terminologiaques complexes, nous proposons de les lemmatiser en fonction de la base du terme (Petit, 2003)[362] ; par exemple, le terme « كروماتوغرافيا العمود = krûmâtûġâfiyâ al'amûd = chromatographie sur colonne » a pour base « كروماتوغرافيا = krûmâtûġâfiyâ = chromatographie » (cf. 4.2.3.4 Lemmatisation de l'emprunt).

4.2.3.3 Lemmatisation de l'adjectif

Quant aux adjectifs, pouvant être aussi des termes à l'état libre (Lelubre, 2005)[363], les grammairiens arabes les ont assimilés aux noms, puisque « du point de vue de la forme, *l'adjectif présente tous les caractères généraux du nom* : en particulier, il est marqué en genre, en nombre, en cas et en détermination » (Kouloughli, 1994)[364]. Il s'agit alors de la même lemmatisation que celle du nom.

4.2.3.4 Lemmatisation de l'emprunt

Les emprunts ne sont pas reconnus par l'analyseur Xerox ; il n'y a pas de renseignements associés (cf. Figure 31 : Analyse morphologique sur Xerox du terme « كيرالي = kîrâlî = chiral »).

Voici l'exemple du terme « كيرالي = kîrâlî = chiral », issu de l'exemple précédent (cf. Tableau 14 : Analyse du terme « مركب = murakkab = composé ») :

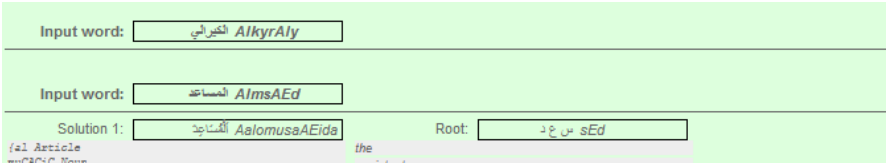


Figure 31 : Analyse morphologique sur Xerox du terme « كيرالي = kîrâlî = chiral »

Par conséquent, les emprunts sont indexés en indiquant le genre, le nombre et le cas.

4.2.3.5 Lemmatisation de la particule

Les particules, entités servant à situer les événements et les objets par rapport au temps, sont généralement les mots outils pour une langue donnée. Dans la division tripartite opérée

[362] Lemmatisation et figement lexical
[363] Le statut de l'adjectif en langue de spécialité
[364] Grammaire de l'arabe d'aujourd'hui, p. 101

par la Tradition grammaticale arabe, la particule est définie par opposition au nom et au verbe, impliquant que tout ce qui n'est ni nom ni verbe est particule. Cela dit, toute particule est mot-outil mais tout mot-outil n'est pas particule. Le terme 'particule' est proposé par la tradition orientaliste pour traduire « حرف = *ḥ*arf » (Dichy, 1990)[365].

De ce fait, les particules forment un ensemble de mots (article défini, prépositions, coordonnants, pronoms…) et sont classées selon leur fonction dans la phrase, mais également selon leur place dans le mot (proclitiques, préfixes, pré-bases …), puisque certaines particules peuvent également porter des préfixes et suffixes, compliquant leur identification ; par exemples, le coordonnant « و = wa = et » ou la préposition : « ب = bi= par ».

À l'écrit, il est parfois difficile de faire la différence entre un proclitique et un caractère appartenant à la racine de certains mots.

D'autre part, nous distinguons deux catégories pour la flexion des particules : les mots outils non déclinables ou invariables comme « على = ʿalâ = sur » (leur lemme est le mot-outil non déclinable) et les mots outils déclinables ou variables, suivant le système de déclinaison à trois cas selon leurs fonctions dans la phrase, comme « ذو = *ḏû* = qui a », leur lemme est le mot-outil au nominatif.

À présent, nous pouvons dire que nous avons identifié les termes du domaine de la chimie ; chacun de ces termes est indexé en indiquant sa forme (terme simple ou terme complexe), sa partie du discours (nom ou verbe) et son lemme (verbe à la troisième personne du singulier masculin de l'accompli actif ou nom au singulier).

Il semblerait que nous possédons tous les éléments nécessaires pour constituer les fiches terminologiques de chaque terme du domaine de la chimie afin de construire une grammaire d'identification de ces termes. Cependant, il manque un élément essentiel : l'attribution d'une classe pour chaque d'un terme, définissant sa place dans le domaine de la chimie.

Dans le chapitre suivant, nous présentons la classification du domaine de chimie.

[365] L'écriture dans la représentation de la langue : la lettre et le mot en arabe

Chapitre 5 : Classification du domaine de la chimie

Le point de départ du travail terminologique est la détermination de la limite du domaine étudié, c'est-à-dire, ce qui fait partie du domaine de la chimie et ce qui n'en fait pas partie. Pour cela, nous allons prendre en considération l'historique du développement de la chimie, axe nous permettant de concevoir l'organisation du domaine. Nous consultons également différents types de classification et nous aboutissons à une classification, qui permet d'organiser tous ces termes de la chimie en fonction de leur usage dans les textes afin de modéliser leur comportement syntaxique et sémantique par des règles de grammaire.

5.1 Domaine

La détermination d'un domaine permet de préciser le secteur ou la sphère dans lequel seront identifiés les termes, les concepts et leurs relations. De Bessé nous éclaircit sur cette notion en définissant le domaine :

> *« Une structuration des connaissances [*permettant*] d'identifier, de délimiter, de dénommer [...] un système conceptuel » (De Bessé, 2000)*[366].

Le système conceptuel est constitué du terme, du concept, de la définition et du domaine. Sans l'attribution d'un domaine, ces éléments ne peuvent exister et/ou avoir la signification souhaitée. Étant un constituant du concept, le domaine permet d'obtenir des informations sur le concept et d'indiquer le système conceptuel auquel il appartient ; il forme alors un couple indivisible avec la définition permettant d'identifier les concepts. Chaque concept est défini en fonction du domaine étudié ; il est alors considéré comme un « ensemble organisé de concepts » (De Bessé, 2000)[367].

Par exemple, le terme « ماء = *mâ'* = eau » dans le domaine de la chimie, est défini comme étant une molécule composée d'un atome d'oxygène et de deux atomes d'hydrogène, sa masse est de 18 g/mol et elle peut être sous forme liquide, solide ou gaz. Par comparaison, dans le domaine de la cuisine, l'eau est définie comme un ingrédient liquide, incolore et inodore.

[366] Le domaine, p. 183
[367] Idem

5.1.1 Limite du domaine

Le domaine peut et doit être délimité afin de répondre aux objectifs demandés. En effet, la délimitation d'un domaine dépend de la vision des connaissances, des besoins des utilisateurs et des consommateurs :

> *« Il existe plusieurs façons de procéder au découpage des connaissances et des activités, qui correspondent à plusieurs points de vue. Les domaines n'ont pas d'existence par eux-mêmes. Ils sont délimités du point de vue du chercheur, de l'ingénieur, du technicien, de l'amateur ». (De Bessé, 2000)*[368]

Au terme « ماء = *mâ'* = eau », nous pouvons ajouter comme renseignement à sa définition qu'elle possède une structure tétraédrique, qu'elle est utilisée comme solvant, qu'elle est un composé amphotère, c'est-à-dire, qu'elle peut être une base ou un acide…

Pour un travail terminologique, il est impératif de limiter le domaine étudié afin d'identifier les termes qui lui appartiennent et ceux qui sont extérieurs au domaine :

> *« Il convient ainsi de délimiter et de caractériser ce domaine. Un domaine de spécialité est contigu à d'autres domaines, et est souvent à l'intersection d'autres domaines, avec lesquels il a des éléments communs). » (Lelubre, 1992)*[369]

Notre domaine de spécialité étudié est la chimie ; il est contigu à certains domaines et est à l'intersection d'autres, avec lesquels il a des éléments communs, comme la physique, ou encore la biologie, par exemple pour la synthèse d'un corps vivant. Tout ceci ne facilite pas notre travail de délimitation.

Rappelons qu'un terme ne peut être défini qu'en fonction de son domaine de spécialité :

> *« Toute étude terminologique doit prendre en compte le domaine de spécialité dont relèvent les termes étudiés : les unités référentielles (concepts) qui leur correspondent sont définies dans le cadre de ce domaine de spécialité ». (Lelubre, 1992)*[370]

5.1.2 Domaine et sous-domaine

Le domaine peut contenir plusieurs sous-domaines qui peuvent eux-mêmes être décomposés :

[368] Le domaine, p. 187
[369] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 93
[370] Idem

« Le domaine lui-même comprend plusieurs sous-domaines, qui peuvent à leur tour se subdiviser en sous-domaines. [...] il s'agit donc pour le terminologue de structurer le domaine relatif à la terminologie étudiée ». (Lelubre, 1992)[371]

Concernant notre domaine de la chimie, nous pouvons avoir comme sous-domaines la chimie organique, la chimie minérale, la chimie analytique … En effet, les sous-domaines déterminés doivent permettre de classer tous les termes du domaine. Par conséquent, le choix des sous-domaines est déterminant et essentiel pour la construction de la classification du domaine. De plus, ce dernier peut avoir plusieurs constructions possibles ; en effet, cela dépend de deux caractéristiques : son auteur (qui a construit cette classification du domaine ?) et de son objectif (pourquoi a-t-on construit la classification du domaine ?).

La première caractéristique (l'auteur) est due au fait que chaque personne à sa propre représentation du monde et la décrit à sa manière en utilisant des mots et des formes qui lui sont personnels. La seconde caractéristique (l'objectif) guide l'auteur dans sa construction de la classification du domaine et le limite dans le choix des sous-domaines.

Par exemple : « أزوت= ʾazût = azote » est un gaz à température ambiante et à pression atmosphérique ; mais, à -195,79°C, il est liquide.

- Si notre objectif est la classification des éléments chimiques, les sous-domaines choisis sont : solide, liquide, gaz.
  - Soit nous effectuons la classification des éléments chimiques à température ambiante et à pression atmosphérique et nous ne citons l'azote que dans le sous-domaine gaz (cf. Figure 32 : Classification des éléments chimiques en fonction de leur état à température et à pression ambiante) :

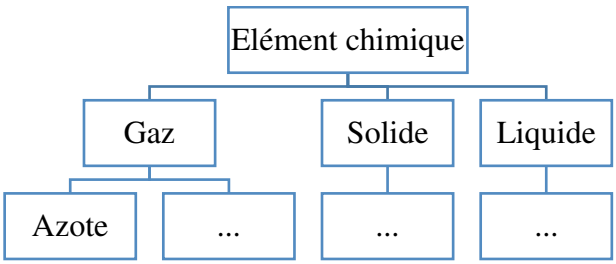


Figure 32 : Classification des éléments chimiques en fonction de leur état à température et à pression ambiante

  - Soit nous effectuons la classification des éléments chimiques quelles que soient la température et la pression et nous citons l'azote dans les sous-domaines gaz et

[371] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation

liquide (cf. Figure 33 : Classification des éléments chimiques en fonction de leur état) :

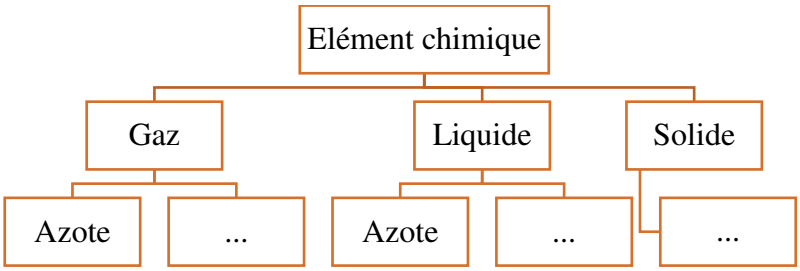


Figure 33 : Classification des éléments chimiques en fonction de leur état

- Si notre objectif est la classification des gaz, nous limitons notre étude aux gaz (cf. Figure 34 : Classification des gaz) :

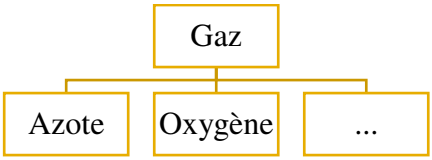


Figure 34 : Classification des gaz

En ce qui nous concerne, notre travail de fouille de textes a pour but l'extraction d'information ciblant les termes de la chimie en arabe. Par conséquent, notre classification doit couvrir tout le domaine de la chimie.

Cette classification obtenue pour un domaine est appelée 'arbre du domaine' :

> *« Puis à l'intérieur de chaque sous-domaine, il convient d'établir les relations existantes entre les unités référentielles relatives à ce sous-domaine (en tenant compte du fait qu'éventuellement une unité référentielle peut relever de deux ou plusieurs sous-domaines ...). C'est ce qu'on appelle l'arbre du domaine ». (Lelubre, 1992)*[372]

Cet arbre de domaine est représenté par un schéma arborescent prenant en compte les relations entre toutes les unités référentielles du domaine.

La construction de l'arbre de domaine présente des difficultés quant à la classification d'une même unité référentielle qui peut se retrouver dans plusieurs catégories :

> *« L'établissement d'un arbre du domaine n'est pas une tâche facile [..] car il s'agit d'établir un système de classement et qu'en cette matière, un même objet peut être envisagé sous une grande diversité d'angles. […] Par ailleurs, certains éléments pourront trouver place dans plus d'un embranchement ;*

[372] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation

le terminologue devra alors décider s'il doit les répéter ou ne les greffer qu'à un seul nœud […] ». (Rondeau, 1983)[373]

Par exemple, le terme « أكسجين = *'uksi*jîn = oxygène » présente plusieurs caractéristiques ; il s'agit d'un élément chimique, appartenant au groupe des chalcogènes qui fait partie des non-métaux et est un gaz. Par conséquent, si notre arbre de domaine possède les sous-domaines 'élément chimique', 'chalcogènes', non-métaux' et 'gaz', alors ce terme sera répété dans chacune de ces catégories. Or, à chaque terme doit être attribuée une et uniquement une seule étiquette dans la classification et c'est par l'intermédiaire de relations que les termes sont spécifiés et caractérisés.

5.2 Chimie

La chimie est définie comme la science qui étudie la matière, ses propriétés et les transformations qui peuvent s'y produire. Cette science est partout avec nous où que nous soyons et quoique nous fassions :

« Plus que toute autre science, la chimie tient une place considérable dans notre vie de tous les jours. Toutes les réactions qui participent au maintien de la vie mettent en jeu des composés chimiques. Quotidiennement, on se trouve en contact avec des substances telles que l'essence, l'huile, le bois, le papier, des vêtements, des médicaments, des parfums et des odeurs. Tous les jours, dans les journaux, à la radio ou à la télévision, il est fait mention de polyéthylène, de résine époxy, de nicotine, de saccharine, de corps gras insaturés, de cholestérol, d'indice d'octane…bref, la chimie n'est plus une branche de la science réservée aux seuls chimistes, médecin, dentiste, pharmacien, infirmière, ou agriculteur. Elle fait partie de notre culture technologique ». (Friedli, 2002)[374]

En d'autres termes, c'est l'air que nous respirons ; ce fluide gazeux parfait, incolore, inodore et invisible est formé principalement d'oxygène et d'azote. C'est également la nourriture que nous mangeons ; les fruits et les légumes ou encore les plats cuisinés sont obtenus par l'utilisation de certains produits chimiques. C'est aussi les habits que nous portons ; ces matières douces, chaudes, élastiques et colorées sont constituées de composés chimiques. Et c'est les appareils et les outils que nous utilisons ; la voiture ou le téléphone ou encore le réfrigérateur sont composés d'au moins un élément chimique.

[373] Introduction à la terminologie, p. 70
[374] Chimie générale pour ingénieur, p. IX

5.2.1 Historique de la chimie

L'étude de la chimie par son historique et son évolution nous permet de voir comment s'organise ce domaine autour de ses concepts fondamentaux qui évoluent avec le temps, ainsi que leur importance dans l'organisation du domaine :

> *« L'élément historique est l'un des éléments importants à prendre en considération dans l'établissement d'une classification dans un domaine de la connaissance humaine, c'est-à-dire, la façon dont ce domaine s'est constitué et a évolué avec le développement général de la société, en particulier sur le plan scientifique et technique ». (Lelubre, 1992)*[375]

Pour comprendre l'histoire de la chimie, nous ne contentons pas de regarder les récentes découvertes des chimistes. Nous traversons les siècles pour voir les débuts de la chimie, notamment l'apport des arabes pour ce domaine.

Contrairement aux autres sciences comme les mathématiques ou l'astrologie, la chimie ne devient une « conception purement rationnelle » que seulement depuis le XVIII^{ème} siècle grâce aux chimistes Lavoisier, Berthollet, Guyton-Morveau et de Fourcroy ; mais, en attendant son apogée, elle sort de sa « forme religieuse et mystique » et apparait comme une science mixte, l'alchimie :

> *« L'évolution qui s'est faite à cet égard, depuis les Orientaux jusqu'aux Grecs et jusqu'à nous, n'a pas été uniforme et parallèle dans tous les ordres. Si la science pure s'est dégagée bien vite dans les mathématiques, son règne a été plus retardé dans l'astronomie, où l'astrologie a subsisté parallèlement jusqu'aux temps modernes. Le progrès a été surtout plus lent en chimie, où l'alchimie, science mixte, a conservé ses espérances merveilleuses jusqu'à la fin du siècle dernier. » (Berthelot, 1885)*[376].

Située vers le III^{ème} siècle, l'alchimie tient son origine de la transmutation des métaux précieux née des pratiques des orfèvres égyptiens pour les imiter et les falsifier (Berthelot, 1889)[377]. Cet ensemble de procédés industriels très perfectionnés existait en Babylonie comme en Egypte et était commun d'ailleurs aux Phéniciens et aux populations syriennes, intermédiaires entre l'Egypte et la Babylonie. Ils se sont conservés par voie traditionnelle jusqu'aux Arabes et aux Persans modernes. Il s'agit là d'une autorité différente reprenant les auteurs alchimistes et proposant une science plus méthodique, plus avancée et par conséquent postérieure à celle des alchimistes. L'alchimie arabe permet d'être sur un terrain plus solide et d'atteindre le sommet

[375] La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, p. 81
[376] Les origines de l'alchimie, p. VII
[377] Introduction à l'étude de la chimie des anciens et du moyen âge, p. VI

de la science chimique notamment avec Jabir b. Hayyan et Abu Bakr al-Razi (Anawati, 1997)[378]. Enfin, ce sont les Arabes de Syrie et d'Espagne qui l'ont enseignée à l'Occident, et non par seulement par la traduction des écrits grecs (Halleux, 1997)[379], puisque « l'alchimie de l'Occident latin ne doit à peu près rien aux Grecs ; aux Arabes, elle doit à peu près tout. » (Ruska, 1931)[380].

Différentes théories ont vu le jour durant cette période : de la doctrine des quatre éléments (terre, eau, air et feu), définissant la matière et ses transformations à la théorie atomique, en passant par la théorie du mercure des philosophes. Mais, avec les nouvelles découvertes relatives aux matières explosives et à l'électricité, permettant d'atteindre de nouvelles températures et de communiquer à la matière en mouvement une activité et une force vive, un progrès immense et inattendu a été accompli en chimie, mettant un terme au rêve antique de la transmutation.

Le changment radical en chimie est alors la mise au point d'une terminologie avec Lavoisier, en définissant les substances chimiques exclusivement par leur composition.

À partir de là, la chimie se constitue comme une science positive et définit l'existence définitive et immuable de soixante-six éléments distincts, des éléments isomères et des polymères ; leur classification en familles naturelles et en séries périodiques, donne naissance au tableau périodique de Mendeleïev.

S'ensuit le développement de la chimie organique avec la synthèse de l'urée par Friedrich Wöhler en 1818 et la synthèse de la mauvéine par Henry Perkin en 1856, créant une véritable révolution scientifique et technologique qui s'est alors développée.

La chimie était essentiellement une science de l'analyse cherchant à découvrir la composition de la matière. Mais avec la nomination de Berthelot au Collège de France, la chimie marque une très nette rupture avec le passé et une nouvelle voie s'ouvre, favorisant la synthèse plutôt que l'analyse, laissant apparaitre de nouvelles disciplines, comme la thermochimie.

D'autres découvertes telles que la théorie atomique de Dalton, la loi des volumes de Gay-Lussac, la théorie électrochimique de Davy, celle des hydracides et celle des radicaux ou encore celle du noyau… (Ladenburg, 1909)[381] ont vu le jour. En raison du nombre croissant de ces découvertes dans la chimie et de ces idées fondamentales, une nécessité d'élaborer les règles

[378] L'alchimie arabe, p. 122
[379] La réception de l'alchimie arabe en occident, p. 154
[380] Turba Philosophorum : Ein Beitrag zur Geschichte der Alchemie, cité par Halleux, idem, p. 153
[381] Histoire du développement de la chimie depuis Lavoisier jusqu'à nos jours

de formation de la nomenclature qui seraient valables au niveau international a réuni les chimistes et la constitution de l'Union Internationale de Chimie Pure et Appliquée (UICPA) (cf. 1.4.2.1.2 Union Internationale de Chimie Pure et Appliquée (UICPA)) et la Commission de Nomenclature de la Chimie Organique (Leigh et al., 2001)[382]. Ces organismes internationaux systématisent les méthodes de dénomination des composés et également des données basiques de la chimie.

5.2.2 Classification de la chimie

La classification du domaine étudié est une étape primordiale dans un travail de fouille de textes qui consiste à organiser les termes du domaine afin d'en faciliter l'accès et l'étude :

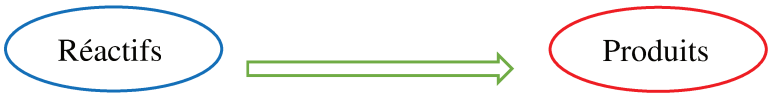
> *« Distribution d'un ensemble de connaissances, d'entités ou d'objets en vue d'en faciliter l'accès ou l'étude et ce, au moyen de critères alphabétiques, associatifs, hiérarchiques, numériques, idéologiques, spatiaux, chronologiques, etc. » (Pavel et al., 2001)[383].*

Cette classification de la chimie requiert une étude minutieuse afin de pouvoir comprendre les principes du domaine purement scientifique. Selon Wüster, ce travail ne peut être effectué que par des spécialistes du domaine, étant les seuls à posséder les connaissances nécessaires pour réaliser ces classifications.

Depuis son origine, les chimistes ont cherché à organiser le domaine de la chimie en plusieurs classes et/ou catégories. Ils se réfèrent à la matière des corps pour classer les éléments, en adoptant des notations et des nomenclatures particulières, construites d'après des méthodes précises, permettant d'établir des listes des métaux, des alliages et des techniques utilisées.

La chimie peut être organisée selon trois axes principaux (cf. Figure 35 : Représentation générale d'une réaction chimique) :

- Réactif : un réactif est un composé chimique mis en présence avec d'autres composés et se transforme en produit
- Produit : un produit est un composé chimique, récupéré à la fin de la réaction, résultant de l'interaction des réactifs
- Réaction : une réaction est la transformation des réactifs en produits



[382] Principes de nomenclature de la chimie. Introduction aux recommandations de l'IUPAC, p. 1
[383] Précis de terminologie, p. 116

Qui se transforment

Figure 35 : Représentation générale d'une réaction chimique

Ces trois axes sont liés entre eux ; réactifs, produits et réactions sont la base de la chimie. Cependant, établir une classification du domaine à partir de ce schéma ne permet pas de répertorier tous les termes de la chimie. En effet, ces trois axes (réactif, produit et réaction) restreignent le domaine de la chimie et ne prennent pas en compte son environnement (Propriétés, conditions, préparation…). Par conséquent, une organisation du domaine selon les connaissances et les travaux des chimistes est plus appropriée.

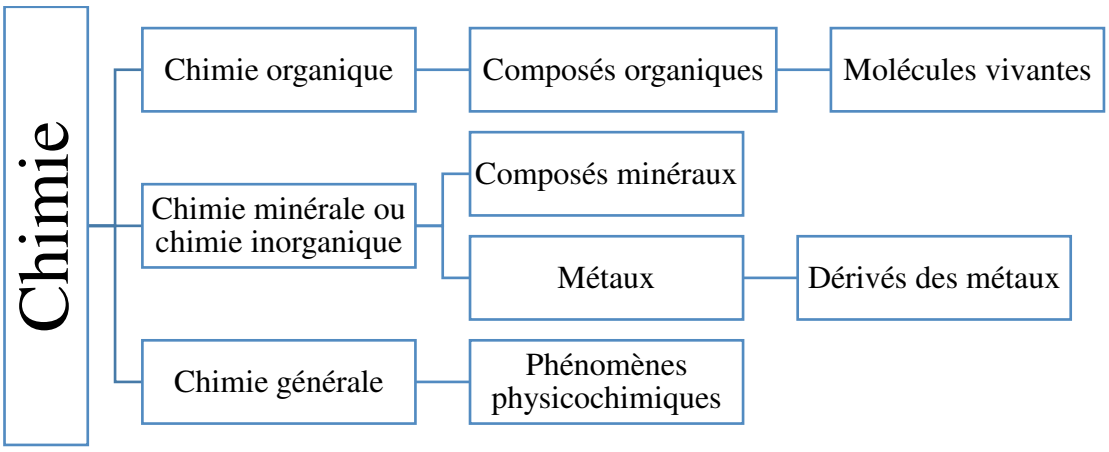
Une première organisation du domaine est possible en prenant en compte les bases de la chimie ainsi que les domaines partagés avec la chimie :

> *« La chimie a longtemps été scindée en chimie générale, chimie minérale et chimie organique ». (Gordon et al., 2012)*[384]

Elle est divisée en trois sous-domaines :

- La chimie organique : étude des composés organiques, incluant les molécules vivantes de la biologie, fondée en 1879
- La chimie minérale ou chimie inorganique : par opposition à la chimie organique, étude des composés minéraux, des métaux et leurs dérivés, fondée en 1842
- La chimie générale : étude des phénomènes chimiques ou physicochimiques faisant appel à la physique et aux mathématiques, fondée en 1809

Voici la classification obtenue, en prenant en compte les trois sous-domaines (cf. Figure 36 : Classification de la chimie avec trois sous-domaines) :



[384] Le développement de la chimie au XXe siècle

Figure 36 : Classification de la chimie avec trois sous-domaines

Cependant, avec le développement et l'avancée de la recherche mais aussi la découverte de nouvelles méthodes de synthèse, d'analyse…, la classification de la chimie s'est vue évoluer et modifier. Sa classification est beaucoup plus riche et est composée de nombreux sous-domaines qui prennent en compte les évolutions du domaine ainsi que les domaines frontières de la chimie ; ces sous-domaines peuvent se subdiviser :

- La chimie générale : c'est l'étude des phénomènes chimiques, tels que la thermodynamique, l'atomistique et les équilibres en solution.
- La chimie organique et bio-organique : c'est la chimie du carbone, de la liaison en générale carbone-carbone avec souvent des hydrogènes et quelques exceptions, comme le graphite. Elle permet de synthétiser des édifices carbonés stables, du plus simple, le méthane CH_4, jusqu'au plus complexe que l'on obtient par synthèse totale ou que l'on trouve dans les organismes biologiques, comme les acides aminés.
- La chimie inorganique et bio-inorganique : anciennement chimie minérale (chimie des minéraux), c'est la chimie de tous les éléments de la classification périodique des éléments. Elle synthétise, elle étudie, elle utilise les objets chimiques les plus divers (moléculaires et solides), leur structure, leur réactivité ... Elle comprend comme sous-domaine la métallurgie, la chimie des solides et bien d'autres encore.
- La chimie physique : c'est l'étude des phénomènes physicochimiques ; elle peut se diviser à son tour en plusieurs grands ensembles : la cinétique, la catalyse, l'électrochimie. Elle fournit aux autres disciplines de la chimie un grand nombre d'outils théoriques et instrumentaux. Son nom dérive de sa grande proximité avec les concepts et les moyens de la physique.
- La chimie analytique : ce sont les méthodes et techniques d'études structurales et spectroscopiques.
- La chimie nucléaire : la chimie nucléaire a un statut particulier en ce qu'elle traite des phénomènes relatifs au noyau de l'atome (radioactivité, radiochimie, fusion et fission nucléaires...) alors que les autres branches de la chimie reposent essentiellement sur les propriétés des électrons. La chimie nucléaire est particulièrement importante dans le domaine de la santé, comme la radiochimie la curiethérapie (un domaine entièrement créé par Marie Curie en s'appuyant sur ses travaux) ...

- Nouvelles catégories : la chimie de l'atmosphère (détection et analyse des molécules et des particules dans l'atmosphère) ou encore la chimie verte (une approche de la chimie qui doit tendre vers une pratique plus soucieuse de l'environnement), … sont le fruit de l'évolution de la discipline et de ses rapports avec les sciences voisines et les technologies. L'accent mis sur les objets étudiés (molécules, assemblages de molécules, solides - moléculaires, ioniques, métalliques ...-) conduit à une structuration en chimie moléculaire, supramoléculaire et du solide où cohabitent expérience et théorie.

Nous proposons la classification suivante, qui prend en compte l'évolution du domaine (cf. Figure 37 : Classification de la chimie prenant en compte l'évolution du domaine) :

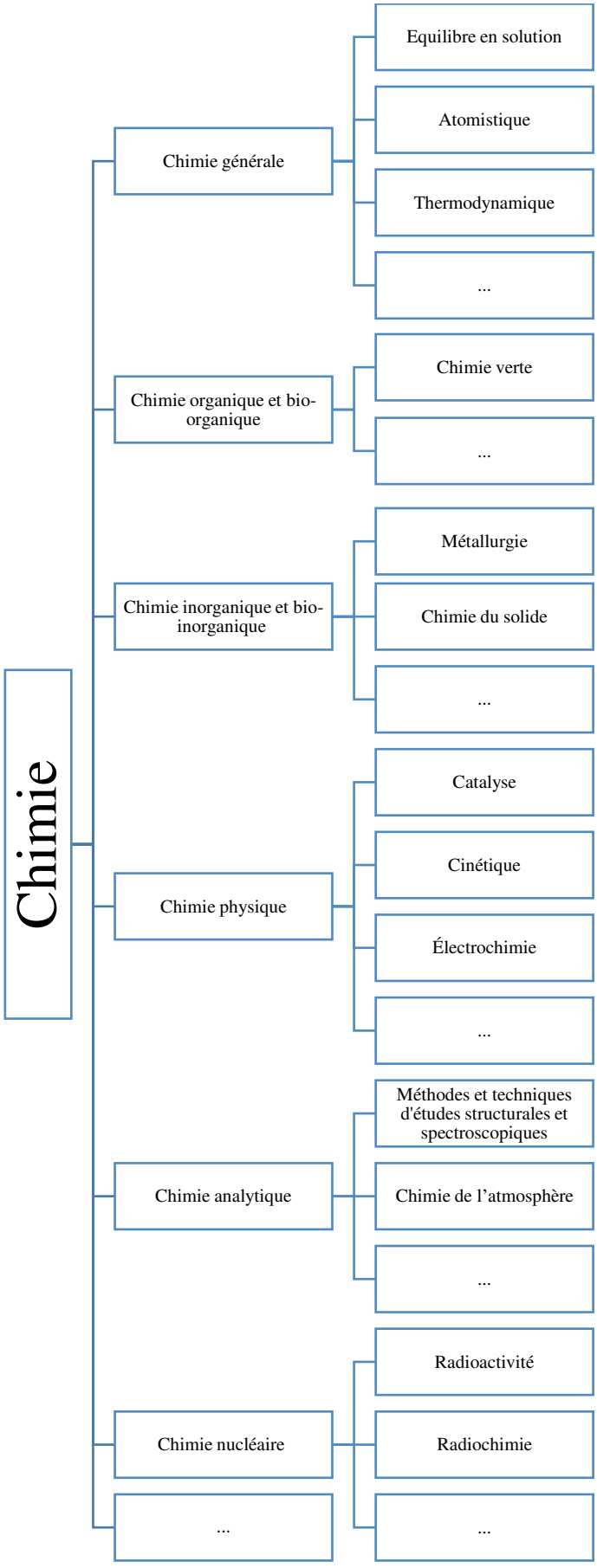


Figure 37 : Classification de la chimie *prenant en compte l'évolution du domaine*

Cette classification permet de différencier les branches de la chimie, ses expériences ainsi que ses outils.

D'une part, nous vérifions si cette classification est en accord avec les besoins actuels du domaine en la comparant avec des programmes d'enseignement secondaire et supérieur en chimie et/ou avec des sommaires des livres de chimie autant en français qu'en arabe. Ces programmes et ces livres de chimie abordent les principes fondamentaux de la chimie ; mais, en fonction de l'année d'enseignement et de la spécialisation de la discipline, les thèmes sont plus ou moins développés (cf. Annexe 4 : Programme de chimie).

Ces ouvrages traitent de la constitution de la matière en examinant la structure de l'atome, unité de base de tout élément présent dans la matière, et les différents modèles qui lui sont associés, ce qui a mené à la création du tableau périodique. Ce tableau et ses regroupements (groupes et périodes) sont utilisés dans la description de la structure et des propriétés de chaque élément. Du point de vue microscopique, la matière est traitée comme un rassemblement d'atomes et de molécules et du point de vue macroscopique, elle détermine leurs propriétés générales. Des transformations chimiques se produisent dans cette matière ; ce sont des réactions chimiques qui sont représentées par des équations chimiques. Elles doivent être interprétées et équilibrées à l'aide de calculs stœchiométriques mettant en jeu la relation quantitative lors de réactions chimiques, le calcul du pourcentage massique et la détermination de formules moléculaires à partir de données expérimentales. Ces données expérimentales fournissent la structure et la forme de molécules simples et complexes ainsi que celles des ions.

D'autre part, nous confrontons la classification obtenue avec les applications de la fouille de textes réalisées en chimie telles que les thésaurus, les ontologies... Les travaux effectués dans ce domaine répondent à des problématiques précises telles que la détermination des domaines intervenant dans la chimie, les relations entre les réactions chimiques et les éléments de ces réactions, ou encore la composition de molécules (cf. Figure 38 : Exemple d'ontologie en chimie : composition de molécules (Gandon, 2008))

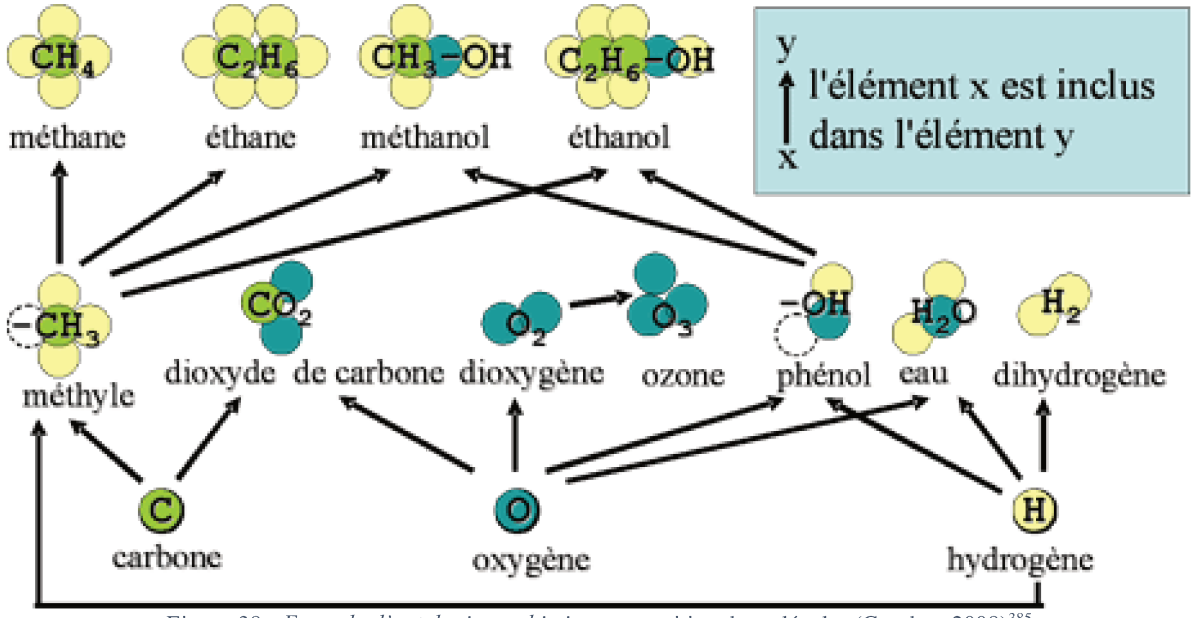


Figure 38 : *Exemple d'ontologie en chimie* : composition de molécules (Gandon, 2008)[385]

En d'autres termes, notre classification doit prendre en compte autant les principes fondamentaux de la chimie que les (nouveaux) besoins de ce domaine.

5.3 Classification adoptée

Cette classification est constituée de plusieurs catégories qui peuvent elles-mêmes se subdiviser en sous-catégories ; leur détermination et leur hiérarchisation doivent permettre de classer tous les termes du domaine, notamment ceux identifiés dans notre corpus.

5.3.1 Classification des entités chimiques

Dans notre corpus, nous identifions des termes tels que 'acide', 'alcool', 'titane', 'sodium', … Ces termes peuvent être classés en deux catégories : d'une part, les éléments chimiques du tableau périodique de la chimie qui regroupe plus d'une centaine de termes, et d'autre part, les composés chimiques qui sont la combinaison de deux ou plusieurs éléments chimiques (Albeiriss, 2016)[386].

5.3.1.1 Élément chimique

Les éléments chimiques peuvent être classés :

- soit par ordre alphabétique

[385] Graphes RDF et leur Manipulation pour la Gestion de Connaissances
[386] Étude Terminologique de la Nomenclature de la Chimie en Arabe dans une Approche de Fouille de Textes

- soit par le nombre d'occurrence en se référant à la lexicologie
- soit par ordre chronologique de découverte de l'élément chimique
- soit en fonction de leurs propriétés chimiques
- soit comme présentés dans le tableau périodique

Pour des raisons didactiques, nous utiliserons la classification du tableau périodique. Ce tableau périodique des éléments, ou tableau de Mendeleïev, a été créé en 1869 par le Russe Dmitri Mendeleïev. Ce tableau regroupe tous les éléments chimiques connus, classés en fonction de leur nombre de protons, ou numéro atomique. Le tableau est nommé tableau périodique, car nous retrouvons les éléments avec les mêmes propriétés à des intervalles réguliers (donc en colonne). Les éléments présents dans une même colonne ont des propriétés chimiques semblables. On distingue également plusieurs familles : les métaux vrais (les métaux alcalins et les métaux alcalino-terreux), les métaux de transition, les métalloïdes, les non-métaux, les halogènes, les gaz nobles, les lanthanides, les actinides et les transuraniens. Ce tableau périodique est un référentiel universel auquel peuvent être rapportés tous les types de comportements physique et chimique des éléments. En novembre 2016, sa forme standard comportait 118 éléments, allant de l'hydrogène $_{1}$H à l'oganesson $_{118}$Og.

Pour plus de précisions, nous ajouterons l'état des éléménts chimiques, c'est-à-dire, liquide, solide ou gaz, information capitale pour un chimiste, puisqu'un élément chimique a au moins un état et peut avoir jusqu'à 3 états (solide, liquide et gaz).

Notons que nous considérons ici un élément chimique comme la molécule chimique contenant un seul type d'élément (cf. 2.2.2.2.22.2.2.2.2 Métonymie).

Voici par exemple le terme « أزوت = ’azût = azote » : il appartient à la sous-catégorie ‘Non-métaux’ de la catégorie ‘Élément chimique’ (cf. Figure 39 : Classe du terme « أزوت = ’azût = azote ».

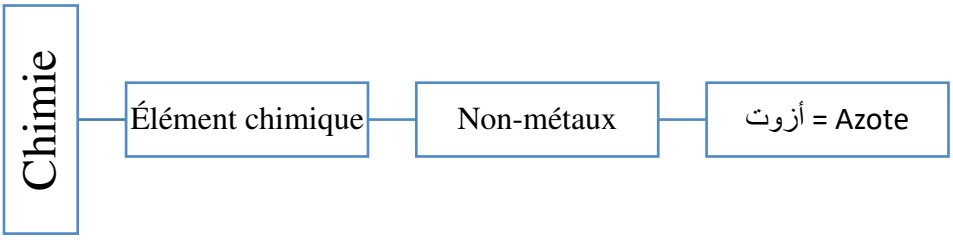


Figure 39 : Classe du terme « أزوت = ’azût = azote »

5.3.1.2 Composé chimique

Quant aux composés chimiques, ils sont classés selon leur nomination comme décrit par le système UICPA (cf. 1.4.2.1.21.4.2.1.2 Union internationale de chimie pure et appliquée (UICPA)). Ainsi, trois classes sont définies : les hydrocarbures, les fonctions chimiques et les espèces chimiques.

- Les hydrocarbures (HC)

Les hydrocarbures ne sont formés que de carbone et d'hydrogène. Ils peuvent être soit acycliques, soit aromatiques et des ramifications peuvent être ajoutées.

Par exemple le terme « ن-هكسان = n-haksân = n-hexane » se voit attribuer la classe 'Composé chimique', la sous-classe 'Hydrocarbure', la catégorie 'Hydrocarbure acyclique' et la sous-catégorie 'Alcane' (cf. Figure 40 : Classe du terme « ن-هكسان = n-haksân = n-hexane ») :

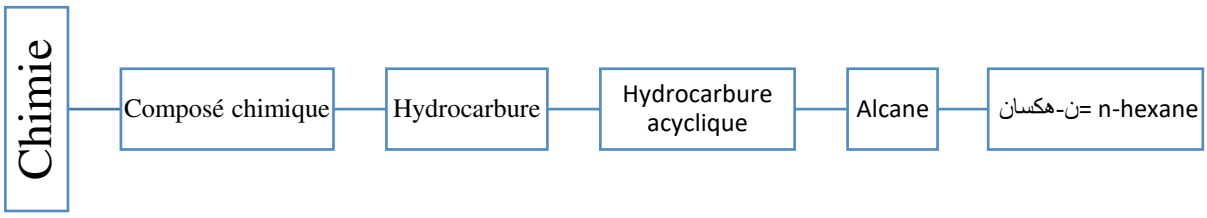


Figure 40 : Classe du terme « ن-هكسان= n-haksân = n-hexane »

- Les fonctions chimiques

Elles permettent la détermination du nom d'une molécule fonctionnalisée. Nous pouvons citer les groupes fonctionnels principaux : les alcools, les éthers, les aldéhydes, les cétones, les acides carboxyliques, les esters, les amines et les amides ; cependant, tout composé chimique n'a pas nécessairement de groupe fonctionnel.

Voici l'exemple du terme « إيثانول = *'îṯânûl* = éthanol » qui se voit attribuer la classe 'Composé chimique', la sous-classe 'Fonction chimique', la catégorie 'Alcool' (cf. Figure 41 : Classe du terme « إيثانول = 'îṯânûl = éthanol ») :

Figure 41 : Classe du terme « *إيثانول = ʾiṯânûl* = éthanol »

- Espèces chimiques

D'autres constituants de la matière existent (ou vont être découverts) ; cependant, ce ne sont ni des éléments chimiques, ni des hydrocarbures et ils ne possèdent pas de fonctions chimiques. Ce sont des substances chimiques telles que l'eau ou le chlorure de sodium ou encore les sulfates ; nous ajoutons alors une nouvelle catégorie 'Espèce chimique'.

Voici l'exemple du terme « كبريتات الصوديوم = *kibrîtât alṣûdyûm* = sulfate de sodium » qui se voit attribuer la classe 'Espèce chimique', la sous-classe 'Composé ionique' (cf. Figure 42 : Classe du terme « كبريتات الصوديوم = kibrîtât alṣûdyûm = sulfate de sodium ») :

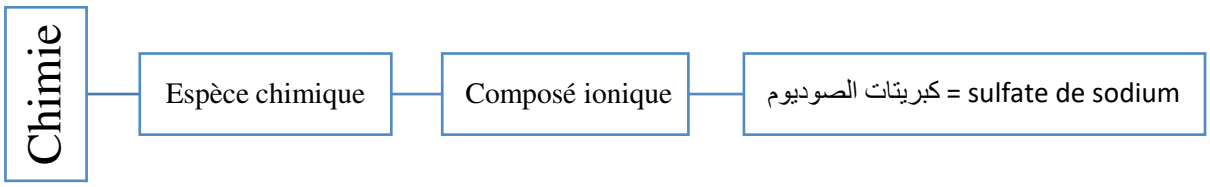


Figure 42 : Classe du terme « *كبريتات الصوديوم = kibrîtât alṣûdyûm = sulfate de sodium* »

Ainsi, nous définissons la classification de la matière (cf. Figure 43 : Classification de la matière), classe constituée des sous-classes 'Éléments chimiques' et 'composés chimiques', et nous prenons en compte la structure atomique de la matière en ajoutons les sous-classes 'Atome' et 'Liaison'.

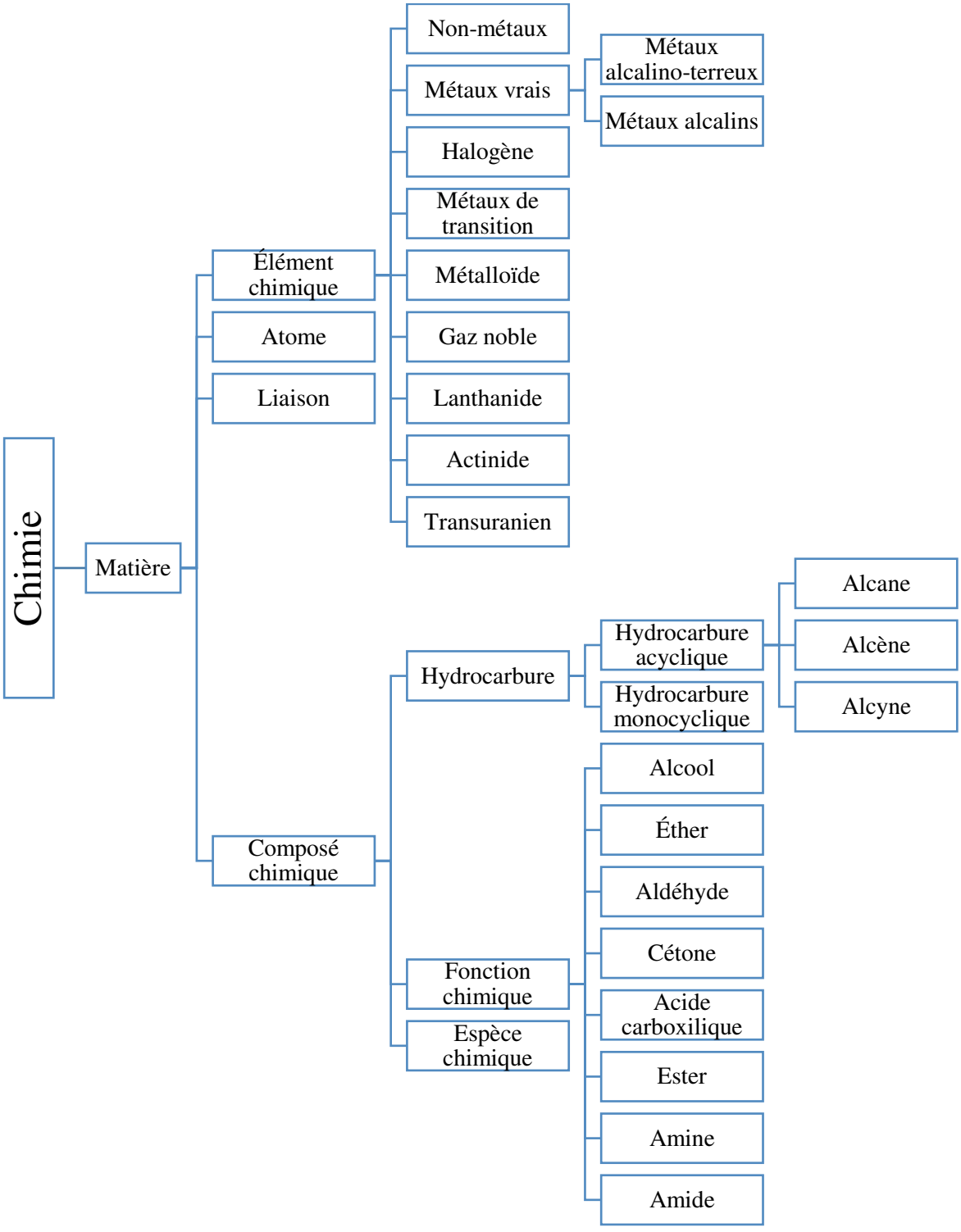


Figure 43 : Classification de la matière

5.3.2 Classification des méthodes expérimentales

Dans notre corpus, nous identifions des termes tels que ‘réaction chimique’, ‘pureté’, ‘hydrogénation asymétrique’, ‘condensation aldolique’, ‘oxydation’, ‘spectromètre infrarouge’, ‘rendement chimique’, … Ces termes peuvent être classés en deux catégories :

d’une part, les réactions chimiques mises en jeu lors de ces transformations chimiques, et d’autre part, les procédures chimiques à suivre pour les effectuer, en indiquant les équipements et les méthodes d’analyse employés. Il existe ici une relation entre une procédure chimique et une réaction chimique.

5.3.2.1 Réaction chimique

Une réaction chimique est un processus au cours duquel certaines espèces chimiques se transforment en d'autres ; ce processus présente la particularité de conserver intégralement la matière, d’où la maxime attribuée à Lavoisier « rien ne se perd, rien ne se crée, tout se transforme ». Décrite grâce à une équation chimique, une réaction chimique correspond à la rupture et à la formation d'une ou de plusieurs liaisons chimiques, ou au transfert d'électrons, ou encore aux interactions entre anions et cations. Il existe exactement 37 types de réactions chimiques[387] et voici les principales : les réactions acido-basiques, les réactions d’oxydoréductions, la combustion, l’hydrogénation.

Par exemple, le terme « هدرجة لامتناظرة = hadrajat lâmutanâẓirat = hydrogénation asymétrique » se voit attribuer la classe ‘Réaction chimique’, la sous-classe ‘Hydrogénation’ (cf. Figure 44 : Classe du terme « هدرجة لامتناظرة = hadrajat lâmutanâẓirat = hydrogénation asymétrique ») :

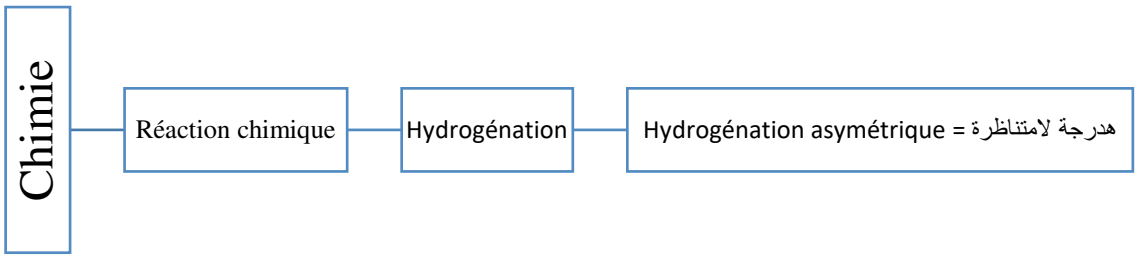


Figure 44 : Classe du terme « *هدرجة لامتناظرة* = hadrajat lâmutanâẓirat = hydrogénation asymétrique »

5.3.2.2 Procédure chimique

Une procédure chimique correspond aux instructions à suivre pour réaliser les réactions chimiques ; il s’agit de la préparation d’une solution afin d’obtenir le produit final. Pour cela, il faut peser les réactifs et peut-être ajouter des solvants ; le mélange est éventuellement

[387] http://villemin.gerard.free.fr/aScience/Chimie/aaaExpli/Reaction.htm

transvasé afin de récupérer le produit de réaction par séchage, ou par filtration, ou encore par distillation, …

Par exemple, le terme « ترشيح = *taršîḥ* = filtration » se voit attribuer la classe ‘Procédure chimique’ (cf.Figure 45 : Classe du terme « ترشيح = taršîḥ = filtration ») :

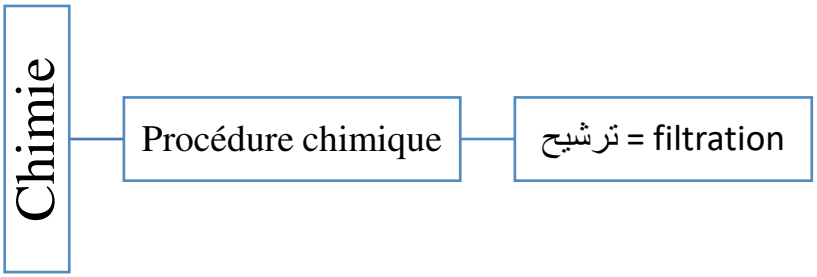


Figure 45 : Classe du terme « *ترشيح* = *taršîḥ* = filtration »

5.3.2.3 Équipement

Ces procédures chimiques nécessitent l’utilisation d’un équipement précis afin d’obtenir un meilleur rendement. Cela désigne la verrerie (équipements généralement en verre), les instruments et les appareils utilisés en laboratoire par les chimistes.

Par exemple, le terme « قمع فصل = qimʿ faṣl = ampoule à décanter » se voit attribuer la classe ‘Équipement’, la sous-classe ‘Verrerie’ (cf. Figure 46 : Classe du terme « قمع فصل = qimʿ faṣl = ampoule à décanter ») :

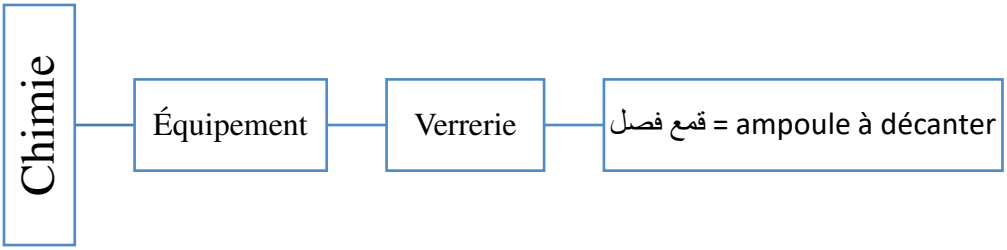


Figure 46 : Classe du terme « *قمع فصل* = qimʿ faṣl = ampoule à décanter »

5.3.2.4 Méthode d'analyse

Après avoir obtenu le produit final, il faut vérifier qu'il s'agisse du bon produit. Différentes analyses sont possibles telles que la spectroscopie ou la chromatographie.

Par exemple, le terme « كروماتوغرافيا العمود = krûmâtûġâfiyâ alʿamûd = chromatographie sur colonne » se voit attribuer la classe 'Méthode d'analyse', la sous-classe 'Chromatographie' (cf. Figure 47 : Classe du terme « كروماتوغرافيا العمود = *krûmâtûġ*râfiyâ alʿamûd = chromatographie sur colonne ») :

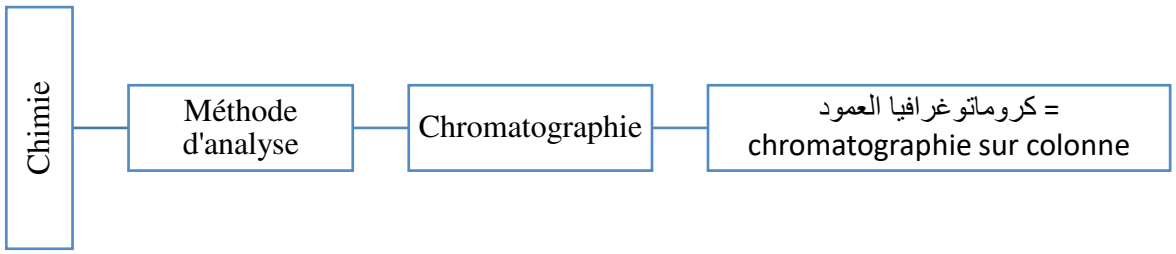


Figure 47 : Classe du terme « *كروماتوغرافيا العمود* = krûmâtûġrâfiyâ alʿamûd = chromatographie sur colonne »

Ainsi, nous obtenons notre classification de la chimie (cf. Figure 48 : Classification adoptée). Elle est composée des six classes 'Matière', 'État', 'Réaction chimique', 'Méthode d'analyse', 'Procédure chimique' et 'Équipement' et chacune de ses classes est constituée de sous-classes qui peuvent également avoir des catégories et des sous-catégories.

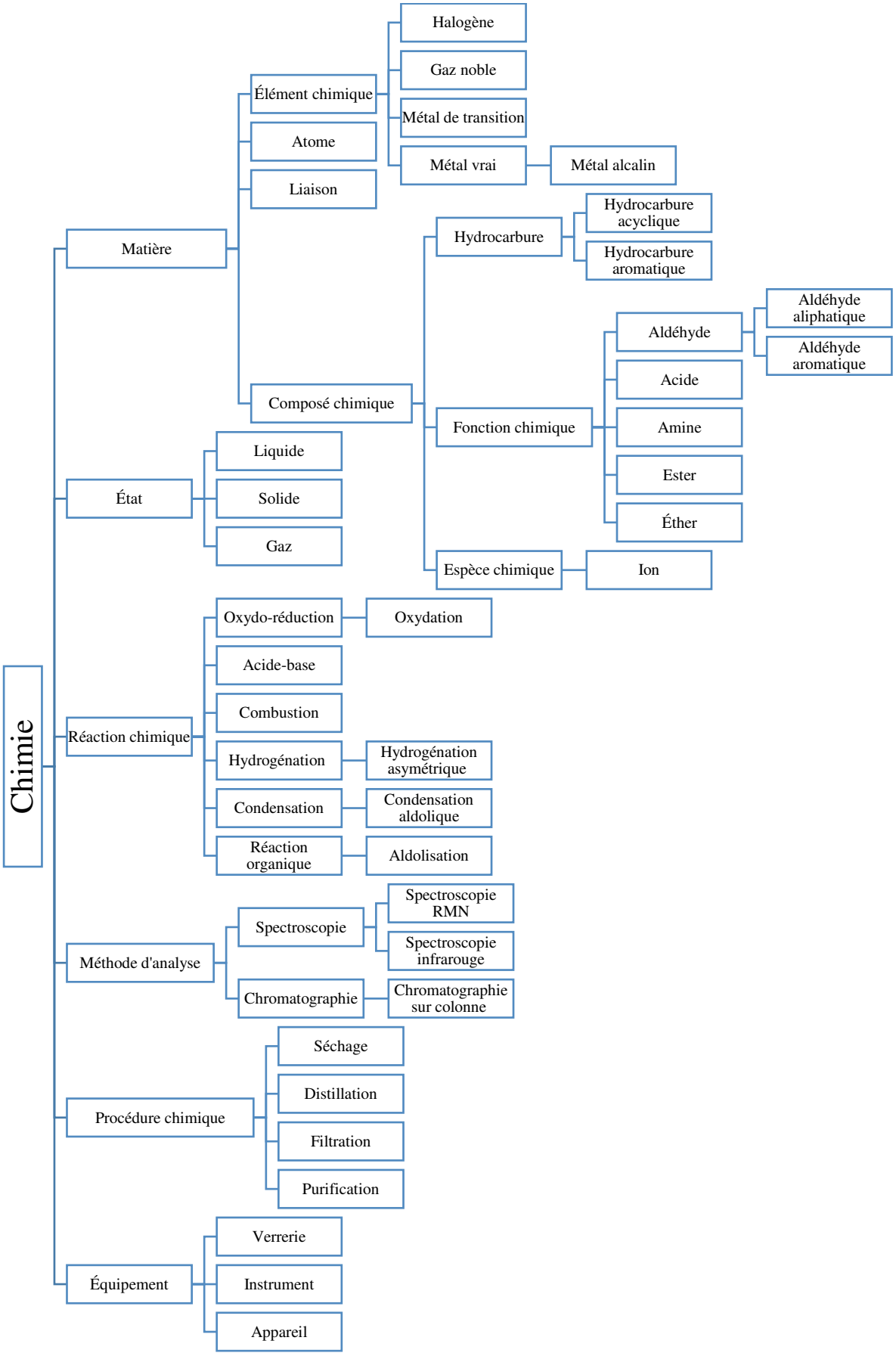


Figure 48 : Classification adoptée

Cette classification nous permet de prendre en compte tous les termes de notre corpus. Par conséquent, nous pouvons la considérer comme la classification de la chimie. De plus, nous l'avons soumis à un autre expert chimiste qui l'a également validé. Nous avons conçu cette classification de telle sorte que l'on puisse la compléter par d'autres termes de la chimie non présents dans le corpus en créant de nouvelles classes et sous-classes.

À partir du corpus recueilli et des quelques outils informatiques dont nous disposons, nous avons identifié et classé tous les termes de la chimie de notre corpus. À présent, nous pouvons construire une grammaire d'identification pour ces termes.

Partie III : Modélisation de l'extraction des unités terminologiques de la chimie et de leurs relations morphosyntaxiques

À partir de notre étude terminologique exposée dans la première partie et la classification construite dans la deuxième, nous pouvons maintenant développer une méthode d'extraction des unités terminologiques de la chimie en arabe ainsi que leurs relations morphosyntaxiques.

Dans cette dernière partie, nous nous intéressons d'abord à la modélisation de l'extracteur terminologique par l'établissement de la grammaire d'identification des unités terminologiques de la chimie en arabe. Puis, nous décrivons une des applications de ce travail, à savoir, une ontologie afin de représenter les connaissances du domaine étudié.

Chapitre 6 : Fouille de textes

Dans ce chapitre, nous présentons la modélisation de l'extracteur terminologique par l'établissement de la grammaire d'identification des unités terminologiques de la chimie en arabe. Pour cela, nous définissons la notion de fouille de textes et ses défis dans la recherche d'aujourd'hui, en exposant ses contributions et ses applications. Puis, nous présentons la méthode de modélisation de l'extracteur terminologique avant de terminer par quelques pistes du traitement informatique.

6.1 Fouille de textes : définitions

Traduction approximative de l'anglais « text mining », la fouille de textes peut être définie comme étant l'acquisition des connaissances (données) enfouies (fouille) dans des corpus de textes, impliquant que la fouille de textes peut désigner différents types d'activités et/ou englobant l'ensemble de ces activités (Toussaint, 2011)[388]. Mais, la plupart du temps, elle fait référence à la recherche d'information ou à l'extraction de connaissances (Prince et al., 2007)[389]. Cependant, certains chercheurs considèrent que la fouille de textes n'est pas une étape dans le processus d'extraction de connaissances (Roche, 2004)[390]. Mais, nous adoptons l'avis de Toussaint considérant « la fouille de textes *comme un paradigme de l'extraction de* connaissances à partir de données pour lequel les données sont des textes » et qu'elle doit « *permettre à un expert ou un ensemble d'experts d'un domaine d'avoir une vision synthétique* de son domaine, et notamment des dernières avancées en les représentant dans un langage de représentation de connaissances » (Toussaint, 2011)[391].

Les principaux buts de la fouille de textes sont :

- *« Le recueil de renseignements fiables par la découverte de termes et de concepts présents dans les textes ;*

[388] Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances, p. 3

[389] Le Défi fouilles de textes : quels paradigmes pour la reconnaissance d'auteurs ?, p. 3

[390] Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes

[391] Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances, p. 3

- *Le recueil des relations ou règles d'associations entre les termes et les concepts précédents ;*
- *La détection de tendances à partir de textes ;*
- *Le regroupement et la sélection de documents en fonction de concepts ou thèmes communs ;*
- *La production de résumés (d'un seul document ou d'un ensemble de documents) ;*
- *La réponse à des questions précises (« Quand la France a-t-elle gagné la Coupe du Monde de Football ?») ou plus générales (« Parlez-moi de Che Guevara ») en utilisant une masse de textes comme source d'informations. » (Pillet, 2000)*[392]

6.1.1 Historique de la fouille de textes

A l'origine, l'informatique et la linguistique sont deux domaines qui ne devaient pas se rencontrer (Tellier, 2007)[393] ; mais, le développement de l'informatique dans les années trente et la guerre froide entre les Américains et les Russes ont été les éléments déclencheurs pour donner naissance à la traduction automatique. En effet, comprendre la langue de l'ennemi a été un besoin nécessaire pour remporter la guerre ; la légende raconte qu'en partant de « The spirit is willing but the flesh is weak » (l'esprit est fort mais la chair est faible), après un aller-retour russe-anglais, on obtint alors « The vodka is strong but the meat is rotten » (la vodka est forte mais la viande est pourrie). De ce fait, des recherches sont mises en place dans le traitement de la langue en vue d'améliorer et d'obtenir un système de traduction automatique.

Cependant, ce projet ambitieux mène à une impasse et il faut attendre les années 70-80 pour parler du Traitement Automatique de la Langue (TAL) qui s'intéresse à la compréhension automatique de textes afin de saisir le sens global : les notions de « scripts », de « réseaux sémantiques », de « graphes conceptuels » et de « sémantique » apparaissent.

Mais, face aux problèmes théoriques et pratiques, aux ambiguïtés et à la difficulté de compréhension du texte, des nouvelles orientations sont proposées avec l'émergence de la linguistique de corpus dans les années 90, notamment la fouille de textes (extraction d'information).

[392] Méthodologie d'extraction automatique d'information à partir de la littérature scientifique en vue d'alimenter un nouveau système d'information

[393] Introduction au TALN et à l'ingénierie linguistique

Son succès grandissant, la fouille de textes se développe grâce à ses nombreuses applications concrètes et relativement spectaculaires (Prince et al., 2007)[394] ; elle devient pluridisciplinaire en touchant tous les domaines : statistiques, analyse des données et reconnaissance de formes, apprentissage, intelligence artificielle, traitement automatique de la langue, économie, ingénierie documentaire, terminologie… (Ibekwe-Sanjuan, 2007)[395]. Parmi ces domaines, le traitement automatique de la langue et la terminologie nous intéressent.

Comme la fouille de textes, le traitement automatique de la langue (TAL) se positionne à la frontière entre la linguistique et l'informatique où ses applications sont assez diverses : la traduction automatique, la génération automatique de textes, la représentation du contenu d'une phrase ou d'un discours... C'est ce dernier point qui lie le TAL à la fouille de textes et à notre travail, puisque l'objectif de cette application du TAL est de savoir comment extraire et représenter des éléments de sens à partir de textes. Le TAL suggère de modéliser, parfois de façon très précise, les phénomènes de langues dans un cadre logique et en proposant de représenter la sémantique d'une phrase et/ou d'un discours sous forme logique. Cela conduirait, tout simplement, à traduire une forme linguistique en une forme logique. Mais, dans la pratique, cela est plus compliqué, notamment en raison de la complexité de la langue, et requiert la mise en place de règles de grammaire pouvant être associées à des méthodes statistiques. De plus, « *le TAL est souvent considéré dans un processus de fouille de textes comme l'étape* permettant de passer du texte à des données pour que des processus de fouille de données puissent être appliqués » (Toussaint, 2011)[396]. Cela implique que notre travail de recherche s'inscrit autant dans la fouille de textes que dans le TAL.

Depuis plus d'une dizaine d'années, les travaux en terminologie se sont vus de plus en plus associés à la fouille de textes, notamment avec des approches sur corpus (linguistique de corpus) (cf. 1.2.3 Terminologie et linguistique de corpus) et des méthodes statistiques pour la recherche de patrons syntaxiques (cf. 1.2.4 Terminologie et ontologie). Cela joue un rôle essentiel dans l'accès au contenu des textes dans des domaines spécialisés, puisque le terme est construit à partir des manifestations linguistiques en corpus qui permettent de définir le concept (Aussenac-Gilles, 2005)[397]. De ce fait, notre travail de recherche suit cette continuité en associant la terminologie à la fouille de textes.

[394] Le Défi fouilles de textes : quels paradigmes pour la reconnaissance d'auteurs ?, p. 3

[395] Fouille de textes : méthodes, outils et applications

[396] Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances, p. 9

[397] Méthodes ascendantes pour l'ingénierie des connaissances

6.1.2 Application de la fouille de textes

La fouille de textes réunit et intègre dans ses applications des méthodes de recherche d'information et d'extraction d'information ; nous proposons de les présenter.

6.1.2.1 Recherche d'information

Issue de la recherche documentaire, la recherche d'information (RI) est la problématique la plus ancienne de la fouille de textes. Elle se définit par un ensemble de méthodes et d'outils, permettant à un utilisateur de chercher des données pertinentes pour un objectif spécifique, par la formulation d'une requête en langue naturelle ou avec des mots-clés, afin de sélectionner les documents répondant à ses critères.

Classiquement, un processus de RI est composé de plusieurs fonctions : « *l'indexation des* documents et des requêtes, la mise en correspondance requête-documents avec un ordonnancement des documents quand le modèle le permet, la restitution des documents reconnus pertinents par rapport à la requête » (Champclaux, 2010)[398].

Cependant, ces outils et ces méthodes de la RI (analyse ou caractérisation automatique de documents, classification automatique, structure des documents et des fichiers…) sont souvent attribués à la fouille de textes :

> *« Les méthodes de classification et de cartographie appliquées aux mots-clés ou aux termes d'un domaine sont assez souvent qualifiées d'outil de fouille de textes […] De ce fait, il y a souvent confusion entre d'une part, la classification (non supervisée) de textes, la catégorisation (classification supervisée) de textes et l'identification de thèmes, ou encore la détection de tendances avec, d'autre part, le processus de construction de connaissances à partir de textes dont ils peuvent éventuellement faire partie. (Toussaint, 2011)*[399]

Or, un système de RI doit pouvoir proposer en réponse à une requête les documents les plus pertinents et cela doit pouvoir s'appliquer à de gros volumes de données, ce qui implique que le domaine étudié est faiblement pris en compte. Mais, des améliorations sont en cours de développement, puisque le domaine de recherche est très actif, notamment avec la percée de l'ontologie utilisant des relations lexicales et sémantiques. Cela permet d'obtenir les résultats d'une requête présentée sous forme d'une liste structurée (classe, hiérarchie), ainsi que de coder

[398] Un modèle de recherche d'information basé sur les graphes et les similarités structurelles pour l'amélioration du processus de recherche d'information, p. 34

[399] Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances, p. 5

l'information contenue dans un document, pouvant se faire au niveau des mots contenus dans ce document, qui correspondent aux termes d'un domaine.

De ce fait, notre travail de recherche peut s'intégrer également au domaine de la RI.

6.1.2.2 Extraction d'information

L'extraction d'information (EI) consiste à analyser un texte donné afin d'en extraire des informations spécifiques qui peuvent être des mots et/ou des séquences de mots.

Son principe repose sur trois grandes étapes : l'identification des fragments de textes pertinents, contenant une information, la définition de la structure de représentation de l'information et le développement des règles, permettant d'identifier l'information, afin de remplir la structure proposée (Nédellec et al., 2001)[400]. Cela permet de décomposer l'identification d'informations, parfois complexes, en des sous-problèmes simples, afin de reconnaitre des entités nommées, de résoudre des coréférences, d'extraire des propriétés, d'identifier des relations et/ou des évènements (Humphreys et al., 2000] [401]:

Pour cela, deux approches sont possibles. D'une part, les techniques statistiques à base d'apprentissage, utilisées pour détecter les régularités sur le texte où les informations recherchées ont été préalablement annotées ; cependant son inconvénient principal est le manque de corpus annoté et que selon la taille du corpus, les résultats varient (Serrano et al., 2011) [402]. D'autre part, l'extraction basée sur des techniques linguistiques, exploitée par le TAL, dont le principe repose sur « *l'utilisati*on de grammaires formelles construites par la *main d'un expert*-linguiste » (Serrano et al., 2011) [403]. Ainsi, un système hybride alliant ces deux approches pour s'affranchir de leurs limites respectives est possible et permet d'obtenir des meilleurs résultats.

Pour le domaine de la terminologie, Toussaint (2011)[404] parle d'extraction terminologique, libre ou contrôlée ; la première permet d'obtenir des candidats termes, privilégiant un silence faible au détriment du bruit qui est assez élevé. Quant à la seconde, elle permet de reconnaître les termes, générant assez peu de bruit mais plus de silence. Afin d'optimiser l'extraction terminologique, une recherche des variantes des termes, représentant entre 15 et 35% des

[400]Sentence filtering for information extraction in genomics, a classification problem

[401] Two applications of information extraction to biological science journal articles : enzyme inter-actions and protein structures

[402] Extraction de connaissances pour le renseignement en sources ouvertes, p. 4

[403] Idem

[404] Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances, p. 57

occurrences des termes, selon le domaine et le type de corpus, permet de réduire le silence et le bruit (Daille, 2002)[405] ; cela suppose que le corpus soit étiqueté et lemmatisé.

Nous proposons de combiner un système hybride, qui prendra en compte les techniques statistiques à base d'apprentissage et les techniques linguistiques, à une extraction terminologique contrôlée, naviguant entre les termes et leurs variantes, suivant une stratégie descendante, basée sur des patrons terminologiques.

6.2 Modélisation de l'extraction terminologique

La modélisation des connaissances à partir de textes consiste en la représentation des informations recherchées dans un modèle donné. De ce fait, elle exige une certaine habileté dans l'utilisation des textes et des modèles, puisque de nombreux aller-retour entre l'étude linguistique et la modélisation sont nécessaires afin d'atteindre l'objectif visé.

Quant à l'extraction d'information, elle consiste à utiliser un programme informatique qui reçoit en entrée un corpus de textes portant sur un domaine donné, et qui doit en sortie proposer l'information recherchée, sous forme de mots et/ou de séquences de mots, extraite de ce corpus ; ce programme se focalise sur le repérage des termes et identifie les patrons morphosyntaxiques en s'appuyant sur des règles de grammaire.

6.2.1 Unité terminologique

Comme nous l'avons vu, l'unité terminologique, est l'unité signifiante constituée d'un mot, l'unité terminologique simple (UTS) ou de plusieurs mots, l'unité terminologique complexe (UTC) (cf. 1.1.1 Terme).

6.2.1.1 Unité terminologique simple (UTS)

L'identification des unités terminologiques simples (UTS) et la recherche d'attestations sont extrêmement simples, sans grande exigence technique et linguistique. Généralement, elles sont examinées minutieusement à la main par le chercheur et les besoins en outillage pour ce type de travail sont assez légers, puisque bien souvent, un simple concordancier fait amplement

[405] Découvertes linguistiques en corpus

l'affaire, si ce n'est pas la fonction 'rechercher' du logiciel de traitement de texte (Tanguy, 2012)[406].

Rappelons que pour notre domaine, ces UTS correspondent à des noms désignant notamment des éléments chimiques, comme « أكسجين = uksijîn = oxygène », ou à des verbes indiquant notamment une procédure chimique, comme « بلور = balwara = cristalliser ». Ce sont des termes du domaine, mais également des déclencheurs des termes du domaine, puisque ces UTS peuvent être un élément d'une UTC, soit la base, soit l'extension, permettant d'identifier les UTC du domaine, comme « قمع فصل = qimʿ faṣl = ampoule à décanter ». Les UTS de la chimie ont été identifiées et nous les avons encodées dans un éditeur, sous forme de tableau (cf. 4.2.2.1**Erreur ! Source du renvoi introuvable.** Identification des termes de la chimie).

Grâce à la mise en place de la classification de la chimie (cf. 5.3 Classification adoptée), à l'analyse de la formation des termes en arabe (cf. 2.1 Système de la langue arabe) et à l'aide des logiciels AntConc pour étudier les cooccurrences, Kawâkib pour analyser les racines et Xerox pour réaliser l'analyseur morphologique (cf. 4.2 Analyse des formes), nous détenons toutes les informations nécessaires pour rédiger les fiches terminologiques de chaque UTS du corpus.

6.2.1.1.1 Fiche terminologique du verbe

Dans notre travail, nous avons considéré le verbe comme un terme (cf. 2.1.2.12.1.2.1 Modus personnel ou verbe et 4.2.3.1 Lemmatisation du verbe) et nous classons les verbes en trois catégories : les verbes très spécialisés, comme « بلور = balwara = cristalliser », les verbes employés dans le domaine étudié, comme « شرّح = *šarraḥa* = filtrer » et les verbes énonciateurs, comme « شكّل = *šakkala* = constituer » (Albeiriss, 2017)[407]. Nous prenons en compte leur structure argumentale (complément et circonstant) permettant d'identifier et/ou de vérifier les termes de nature nominale du domaine et d'indiquer, s'il y a lieu, la ou les prépositions privilégiées.

Voici l'exemple d'une fiche terminologique pour « جفّف = jaffafa = sécher » :

[406] Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes, p. 55

[407] Terminologie basée sur corpus : les verbes de la chimie en arabe

جفّف= jaffafa = sécher			
Classe, sous-classe	Verbe polysémique dont au moins un sens est spécialisé		
Définition	Rendre un composé chimique sec en faisant évaporer le liquide		
Structure argumentale	~ X (Composé chimique) prép Y (Composé chimique) prép Z (Grandeur) ~ prép X (Composé chimique)		
Réalisation linguistique des arguments	مدة 24 ساعة muddat 24 sâ ʼat pendant 24 heures	فوق $CaCl_2$ fawqa $CaCl_2$ sur $CaCl_2$ فوق كبريتات الصوديوم Na_2SO_4 fawqa kibrîtât alṣûdyûm sur du sulfate de sodium Na_2SO_4.	الإيثر al ʼîṯr l'éther
Contexte	يجفف الإيثر فوق $CaCl_2$ مدة 24 ساعة. (Réf. 1.b) yujaffafu al ʼîṯr fawqa $CaCl_2$ muddat 24 sâ ʼat. L'éther est séché sur $CaCl_2$ pendant 24 heures.		
	و تجفف فوق كبريتات الصوديوم Na_2SO_4. (Réf. 1.b) wa tujaffafu fawqa kibrîtât alṣûdyûm Na_2SO_4. Et il est séché sur du sulfate de sodium Na_2SO_4.		
	نقوم بتجفيفها فوق كبريتات المغنيسيو $MgSO_4$. (Réf. 1.b). naqûm bitajfîfihâ fawqa kibrîtât almaġnîsyûm $MgSO_4$. Nous effectuons son séchage sur du sulfate de magnésium $MgSO_4$.		
Terme relié sémantiquement / associé	تجفيف		
Construction syntaxique	X est séché sur Y pendant Z		
Modélisation	UTS (Verbe polysémique) + prép + UTS (Composé chimique) UTS (Verbe polysémique) + UTS (Composé chimique) + prép + UTS (Élément chimique)		

Tableau 15 : Fiche terminologique de « جفّف = jaffafa = sécher »

Nous avons rédigé les fiches terminologiques des verbes de notre corpus et nous présentons un échantillon en annexe du domaine de la chimie (cf. Annexe 3 : Fiches terminologiques).

6.2.1.1.2 Fiche terminologique du nom

Le nom est considéré comme un terme (cf. 2.1.2.22.1.2.2 Modus impersonnel et 4.2.3.24.2.3.2 Lemmatisation du nom) ; nous classons le nom selon sa place dans le domaine de la chimie : composé chimique, réaction chimique, procédure chimique, équipement… (cf. 5.35.3 Classification adoptée). Nous prenons en compte sa fonction nominale (agent, objet,

instrument, lieu, nom d'action…) et sa construction syntaxique, sachant que ces noms peuvent être des emprunts.

Voici l'exemple d'une fiche terminologique pour « تيتانيوم = tîtâniyûm = titane » :

تيتانيوم = tîtâniyûm = titane	
Classe, sous-classe	Élément chimique, métal de transition, solide
Définition	Élément chimique, qui a l'état stable, est sous forme de métal blanc à éclat métallique et possède une densité égale à 4,5, une résistance mécanique assez faible associée à une grande ductilité et une bonne résistance à la corrosion.
Catégorie lexicale	Nom
Construction syntaxique	Ø
Contexte	نضع 0.78 غ، 3.59 ميلي مول من معقد التيتانيوم (Réf 1.b)
Termes reliés sémantiquement / associés	تيتان - Ti
Modélisation	UTS (Elément chimique)

Tableau 16 : Fiche terminologique de « *تيتانيوم* = tîtâniyûm = titane »

Nous avons rédigé les fiches terminologiques des noms de notre corpus et nous présentons un échantillon en annexe (cf. Annexe 3 : Fiches terminologiques).

6.2.1.2 Unité terminologique complexe (UTC)

L'identification et la détermination de ces unités terminologiques complexes (UTC) sont beaucoup plus exigeantes que celles des UTS, ce qui implique que leur extraction doit être sensiblement plus sophistiquée :

> *« Or le fait de travailler sur une structure, et non une forme, particulière de la langue complique la recherche automatique d'attestations, puisqu'elle ne permet pas la simple recherche par chaîne de caractères qu'on peut pratiquer à l'aide de n'importe quel logiciel de traitement de texte ou moteur de recherche. » (Leroy, 2004)*[408]

Ces UTC sont considérées en informatique comme des séquences de mots qui se répètent normalement plus d'une fois côte à côte dans un texte, appelées segments répétés ; puisqu'ils apparaissent souvent ensemble d'une manière statistiquement significative, ils ont une grande chance de former des UTC. De ce fait, ces UTC sont structurées en 'réseau terminologique' en se basant sur la décomposition de ces termes en 'tête (T)' et 'expansion (E)' et reliant chaque

[408] Extraire sur patrons : allers et retours entre analyse linguistique et repérage automatique, p. 26

élément de l'UTC à sa tête et à son expansion, et réciproquement (Harrathi, 2009)[409]. Pour notre analyse de ces UTC, nous adoptons 'base' et 'expansion' pour les éléments composant une UTC et 'schéma morphosyntaxique ' la structure de l'UTC (cf. 2.1.3.22.1.3.2 Formation des unités terminologiques complexes).

Ces UTC de la chimie ont été identifiées et nous les avons encodées dans un éditeur, sous forme de tableau (cf. **Erreur ! Source du renvoi introuvable.**4.2.2.1 Identification des termes de la chimie).

De la même manière que pour les UTS, nous rédigeons les fiches terminologiques de chaque UTC du corpus et voici un exemple.

ثلاثي عنق = *ṯulâṯî ʿunq* = tricol	
Classe, sous-classe	Equipement, verrerie
Définition	Récipient largement utilisé en verrerie de laboratoire, constitué de trois cols.
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Expansion d'annexion
Modélisation	UTC = UL + UTS
Contexte	في ثلاثي عنق ذو سعة 100 مل، نضع 0.78 غ. (Réf 1.c)
Termes reliés sémantiquement / associés	عنق - ثنائي عنق

Tableau 17 : Fiche terminologique de « *ثلاثي عنق = ṯulâṯî ʿunq* = tricol »

Nous avons rédigé les fiches terminologiques des UTC de notre corpus et nous présentons un échantillon en annexe (cf. Annexe 3 : Fiches terminologiques).

Rappelons que pour notre domaine, ces UTC désignent, notamment des composés chimiques, comme « بيرمنغنات البوتاسيوم = *bîrmanġanât albûtâsyûm* = permanganate de potassium » (cf. 2.1.3.2 Formation des unités terminologiques complexes (UTC)). Mais, tous les éléments d'une UTC ne sont pas nécessairement des UTS du domaine et sont alors des unités lexicales (UL), comme « مدة التفاعل = muddat altafâʿul = temps de réaction », où « muddat » est une unité lexicale (UL), considérée comme un déclencheur d'une UTC du domaine de la chimie.
Nous allons modéliser les différentes relations entre les éléments de ces UTC par des patrons morphosyntaxiques.

[409] Extraction de concepts et de relations entre concepts à partir des documents multilingues : Approche statistique et ontologique, p. 73

6.2.2 Patron morphosyntaxique

Un patron morphosyntaxique est la modélisation des différentes relations entre les éléments des syntagmes, aussi bien sur le plan grammatical que sur le plan sémantique, permettant d'identifier les types de constructions morphosyntaxiques, afin de les formaliser autant que possible de façon à en déduire des règles opératoires de dépistage, appelée « règles de grammaire » (cf. 6.2.36.2.3 Règle de grammaire). Ces patrons morphosyntaxiques, ou patrons de surface, doivent couvrir une grande quantité de données, en offrant une grande variété de structures syntaxiques et en proposant des solutions pour résoudre les problèmes de l'ambiguïté. Ils sont construits à partir d'éléments lexicaux, syntaxiques et des types sémantiques, en prenant en compte les indications de séquentialité dans le texte (Toussaint, 2011)[410], l'objectif étant de définir des patrons correspondant à des structures qui expriment une relation morphosyntaxique précise entre des unités lexicales et/ou terminologiques.

À partir de notre analyse terminologique des UTC de la chimie (cf. 2.1.3.2 Formation des unités terminologiques complexes), nous modélisons les différentes relations entre les éléments de ces UTC par des patrons morphosyntaxiques nominaux, adverbiaux et verbaux.

6.2.2.1 Patron morphosyntaxique nominal

Ces patrons syntaxiques nominaux sont de trois types : l'expansion d'annexion, l'expansion complétive et l'expansion d'identification (Albeiriss, 2017)[411]. De plus, ils peuvent se combiner entre eux (cf. 2.1.3.22.1.3.2 Formation des unités terminologiques complexes).

6.2.2.1.1 Expansion d'identification

Ce patron syntaxique est un syntagme nominal, formé d'une base nominale, pouvant correspondre à une UTS, et d'une expansion adjectivale, pouvant être une UTS. Cette expansion compte deux patrons syntaxiques :

- UTS (N) + UTS (Adj) : « مركب عضوي = murakkab ʿuḍwî = composé organique »
- UL (N) + UTS (Adj) : « وظيفة حمضية = waẓîfat ḥamḍiyyat = fonction acide »

À ces patrons syntaxiques, nous précisons la classification des UTS :

- Terme (Phénomène Physique) + Terme (Composé chimique)Adj : « تكاثف الدولي = takâ*ṯuf* 'aldûlî = condensation aldolique »

[410] Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances, p. 8

[411] Modélisation des unités lexicales complexes d'une langue spécialisée : le cas de la chimie en arabe

Nous avons recensé dans notre corpus ces UTC formées par expansion d'identification et nous l'indiquons dans les fiches terminologiques (cf. Annexe 3 : Fiches terminologiques).

6.2.2.1.2 Expansion d'annexion

Ce patron syntaxique est un syntagme nominal, formé d'une base, correspondant à une UTS, et d'une expansion annective, pouvant être une UTS. Cette expansion compte deux patrons syntaxiques :

- UTS (N) + UTS (Ann) : « طيف الامتصاص = *ṭayf alimtiṣâṣ* = spectre d'absorption »
- UL (N) + UTS (Ann) : « مجموعة الأمين = majmûʿat al'amîn = groupe amine »

À ces patrons syntaxiques, nous précisons la classification des UTS :

- UL (N) + Terme (Physique)Ann : « درجة الإنصهار = darajat al*inṣihâ*r = point de fusion »

Nous avons recensé dans notre corpus ces UTC, formées par expansion d'annexion et nous l'indiquons dans les fiches terminologiques (cf. Annexe 3 : Fiches terminologiques).

6.2.2.1.3 Expansion complétive

Ce patron syntaxique est un syntagme nominal, formé d'une base, correspondant à une UTS, d'une préposition et d'une expansion, pouvant être une UTS. Cette expansion compte un patron syntaxique :

- UL (N) + prép + UTS (N) : « جو من الأزوت = jaww *min al'azût* = atmosphère d'azote »

À ce patron syntaxique, nous précisons la classification des UTS :

- UL (N) + prép + Terme (Composé chimique) : « جو من الأزوت = jaw*w min al'azût* = atmosphère d'azote »

Nous avons recensé dans notre corpus ces UTC, formées par expansion complétive et nous l'indiquons dans les fiches terminologiques (cf. Annexe 3 : Fiches terminologiques).

6.2.2.1.4 Combinaison des patrons morphosyntaxiques

Certaines UTC sont formées d'une base constituée elle-même d'une UTC ou d'une expansion elle-même composée d'une UTC ou les deux.

6.2.2.1.4.1 Expansion d'identification et expansion de coordination

Ce patron syntaxique est un syntagme nominal, formé d'une base nominale, correspondant à une UTS et d'une expansion adjectivale, pouvant être une UTS, suivie d'un coordonnant et d'une expansion, correspondant aussi à une UTS ; à cela, il faut ajouter la possibilité d'avoir une UL à la place d'une UTS :

- UTS (N) + UTS (Adj) + coord + UTS (Adj) : « ألدهيدات أليفاتيـة و عطرية = *'aldahîd*ât *'alîf*âtiyat wa *ʿiṭrîyat* = aldéhydes aliphatiques et aromatiques »

À ce patron syntaxique, nous précisons la classification des UTS :

Terme (Composé chimique) + Terme (Composé chimique)Adj + coord + Terme (Composé chimique) : « أحماض آمينية و مشتقاتها = *'aḥmâḍ 'amîniyat wa mušta*qqâtihâ = acides aminés et leurs dérivés »

Nous avons recensé dans notre corpus ces UTC, formées par expansion d'identification et expansion de coordination et nous l'indiquons dans les fiches terminologiques (cf. Annexe 3 : Fiches terminologiques).

6.2.2.1.4.2 Expansion d'annexion et expansion d'identification

Ce patron syntaxique est formé d'une base nominale, correspondant à une UTS, et d'une expansion annective, composée d'un syntagme nominal, constitué d'une base nominale, pouvant être une UTS ou une UL, et d'une expansion adjectivale, correspondant à une UTS :

- UTS (N) + UTS (Ann) + UTS (Adj) : « تحضير الأحماض الآمينية = *taḥḍîr al'aḥmâḍ al'am*înyyat = préparation des acides aminés »
- UTS (N) + UL (Ann) + UTS (Adj) : « إرجاع الوظيفة الحمضية = irjâʿ alwaẓîfat al*ḥ*am*ḍ*iyyat = réduction de la fonction acide »

À ces patrons syntaxiques, nous précisons la classification des UTS :

- Terme (Physique) + Terme (Physique)Ann + Terme (Physique)Adj + Terme (Physique)Adj : « طيف الرنين النووي المغناطيسي = *ṭ*ayf alranîn alnawawî almag*n*â*ṭ*îsî = spectre de résonance magnétique nucléaire »

Nous avons recensé dans notre corpus ces UTC, formées par expansion d'annexion et d'identification et nous l'indiquons dans les fiches terminologiques cf. Annexe 3 : Fiches terminologiques).

6.2.2.1.4.3 Déclencheur en base annective

Certains déclencheurs, appartenant à une UTC, interviennent comme une base d'annexion ; ils sont considérés comme des bases annectives, pouvant être nominales, adjectivales ou prépositionnelles, et constituent, avec leur expansion annective, un syntagme adjectival. De ce fait, ce patron syntaxique est un syntagme adjectival, formé d'une base nominale, correspondant à une UTS, et d'une expansion adjectivale, constituée d'une base,

correspondant à un déclencheur en base annective, et d'une expansion, pouvant être une UTS ou un UTC :

- UTS (N) + Déclencheur + UL (N) + UL (Adj) : « مركبات ذات سلاسل مستقيمة = murakkabât *ḏât salâsil mustaqîmat* = composés possédant des chaînes linéaires »

À ce patron syntaxique, nous précisons la classification des UTS :

- Terme (Physique) + Déclencheur + UL : « إشعاع تحت الحمراء = *'išʿâʿ taḥta alḥamrâ'* = rayonnement infrarouge »

Nous avons recensé dans notre corpus ces UTC contenant des déclencheurs en base annective et nous l'indiquons dans les fiches terminologiques (cf. Annexe 3 : Fiches terminologiques).

6.2.2.2 Patron morphosyntaxique adverbial

Ce patron morphosyntaxique adverbial est formé d'une expansion complétive (cf. 2.1.3.22.1.3.2 Formation des unités terminologiques complexes) ; il est composé d'un syntagme adverbial formé d'une base, correspondant à une UTS, d'une préposition et d'une expansion, pouvant être une UTS. Cette expansion compte un patron syntaxique :

- UTS + prép + UTS : « قطرة بقطرة = *qaṭrat biqaṭ*rat = goutte à goutte »

À ce patron syntaxique, nous précisons la classification des UTS :

- Terme (Composé chimique) + prép + Terme (Composé chimique) :« قطرة بقطرة = *qaṭrat biqaṭrat* = goutte à goutte »

Nous avons recensé dans notre corpus cette UTC, formée par expansion complétive et nous l'indiquons dans les fiches terminologiques (cf. Annexe 3 : Fiches terminologiques).

6.2.2.3 Cas particulier

En chimie, nous observons de nouvelles constructions grammaticales pour la dénomination des composés chimiques (cf. 2.1.3.2.3 Cas particulier).

Ces UTC, issues de la nomenclature systématique, proposent des constructions en rupture avec l'arabe classique ; il s'agit toujours d'une structure binaire, composée par deux éléments : d'une part, l'UTC est constituée d'une base et d'une expansion, les deux comportant autant une UTS qu'une multiplicité d'UTS, et d'autre part, l'UTC est formée d'une UL et d'une base, comportant autant une UTS qu'une multiplicité d'UTS (Albeiriss, 2016)[412] :

[412] Étude Terminologique de la Nomenclature de la Chimie en Arabe dans une Approche de Fouille de Textes

- UTS (N) + UTC : « حمض –2 أمينو –3 هيدروكسي بوتان ثنائي أويك = acide 2-amino-3-hydroxybutanedioïque »
- UTC + UTS (N) : « هيدروكلوريد إيثيل إستر الغليسين = chlorhydrate d'ester éthylique de glycine »
- UL + UTS (N) : « ثنائي كلورو ميثان = *ṯunâ'î klûrûmîṯân* = dichlorométhane »
- UL + UTC : « ثلاثي إيثيل أمين = *ṯulâṯî 'îṯîl* 'amîn = triéthylamine »

À ces patrons syntaxiques, nous précisons la classification des UTS :

- UL + Terme (Composé chimique) : « مفروق هكسادسنال = mafrûq haksâdasanâl = trans hexadécénal »

Nous avons recensé dans notre corpus ces UTC, formées par de nouvelles constructions et nous l'indiquons dans les fiches terminologiques (cf. Annexe 3 : Fiches terminologiques).

6.2.2.4 Patron syntaxique verbal

Ces patrons syntaxiques verbaux sont de deux types : soit des verbes sans préposition, soit des verbes avec préposition.

Rappelons que pour les patrons syntaxiques verbaux, nous nous intéressons à la structure argumentale du verbe ; habituellement, l'analyse de cette structure argumentale est réalisée sur les verbes à la forme active. Pour notre travail de recherche, nous analysons autant les verbes à la forme active qu'à la forme passive.

6.2.2.4.1 Verbe sans préposition

Ce patron syntaxique est un syntagme verbal, formé d'une base verbale, correspondant à une UTS, suivie d'un argument verbal (sujet), pouvant correspondre à une UTS ou une UTC, et/ou suivi d'un argument verbal (complément), pouvant correspondre à une UTS ou une UTC. Ce syntagme verbal compte deux patrons syntaxiques :

- UTS (V) + UTS (N) : « يرشح المحلول = yura*šš*a*ḥ* alma*ḥ*lûl = la solution est filtrée »
- UTS (V) + UTC (N) : « يسخن أنبوب الاختبار = yusa*ḫḫ*an 'unbûb ali*ḫ*tibâr = le tube à essai est chauffé »

À ces patrons syntaxiques, nous précisons la classification des UTS :

- UTS (Verbe énonciateur) + UTC (Composé chimique) [UL + UTC (Constituant chimique) [UTS (Réaction chimique) + UTS (Composé chimique)ajd]] : « يتشكل ناتج

التفاعل الألدولي = ya*tašakkal* nâtij altafâʿul al*ʾaldûlî* = il se forme le produit de la réaction aldolique »

Nous avons recensé dans notre corpus ces verbes sans préposition et nous l'indiquons dans les fiches terminologiques (cf. Annexe 3 : Fiches terminologiques).

6.2.2.4.2 Verbe avec préposition

Ce patron syntaxique est un syntagme verbal, formé d'une base verbale, correspondant à une UTS, suivie d'un éventuel argument verbal (sujet), correspondant à une UL, une UTS, ou une UTC, d'une préposition, d'un second argument verbal (complément), pouvant être une UL, une UTS ou une UTC et éventuellement d'une seconde préposition avec un troisième argument verbal (circonstant), pouvant être une UL, une UTS ou une UTC. Ce syntagme verbal compte six patrons syntaxiques :

- UTS (V) + prép + UTS (N) : « يتفاعل مع الحمض = yatafâ`al maʿ al*ḥ*am*ḍ* = il réagit avec l'acide »
- UTS (V) + prép + UTC (N) : « تجفف فوق كبريتات الصوديوم = tujaffafu fawqa kibrîtât *alṣûdyûm* = il est séché sur du sulfate de sodium »
- UTS (V) + UTC (N) + prép + UTC (N) : « يقطر ثنائي كلوروميثان فوق هيدرور الكالسيوم = yuqa*ṭṭ*ar *ṯunâʾî klûrûmîṯân* fawqa hidrûr alkalsiyum = le dichlorométhane est filtré sur de l'hydrure de calcium »
- UTS (V) + UTS (N) + prép + UTC (N) : « نسكب الناتج في دورق أحادي العنق = naskub alnâtij fî dawraq *ʾu*ḥâdî alʿunq = nous versons le produit dans un erlenmeyer monocol »
- UTS (V) + UTC (N) + prép + UTS (N) + prép + UL (N) : « يقطر كلورور الأسيل فوق الكينولين بنسبة 1:10 = yuqa*ṭṭ*ar klûrûr al*ʾas*îl fawqa alkînûlîn binasbat 1:10 = le chlorure d'acyle est séché sur de la quinoléine dans les proportions 1:10 »
- UTS (V) + UTS (N) + prép + UTC (N) + prép + UL (N) + UL (N) : « يترك المزيج تحت التحريك المغناطيسي لمدة نصف ساعة = yutrak almazîj *taḥta* al*taḥrîk* almaġnâṭîsî li mudat niṣf *saʾat* = le mélange est laissé sous agitation magnétique pendant une demi-heure »

À ces patrons syntaxiques, nous précisons la classification des UTS :

- Terme (Verbe polysémique) + Terme (Composé chimique) [UL + Terme (Composé chimique) + prép + Terme (Composé chimique) [Terme (Constituant chimique) + Terme (Élément chimique)]] : « يقطر ثنائي كلوروميثان فوق هيدرور الكالسيوم = yuqa*ṭṭ*ar *ṯunâʾî* klûrû*mîṯân* fawqa hidrûr alkalsiyum = le dichlorométhane est filtré sur de l'hydrure de calcium »

Nous avons recensé dans notre corpus ces verbes avec préposition et nous l'indiquons dans les fiches terminologiques (cf. Annexe 3 : Fiches terminologiques).

6.2.3 Règle de grammaire

Une règle de grammaire est l'écriture en langage informatique du patron syntaxique, traduisant linéairement sa construction syntaxique. Cela signifie que la base du patron syntaxique est le premier mot traité : appelé déclencheur ou amorceur, il est généralement un terme ; puis, l'expansion du patron syntaxique est traitée : cette expansion peut être composée d'un ou plusieurs mots, chaque mot est alors traité successivement et ces mots sont généralement des termes. L'ensemble de ces règles de grammaire forme une grammaire d'identification, correspondant à la modélisation en langage informatique des patrons syntaxiques, des modèles de construction des unités terminologiques prenant en compte leurs structures internes, leurs usages et leurs environnements textuels.

À partir des patrons morphosyntaxiques des UTC de la chimie (cf. 6.2.2 Patron morphosyntaxique), nous établissons les règles de grammaire, permettant d'extraire ces patrons morphosyntaxiques, en précisant la classe de chaque élément de la règle de grammaire.

6.2.3.1 Patron morphosyntaxique nominal

Ces patrons morphosyntaxiques nominaux correspondent à l'expansion d'annexion, l'expansion complétive et l'expansion d'identification ainsi que la combinaison de ces expansions (cf. 6.2.2.16.2.2.1 Patron morphosyntaxique 2.1.3.2nominal) (Albeiriss, 2017)[413].

6.2.3.1.1 Expansion d'identification

Ce patron morphosyntaxique d'une expansion d'identification présente une structure binaire, composée d'une base, pouvant être une UL, une UTS ou une UTC, suivie de son expansion adjectivale, pouvant être une UL, une UTS ou une UTC. Les règles de grammaire de ce patron morphosyntaxique sont de deux types et varient en fonction des classes des UTS :

- UTS (N) + UTS (Adj) → UTC (Adj)

Terme (Réaction chimique) + Terme (Composé Chimique)Adj → Terme (Réaction chimique Physique)Adj : « تكـاثف الدولي = takâ*ṯuf* 'aldûlî = condensation aldolique »

- UL (N) + UTS (Adj)→ UTC (Adj)

[413] Modélisation des unités lexicales complexes d'une langue spécialisée : le cas de la chimie en arabe

UL + Terme (Composé Chimique)Adj → Terme (Composé Chimique)adj : « وظيفة حمضية = waẓîfat ḥamḍiyyat = fonction acide »

Ces règles de grammaire permettent d'identifier toutes les UTC formées par expansion d'identification de notre corpus, mais également de reconnaitre de nouvelles UTC formées par expansion d'identification dans d'autres corpus (cf. Annexe 5 : Règles de grammaire).

6.2.3.1.2 Expansion d'annexion

Ce patron morphosyntaxique d'une expansion d'annexion présente une structure binaire, composée d'une base, pouvant être une UL, une UTS ou une UTC, suivie de son expansion nominale, pouvant être une UL, une UTS ou une UTC. Les règles de grammaire de ce patron morphosyntaxique sont de deux types et varient en fonction des classes des UTS :

- UTS (N) + UTS (Ann) → UTC (Ann)

Terme (Composé chimique) + Terme (Elément chimique)Ann → Terme (Composé chimique)Ann : « كبريتات الصوديوم = *kibrîtât alṣûdyûm* = sulfate de sodium »

- UL (N) + UTS (Ann)→ UTC (Ann)

UL + Terme (Composé chimique)Ann → Terme (Composé chimique)Ann : « مجموعة الأمين = majmûʿat al ʾamîn = groupe amine »

Ces règles de grammaire permettent d'identifier toutes les UTC formées par expansion d'annexion de notre corpus, mais également de reconnaitre de nouvelles UTC formées par expansion d'annexion dans d'autres corpus (cf. Annexe 5 : Règles de grammaire).

6.2.3.1.3 Expansion complétive

Ce patron morphosyntaxique d'une expansion complétive présente une structure binaire, composée d'une base, pouvant être une UL, une UTS ou une UTC, d'une préposition, suivie de son expansion nominale, pouvant être une UL, une UTS ou une UTC. Les règles de grammaire de ce patron morphosyntaxique sont de deux types et varient en fonction des classes des UTS :

- UL (N) + prép + UTS (N)→ UTC (N)

UL + préposition + Terme (Composé Chimique) → Terme (Procédure chimique) : « جو من الأزوت = jaww *min al'azût* = atmosphère d'azote »

Ces règles de grammaire permettent d'identifier toutes les UTC formées par expansion complétive de notre corpus, mais également de reconnaitre de nouvelles UTC formées par expansion complétive dans d'autres corpus (cf. Annexe 5 : Règles de grammaire).

6.2.3.1.4 Combinaison des patrons morphosyntaxiques

Certaines UTC sont formées d'une base constituée elle-même d'une UTC ou d'une expansion elle-même composée d'une UTC ou les deux.

6.2.3.1.4.1 Expansion d'identification et expansion de coordination

Ce patron morphosyntaxique de cette combinaison d'expansion d'identification et de coordination présente une structure binaire, composée d'un syntagme nominal formé d'une base nominale, correspondant à une UTS, et d'une expansion adjectivale pouvant être une UTS, suivi d'un coordonnant et d'une expansion, pouvant correspondre à une UTS. À cela, il faut ajouter la possibilité d'avoir une UL à la place d'une UTS. La règle de grammaire de ce patron morphosyntaxique est la suivante et varie en fonction des classes des UTS :

- UTC (N) + UTS (Adj) → UTC (Adj)

Terme (Composé chimique) + Terme (Propriété chimique) + Terme (Propriété chimique) → Terme (Composé Chimique)Adj : « ألدهيدات أليفاتيـة و عطرية = *'aldahîd*ât *'alîf*âtiyyat wa *ʿiṭrî*yyat = aldéhydes aliphatiques et aromatiques »

Ces règles de grammaire permettent d'identifier toutes les UTC formées par expansion d'identification et de coordination de notre corpus, mais également de reconnaitre de nouvelles UTC formées par expansion d'identification et de coordination dans d'autres corpus (cf. Annexe 5 : Règles de grammaire).

6.2.3.1.4.2 Expansion d'annexion et expansion d'identification

Ce patron morphosyntaxique de cette combinaison d'expansion d'annexion et d'identification présente une structure binaire, composée d'une base nominale, correspondant à une UTS, et d'une expansion annective composée d'un syntagme nominal, constitué d'une base nominale pouvant être une UTS ou une UL, et d'une expansion adjectivale correspondant à une UTS. Les règles de grammaire de ce patron morphosyntaxique sont de trois types et varient en fonction des classes des UTS :

- UTS (N) + UL (N) + UTS (Adj) → UTC (N)

Terme (Réaction chimique) + UL + Terme (Composé chimique)Adj → Terme (Composé chimique)Ann « إرجاع الوظيفة الحمضية = irjâʿ alwaẓîfat al*ḥ*am*ḍ*iyyat = réduction de la fonction acide »

- UTS (N) + UTC (Ann) → UTC (Ann)

Terme (Procédure chimique) + Terme (Composé chimique)Ann + Terme (Composé chimique)Adj → Terme (Composé chimique)Ann : « تحضير الأحماض الآمينية = *taḥḍîr al'aḥmâḍ al'am*înyat = préparation des acides aminés »

- UL + UTC (N)→ UTC (N)

UL + Terme (Physique)Ann + Terme (Physique)Adj + Terme (Physique)Adj → Terme (Méthode d'analyse)Ann : « طيف الرنين النووي المغناطيسي = *ṭ*ayf alranîn alnawawî almag*n*â*ṭî*sî = spectre de résonance magnétique nucléaire »

Ces règles de grammaire permettent d'identifier toutes les UTC formées par expansion d'annexion et d'identification de notre corpus, mais également de reconnaitre de nouvelles UTC formées par expansion d'annexion et d'identification dans d'autres corpus (cf. Annexe 5 : Règles de grammaire).

6.2.3.1.4.3 Déclencheur en base annective

Ce patron morphosyntaxique contenant un déclencheur en base annective est un syntagme adjectival, formé d'une base nominale, pouvant correspondre à une UTS, et d'une expansion adjectivale, constituée d'une base, correspondant à un déclencheur en base annective, et d'une expansion, pouvant être une UTS ou une UTC. Les règles de grammaire de ce patron morphosyntaxique sont de deux types et varient en fonction des classes des UTS :

- UTS (N) + prép + UTS (Adj) → UTC (N)

Terme (Physique) + prép + UL (Adj) → Terme (Physique) : « إشعاع تحت الحمراء = *'iš*ʿâʿ *taḥta alḥamrâ'* = rayonnement infrarouge »

- UTS (N) + prép + UTS (Adj) + UTS (Adj) → UTC (N)

Terme (Composé chimique) + prép + UL (Adj) + UL (Adj) → Terme (Composé chimique)Adj : « مركبات ذات سلاسل مستقيمة = murakkabât *ḏâ*t salâsil mustaqîmat = composés possédant des chaînes linéaires »

Ces règles de grammaire permettent d'identifier toutes les UTC contenant des déclencheurs en base annective, mais également de reconnaitre de nouvelles UTC contenant des déclencheurs en base annective dans d'autres corpus (cf. Annexe 5 : Règles de grammaire).

6.2.3.2 Patron morphosyntaxique adverbial

Ce patron morphosyntaxique adverbial est formé d'une expansion complétive (cf. 6.2.2.2 Patron morphosyntaxique adverbial 2.1.3.2nominal).

- UTS (N) + prép + UTS (N) → UTC (N)

Terme (Composé Chimique) + préposition + Terme (Composé Chimique) → Terme (Procédure chimique) : « قطرة بقطرة = *qaṭrat biqaṭrat* = goutte à goutte »

6.2.3.3 Cas particulier

Les règles de grammaire des patrons morphosyntaxiques des cas particuliers correspondent à de nouvelles constructions grammaticales pour la dénomination des composés chimiques (cf. 2.1.3.2.32.1.3.2.3 Cas particulier).

Le patron morphosyntaxique de ces nouvelles constructions présente une structure binaire, composée d'une part, d'une base et d'une expansion, les deux comportant autant une UTS qu'une multiplicité d'UTS ou d'UTC, et d'autre part, d'une UL et d'une base, comportant autant une UTS qu'une multiplicité d'UTS ou d'UTC. Les règles de grammaire de ce patron morphosyntaxique sont de quatre types et varient en fonction des classes des UTS (Albeiriss, 2016)[414] :

- UL + UTS (N) → UTC (N)

UL + Terme (Composé chimique) → Terme (Composé chimique) : « ثنائي كلوروميثان = *ṯunâ'î klûrûmîṯân* = dichlorométhane »

- UTS (N) + UTC (N) → UTC (N)

Terme (Composé chimique) + Terme (Formant chimique) + Terme (Formant chimique) + Terme (Composé chimique) + UL + Terme (Suffixe chimique) → Terme (Composé chimique)(Ann) : « حمض 2– أمينو 3– هيدروكسي بوتان ثنائي أويك = acide 2-amino-3-hydroxybutanedioïque »

- UTC (N) + UTS (N) → UTC (N)

Terme (Composé chimique) + Terme (Composé chimique) + Terme (Composé chimique) + Terme (Composé chimique)Ann → Terme (Composé chimique)(Ann) : « هيدروكلوريد إيثيل إستر الغليسين = chlorhydrate d'ester éthylique de glycine »

- UL + UTC (N)→ UTC (N)

UL + Terme (Composé chimique) + Terme (Composé chimique) → Terme (Composé chimique) : « ثلاثي إيثيل أمين = *ṯulâṯî 'îṯîl* 'amîn = triéthylamine »

[414] Étude Terminologique de la Nomenclature de la Chimie en Arabe dans une Approche de Fouille de Textes

Ces règles de grammaire permettent d'identifier toutes les UTC formées par de nouvelles constructions de notre corpus, mais également de reconnaitre de nouvelles UTC formées par de nouvelles constructions dans d'autres corpus (cf. Annexe 5 : Règles de grammaire).

6.2.3.4 Patron syntaxique verbal

Ces patrons syntaxiques verbaux sont de deux types : soit des verbes sans préposition, soit des verbes avec préposition (cf. 6.2.2.46.2.2.1 Patron syntaxique verbal2.1.3.2).

6.2.3.4.1 Verbe sans préposition

Ce patron syntaxique est un syntagme verbal, formé d'une base verbale correspondant à une UTS, suivie d'un argument verbal (sujet), pouvant correspondre à une UTS ou une UTC, et/ou suivie d'un argument verbal (complément), pouvant correspondre à une UTS ou une UTC. Les règles de grammaire de ce patron morphosyntaxique sont de trois types et varient en fonction des classes des UTS :

- UTS (V) + UTS (N) → UTC (V)

Terme (Verbe spécialisé) + Terme (Composé Chimique) → Terme (Verbe spécialisé) : « يرشح المحلول = yuraša*ḥ* alma*ḥ*lûl = la solution est filtrée »

- UTS (V) + UTC (Ann) → UTC (V)

Terme (Verbe énonciateur) + UL + Terme (Composé chimique)Ann → Terme (Verbe énonciateur) : « يسخن أنبوب الاختبار = yusa*ḫḫ*an 'unbûb ali*ḫ*tibâr = le tube à essai est chauffé »

- UTS (V) + UTC (N) → UTC (V)

Terme (Verbe énonciateur) + Terme (Composé chimique) + Terme (Réaction chimique)Ann + Terme (Composé chimique)Adj → Terme (Verbe énonciateur) : « يتشكل ناتج التفاعل الألدولي = ya*tašakkal* nâtij altafâ'ul al*'aldûlî* = il se forme le produit de la réaction aldolique »

Ces règles de grammaire permettent d'identifier tous les termes verbaux de notre corpus, mais également de reconnaitre de nouveaux termes verbaux dans d'autres corpus (cf. Annexe 5 : Règles de grammaire).

6.2.3.4.2 Verbe avec préposition

Ce patron syntaxique est un syntagme verbal, formé d'une base verbale correspondant à une UTS, suivie d'un éventuel argument verbal (sujet), correspondant à une UTS ou une UTC, d'une préposition, d'un second argument verbal (complément), pouvant être une UTS ou une UTC et éventuellement d'une seconde préposition avec un troisième argument verbal

(complément), pouvant être une UTS ou une UTC. Les règles de grammaire de ce patron morphosyntaxique sont de six types et varient en fonction des classes des UTS :

- UTS (V) + prép + UTS (N) → UTC (V)

Terme (Verbe spécialisé) + prép + Terme (Composé Chimique) → Terme (Verbe spécialisé) : « يتفاعل مع الحمض = yatafâ`l maʿ alḥamḍ = il réagit avec l'acide »

- UTS (V) + prép + UTC (N) → UTC (V)

Terme (Verbe polysémique) + prép + Terme (Composé Chimique) + Terme (Élément Chimique)Ann → Terme (Verbe polysémique) : « تجفف فوق كبريتات الصوديوم = tujaffafu fawqa ki*brîtât alṣûdyûm* = il est séché sur du sulfate de sodium »

- UTS (V) + UTC (N) + prép + UTC (N) → UTC (V)

Terme (Verbe spécialisé) + UL + Terme (Composé chimique) + prép + Terme (Composé Chimique) + Terme (Élément Chimique)Ann → Terme (Verbe spécialisé) : « يقطر ثنائي كلوروميثان فوق هيدرور الكالسيوم = yuqa*ṭṭ*ar *ṯunâ'î klûrûmîṯân* fawqa hidrûr alkalsiyum = le dichlorométhane est filtré sur de l'hydrure de calcium »

- UTS (V) + UTS (N) + prép + UTC (N) → UTC (V)

Terme (Verbe énonciateur) + Terme (Composé chimique) + prép + Terme (Verrerie) + déclencheur + Terme (Verrerie) → Terme (Verbe énonciateur) : « نسكب الناتج في دورق أحادي العنق = naskub alnâtij fî dawraq *'u*ḥâdî alʿunq = nous versons le produit dans l'erlenmeyer monocol »

- UTS (V) + UTC (N) + prép + UTS (N) + prép + UTC (N) → UTC (V)

Terme (Verbe spécialisé) + Terme (Composé chimique) + Terme (Composé Chimique)Ann + prép + Terme (Composé Chimique) + prép + UL → Terme (Verbe spécialisé) : « يقطر كلورور الأسيل فوق الكينولين بنسبة 1:10= yuqa*ṭṭ*ar klûrûr al*'as*îl fawqa alkînûlîn binasbat 1 :10 = le chlorure d'acyle est séché sur de la quinoléine dans les proportions 1/10 »

- UTS (V) + UTS (N) + prép + UTC (N) + prép + UTC (N) → UTC (V)

Terme (Verbe énonciateur) + Terme (Composé chimique) + prép + Terme (Procédure chimique) + Terme (Physique) + prép + UL + UL → Terme (Verbe énonciateur) : « يترك المزيج تحت التحريك المغناطيسي لمدة نصف ساعة = yutrak almazîj *taḥta* al*taḥrîk* almaġn*âṭî*sî li mudat niṣf *sa'at* = le mélange est laissé sous agitation magnétique pendant une demi-heure »

Ces règles de grammaire permettent d'identifier tous les termes verbaux de notre corpus, mais également de reconnaitre de nouveaux termes verbaux dans d'autres corpus (cf. Annexe 5 : Règles de grammaire).

6.3 Quelques pistes du traitement informatique

Nous proposons une immersion dans le traitement informatique pour l'extraction des termes de la chimie en arabe. Après avoir rappelé brièvement cette notion, notamment les éléments essentiels, nous exposons les choix des traitements en fonction de notre recherche, précisant nos remarques et nos observations. Puis, nous présentons une ébauche du traitement informatique.

6.3.1 Traitement informatique

Le traitement informatique établit une correspondance qui permet sans ambiguïté de passer de la représentation graphique d'une information (nombres, textes…) à une autre représentation sous forme binaire (suite de 0 et 1) de la même information suivant un ensemble de règles précises, formant un algorithme.

De la même manière que les nombres et les caractères se sont vus attribuer des représentations sous forme binaire, comme le code ASCII, les mots sont codés afin de décomposer le problème de l'analyse syntaxique d'une phrase en sous-problèmes élémentaires. De façon simplifiée, la résolution de ce problème s'effectue par un enchaînement en cascade d'une suite de mots qui sont identifiés comme une relation morphosyntaxique, impliquant un ordre linéaire dans l'analyse. Les relations morphosyntaxiques que nous avons développées « à la main » mettent en œuvre le recours à la connaissance grammaticale et sémantique, à partir de notre corpus de travail.

L'idée de ce traitement informatique est de mettre les données entre balises de façon à ce que le contenu de l'information soit connu ; il ne s'agit plus d'un texte brut mais d'un texte balisé. Ces langages permettent de décrire des termes ainsi que les relations entre ces termes ; plus ces langages sont de haut niveau, plus leur pouvoir d'expression est important. Nous partons de données peu formalisées en introduisant petit à petit des contraintes de plus en plus fortes, permettant une exploitation de plus en plus efficace des données (Jazzar, 2009)[415].

[415] L'ontologie de l'économie pétrolière en Arabie Saoudite et analyse terminologique anglais-français-arabe

6.3.1.1 Balisage

Première étape du traitement informatique et la plus simple, le balisage des données a pour objectif de structurer le contenu d'un document, à l'aide du langage HTML ; ces balises sont apparues dès les origines du web et s'écrivent de la manière suivante :

<HTML>

.....

<BODY>

.....

</BODY>

</HTML>

6.3.1.2 Langage XML

Le langage XML permet de baliser plus précisément la connaissance exploitable contenue entre les balises du corps du document, c'est-à-dire, </BODY>.... </BODY> ; le choix des balises revient à son auteur en fonction du domaine de connaissance, mais il faudra bien sûr vérifier que le fichier soit bien écrit.

En effet, les choses deviennent rapidement plus complexes, notamment lorsque plus de deux éléments sont impliqués dans l'analyse. Mais le langage XML est très bien doté en langage de manipulation et d'interrogation des données structurées, dans la recherche des termes de la chimie comme : //NP/_[lemma="ذرة"]. Ce traitement exprime la dominance, comme dans les structures hiérarchiques des chemins de fichiers. Quoi qu'il en soit, le panorama des outils disponibles était amplement suffisant pour ce type de données pour les banques d'arbres syntaxiques (Tanguy, 2012)[416].

6.3.1.3 Contrainte

La représentation des connaissances suppose une analyse morphosyntaxique à laquelle subsistent deux difficultés majeures. La première repose sur les connaissances implicites nécessaires à la compréhension d'un texte et la seconde est liée au fait que l'analyse d'un énoncé correspond à la forme linguistique (Toussaint, 2011)[417].

[416] Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes, p. 69

[417] Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances, p. 6

En effet, les langues comme l'anglais possèdent, pour la plupart des mots, une ou deux étiquettes possibles, et quand il y a ambiguïté, le système d'étiquetage peut généralement faire un choix en se basant sur les étiquettes du contexte immédiat, juste avant ou juste après le mot traité. Pour l'arabe, du fait de sa grande complexité morphologique, et de son système alphabétique consonantique, la difficulté est incomparable, car les ambiguïtés sont bien plus fréquentes et le jeu d'étiquettes peut se révéler beaucoup plus vaste, posant différents problèmes pour le traitement informatique :

> *« Les défis de base sont la directionnalité de droite à gauche, les formes de lettres variantes contextuelles, les ligatures, l'utilisation de signes diacritiques et la manipulation bidirectionnelle de chiffres et de caractères romains dans des contextes arabes ». (Habash, 2010[418])*

Rappelons que la langue arabe s'écrit de droite à gauche et ses lettres changent de forme de présentation selon leur position. Les consonnes arabes sont accompagnées dans un mot par des voyelles et selon ces dernières, l'interprétation peut varier. La non-voyellation (4.1.4.2.3.1 Ambiguïté et voyellation) génère une ambigüité aussi bien sur son sens que sur sa fonction dans la phrase ; il existe de nombreuses analyses morphologiques possibles à cause des multiples voyellations alternatives. Composée de trois catégories de mots, les verbes, les noms et les particules, la langue arabe permet de juxtaposer les formes entre elles, ne constituant plus qu'un seul mot (Dichy, 1997)[419] ; l'agglutination (4.1.4.2.3.2 Ambiguïté et agglutination) engendre une ambiguïté aussi bien de sens que de fonction dans la phrase, impliquant une ambiguïté notoire dans un texte. Quant à sa morphologie, l'arabe est une langue à morphologie flexionnelle riche et complexe (Dichy, 1990)[420] ; environ 333 000 combinaisons de traits morphologiques possibles existent pour la langue arabe, tandis qu'en anglais on peut trouver une cinquantaine de combinaisons de traits au maximum (Ghoul, 2011)[421]. Ce système syntaxique, morphologique et sémantique, est assez éloigné de celui des langues européennes et a une tradition grammaticale originale et séculaire, ce qui rend difficile l'adaptation du système informatique, conçu initialement par des développeurs occidentaux.

Nous avons fait le choix initial de la table rase en commençant sans aucune information, c'est-à-dire, sans développer au préalable un lexique syntaxique très riche recensant les propriétés syntaxiques des mots de la langue. Par conséquent, nous pouvons bénéficier indirectement des

[418] Introduction to Arabic Natural Language Processing

[419] Pour une lexicomatique de l'arabe : L'unité simple et l'inventaire fini des spécificateurs du domaine du mot

[420] L'écriture dans la représentation de la langue : la lettre et le mot en arabe

[421] Outils génériques pour l'étiquetage morphosyntaxique de la langue arabe : segmentation et corpus d'entraînement

ressources lexicales éventuellement exploitées par des analyseurs et/ou des étiqueteurs (Bourigault et al., 2005)[422]. Des informations lexicales sont intégrées au fur et à mesure des besoins, comme la liste des affixes et des clitiques, la liste des pré-bases et des post-bases, mais également les tables comportant les différents cas de transitivités, les aspects, les pronoms, les modes, les cas... ainsi que les tables des racines et des schèmes (Ghoul, 2011)[423]. Ces informations, codées sur les entrées lexicales, sont d'ordre catégoriel et flexionnel, ce qui permet de relier formes fléchies, lemmes et informations morphosyntaxiques et morphosémantiques, de sorte que l'utilisateur peut accéder à la structure des unités lexicales construites, mais également aux contraintes sémantiques imposées par le procédé de construction lexicale en jeu. (Namer, 2003)[424].

6.3.2 Application du traitement informatique

Dans un premier temps, nous définissons les groupes des termes de la chimie en fonction de leur catégorie grammaticale (nom / verbe / adjectif) et nous écrivons les règles de la grammaire avec, si besoin, des contraintes. Puis, nous présentons l'analyse que nous souhaitons obtenir pour les informations recherchées. Enfin, nous relevons quelques remarques sur le traitement informatique.

6.3.2.1 Écriture du traitement informatique

Pour la langue arabe, nous choisissons l'encodage UTF-8 :

<?xml version="1.0" encoding="UTF-8"?>

Un programme est composé d'un début et d'une fin ; il ne faut pas oublier de finir le programme, sinon ce dernier ne fonctionnera pas. Des commentaires peuvent être ajoutés entre parenthèses afin de préciser le travail effectué :

<rule ……>

(commentaire)

</rule>

Les classes répertoriées pour le domaine de la chimie sont définies dans le programme sous la forme de groupe, par exemple l'UTC « ذرة الكربون = *ḏ*arrat alkarbûn = atome de carbone :

[422] Syntex, analyseur syntaxique de corpus

[423] Outils génériques pour l'étiquetage morphosyntaxique de la langue arabe : segmentation et corpus d'entraînemen

[424] Le modèle Lstat : ou comment se constituer une base de données morphologique à partir du Web, p. 87

<groups group="cChemicalConstituant" lexCat="C_NOUN">

</groups>

<groups group="cChemicalElement" lexCat="C_NOUN">

</groups>

Des contraintes peuvent être indiquées afin d'extraire l'information recherchée, portant sur les traits suivants :

- Le genre : masculin et féminin
- Le nombre : singulier, duel et pluriel
- Le cas : nominatif, accusatif et génitif
- Le mode : déterminé et indéterminé

Par exemple :

<groups group="cChemicalConstituant" lexCat="C_NOUN">

<lexTrait mode="Undetermined">

<lexTrait mode="Determined">

</groups>

<groups group="cChemicalElementAnn" lexCat="C_NOUN">

<lexTrait mode="Annexion">

</groups>

À présent, les règles de grammaire sont écrites avec, si besoin, les contraintes, et indiquent les mots de gauche (lhs=left hand side) et de droite (rhs=right hand side) en tenant compte que la langue arabe s'écrit de droite à gauche et en reprenant les groupes établis :

<rule name="TermChemistry" lhs="cChemicalElementAnn" rhs="cChemicalConstituant" result="gChemicalConstituant" finit="true">

<constraint group="cNoun" mode="Determined"/>

<constraint group="cNoun" case="Genitive"/>

</rule>

La règle représente l'UTC ذرة الكربون où le lhs est un mudâf (cChemicalElementAnn), c'est-à-dire, que le mode Annexion se trouve dans la définition du groupe et le rhs possède une contrainte, c'est-à-dire, un nom déterminé au génitif ; ceci explique la contrainte dans la règle.

Ces contraintes permettent de désambiguïser le contexte afin d'obtenir l'analyse la plus correcte.

6.3.2.2 Analyse du traitement informatique

Le texte est analysé linéairement, phrase par phrase, syntagme par syntagme, mot par mot, morphème par morphème, en vérifiant s'il s'agit d'un terme de chimie et s'il existe une relation morphosyntaxique entre ces termes, puis en appliquant les règles de grammaire que nous avons construites. Cette analyse est représentée sous la forme d'un arbre morphosyntaxique, exprimé sous la forme de constituants imbriqués. Il ne s'agit pas d'une invention puisque la quasi-totalité des corpus annotés syntaxiquement par des analyseurs sont des corpus arborés ou treebanks (Marcus et al., 1993). Ce format présente l'avantage majeur de comprendre l'ensemble des informations disponibles pour chaque mot (catégorie, lemme) en plus de la fonction de chaque constituant.

Par exemple, voici (à peu près) à quoi ressemblerait dans ce format l'analyse de l'UTC « في علم الكيمياء = fî ʿilm alkîmyâ ʾ = en chimie » :

<file>

<phrase>

<pos start="0" finish="0" content="في علم الكيمياء" group="">

<interpretation>

<morphology category="C_TOOLS" group="cTools" lemma="فِي" root="xxx" string="في" form="في" formv="فِي">

<traitNoun gender="" number="" mode="" case="Genetive"/>

</morphology>

<morphology category="C_NOUN" group="cScienceAnn" lemma="عِلْم" root="علم" string="علم" form="علم" formv="عِلْمَ">

<traitNoun gender="مذكر" number="Singular" mode="Annexion" case="Accusative"/>

</morphology>

<morphology category="C_NOUN" group="cScienceAdj" lemma="كِيمِيَاء" root="xxx" string="الكيمياء" form="كيمياء" formv="كِيمِيَاءِ">

<traitNoun gender="Female" number="Singular" mode="Determined" case="Genetive"/>

<proclitic category="C_PCL_N" string="ال" formv="اَلْ">

<traitNoun gender="" number="" mode="Determined" case="Genetive"/>

</proclitic>

</morphology>

<properties category="gScienceAdj" group="gScienceAdj" lemma="" root="" string=" في علم الكيمياء" form="في علم الكيمياء" formv="فِي عِلْمِ ٱلْكِيمِيَاءِ">

<traitNoun gender="Female" number="Singular" mode="Determined" case="Genetive"/>

</properties>

</interpretation>

</pos>

<pos start="0" finish="0" content=":" group="word">

<interpretation>

<morphology category="PONCTUATION" group="PONCTUATION" lemma="" root="" string=":" form="" formv=""/>

</interpretation>

</pos>

</properties>

</interpretation>

</pos>

</phrase>

</file>[425]

6.3.2.3 Remarque

Des ambigüités ont été relevées lors de l'écriture du codage informatique ; certains points concernent les langues en général, l'arabe en particulier, mais d'autres sont spécifiques au domaine étudié, la chimie.

Ces ambigüités sont récurrentes dans le traitement de la langue ; mais des scripts plus spécifiques à l'arabe et au domaine étudié peuvent être proposés et ajoutés dans les traitements linguistiques (segmentation, lemmatisation, analyse morphosyntaxique) afin de faciliter ce travail de recherche d'information.

[425] Travail réalisé en collaboration avec Technimed

D’une part, l’analyse doit toujours commencer par les mots maximaux afin de les réduire en mots minimaux. Mais en cas d'ambiguïté ou d’absence dans les tables, l’utilisateur est mis à contribution afin d’ajouter à chaque fois le nouveau mot aux tables concernées, ce qui permet le développement de la base de données. Parfois, un mot n’a pas pu être reconnu ; cela peut être dû au fait que le mot comporte une erreur d’orthographe, ou bien que le mot n’existe pas dans les tables, ou alors que le mot est une transcription d’items étrangers. Pour notre travail, ce cas est récurrent en raison de l’emploi d’emprunts pour le domaine de la chimie.

D’autre part, il n’est pas toujours facile de faire la différence entre un proclitique et un caractère appartenant à la racine de certains mots, comme le cas du proclitique « ل = li = pour que » ou encore le cas du préfixe « ت = ta = préfixe verbal de l’inaccompli » ; ceci n’est pas sans désagrément et un découpage est particulièrement utile, voir indispensable, pour les analyses syntaxiques. Par exemple, « ل = li = pour que » a été défini comme appartenant à la liste finie des proclitiques, mais également comme étant la vingt-troisième lettre de l’alphabet arabe, pouvant être contenue dans la base. Si l’analyse n’est pas correcte, l’utilisateur corrige l’analyse et écrit une (ou des) règle(s) de grammaire en définissant des contraintes, si besoin.

En d’autres mots, ce travail doit être supervisé, nécessitant régulièrement des vérifications et des corrections, tant par la complexité de la langue que par sa variation.

En conclusion de ce chapitre, nous avons présenté une méthode pour la modélisation d’un extracteur morphosyntaxique des termes de la chimie en arabe de notre corpus. Nous avons conçu cette méthode de telle sorte que l’on puisse reconnaître des termes dans d’autres corpus. Cette approche par patrons se révèle robuste et permettrait de traiter des phrases longues et syntaxiquement complexes. Ainsi, elle deviendrait une approche incontournable dans le cas de phrases complexes pour lesquelles les risques d’erreur sont trop élevés. Cependant, développer à la main un ensemble de règles pour l’extraction d’information a certes un côté ludique, mais aussi ses propres limites. Obtenir un ensemble de règles stables et exhaustives par rapport à un objectif donné se révèle laborieux et fastidieux, avec une difficulté à garantir la cohérence de l’ensemble des règles, puisque les patrons définis sont très dépendants du domaine, des types de textes, et même de l’application visée. De ce fait, cette méthode est amenée à évoluer et être enrichie par d’autres contributions.

À ce stade du travail, une des applications possibles serait la construction d’une ontologie de la chimie en arabe, application que nous présentons dans le chapitre suivant.

Chapitre 7 : Réseau sémantique

Dans ce chapitre, nous souhaitons représenter les unités terminologiques de la chimie en arabe et leurs relations dans une ontologie. Pour cela, nous rappelons d'abord la notion de réseau sémantique, ses origines et ses applications. Puis, nous présentons la démarche de construction de l'ontologie en précisant les relations hiérarchiques et non-hiérarchiques entre les unités terminologiques de la chimie en arabe, en nous basant sur notre classification du domaine.

7.1 Réseau sémantique : définitions

Le réseau sémantique peut être considéré comme une représentation de l'ensemble des connaissances qu'un individu se construit pour un domaine spécifique, constituée de nœuds, correspondant aux éléments de connaissance, et de liens, représentant les relations sémantiques et/ou syntaxiques, afin de relier ces éléments de connaissance entre eux, comme la synonymie, la métonymie… (Habrant et al., 1999)[426]. Il est alors assimilé à une structure de graphe orienté et étiqueté, plus précisément un multigraphe (Harrathi, 2009)[427], employant différents types de nœuds et différents types de relations entre ces nœuds, représentés par des arcs étiquetés.

Ces réseaux sémantiques permettent une représentation riche et économique, à partir de notions très simples, afin de passer de l'expression linguistique des connaissances, telle que nous autres humains pouvons la considérer, à une représentation formelle et calculable des connaissances, propre à une exploitation informatique (Bachimont, 2000)[428].

7.1.1 Historique du réseau sémantique

Introduit en 1968 par Quillian, le réseau sémantique a pour origine des expériences de psychologie qui ont montré que l'homme semble mémoriser les informations selon un principe d'économie avec un modèle de mémoire associative représentant des relations entre concepts. Par la suite, en 1979, Fahlman introduit les réseaux à propagation de marqueurs pour simuler

[426] Utilisation des réseaux sémantiques pour la navigation dans l'hypertexte, p. 2

[427] Extraction de concepts et de relations entre concepts à partir des documents multilingues : Approche statistique et ontologique, p. 52

[428] Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en Ingénierie des connaissances

le fonctionnement neurobiologique d'un ensemble d'éléments, considérant les nœuds comme des microprocesseurs rudimentaires qui possèdent une mémoire locale. À la même époque, Hendrix propose les réseaux sémantiques partitionnés, offrant la possibilité de délimiter des sous-ensembles de nœuds et de relations appelés des espaces ; ces connaissances situées dans un espace sont locales et non partageables, sauf avec les espaces-fils. Enfin, en 1984, Sowa propose les graphes conceptuels, aujourd'hui largement utilisés en sciences cognitives et dans l'intelligence artificielle en particulier, considérant que toute forme de représentation pourrait être écrite sous forme de graphe conceptuel (GC), s'appuyant sur l'étude de la perception en psychologie (Desclés, 1987)[429].

Le sens d'un concept se réduit alors à sa position relative par rapport aux autres concepts. Il ne prend donc un sens que par rapport à un réseau sémantique modélisant les connaissances générales du système.

Parmi les domaines qui s'intéressent aux réseaux sémantiques, il y a les sciences cognitives et l'intelligence artificielle.

D'une part, les sciences cognitives (SC) s'intéressent aux mécanismes de l'intelligence et ont été une source majeure d'inspiration pour des systèmes de représentation des connaissances, comme nous avons pu le voir avec Quillian et sa proposition de constituer un modèle de la mémoire humaine. Des systèmes de représentation à base de règles ont aussi été proposés et sont utilisés aujourd'hui principalement dans les systèmes experts.

D'autre part, l'intelligence artificielle (IA) a été l'un des premiers domaines à chercher à représenter les connaissances et constitue toujours la base de nombreuses recherches. L'utilisation des graphes en représentation des connaissances pour l'IA vient de l'idée de représenter graphiquement des concepts et leurs liens. Par ailleurs, les graphes sont souvent utilisés pour abstraire les informations pertinentes et se concentrer seulement sur la topologie d'un problème, le graphe constituant un espace du problème.

7.1.2 Application du réseau sémantique

Le réseau sémantique offre plusieurs représentations des connaissances, notamment le thésaurus et l'ontologie.

[429] Réseaux sémantiques : la nature logique et linguistique des relateurs, p. 55

7.1.2.1 Thésaurus

Le thésaurus est un langage documentaire, rassemblant un ensemble de termes structurés, choisis pour leur capacité à décrire un domaine, et sont utilisés afin de décrire, d'une manière précise, le contenu des documents ; ils sont sélectionnés et normalisés pour l'indexation et le classement des documents (Harrathi, 2009)[430].

Base de connaissances lexicales, le thésaurus aborde les termes avec un point de vue différent des dictionnaires de langue, en se basant sur un sens plus relationnel, permettant ainsi d'identifier le bon terme et/ou de décrire une notion par des termes, offrant alors un accès par les termes (entrées), mais également un accès par les notions (Habert et al., 1997)[431].

Fondé sur une structuration hiérarchisée, le thésaurus est constitué d'un vocabulaire, correspondant à l'ensemble des termes d'une langue de spécialité, organisés suivant des relations sémantiques entre les termes du vocabulaire : relations d'équivalence, hiérarchiques et associatives (Hernandez et al., 2006)[432] ; ces termes dénotent les concepts d'un domaine de spécialité, permettant de les nommer facilement, à l'aide d'un terme descripteur, avec la collaboration des terminologues, afin d'identifier les variations en discours du concept et ne retenir qu'une forme canonique (Zacklad, 2007)[433].

WordNet est un bel exemple de thésaurus, conçu à l'origine comme une base lexicale ; mais par la suite, il a été perçu comme un réseau sémantique. Développé depuis 1985 à l'université de Princeton par une équipe de linguistes et de psycholinguistes sous la direction de G. Miller, chaque nœud représente un concept et est constitué par un ensemble de termes synonymes qui sont reliés par des relations sémantiques, comme la relation de synonymie, relation de base dans WordNet. Une tentative d'un WordNet pour l'arabe a vu le jour : AWN : ArabWord Net (Elkateb et al. 2006)[434], en se basant sur l'architecture de WordNet.

Cependant, le recours à un thésaurus soulève plusieurs problèmes, notamment par sa construction, demandant de lourds efforts, lorsqu'il est créé de façon manuelle (Hernandez et al., 2006)[435], mais également par son format non normalisé et par son faible degré de

[430] Extraction de concepts et de relations entre concepts à partir des documents multilingues : Approche statistique et ontologique, p. 49

[431] Les linguistiques de corpus, p. 78

[432] TtoO : une méthodologie de construction d'ontologie de domaine à partir d'un thésaurus et d'un corpus de référence, p. 4

[433] Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI)

[434] Building a WordNet for Arabic

[435] TtoO : une méthodologie de construction d'ontologie de domaine à partir d'un thésaurus et d'un corpus de référence

formalisation (Mizoguchi, 2004)[436]. Pour pallier à ces inconvénients, différentes solutions ont été proposées dans la littérature, notamment en faisant appel à des approches prenant en compte des connaissances linguistiques et des avancées en ingénierie des connaissances, au travers des ontologies.

7.1.2.2 Ontologie

Une ontologie est une représentation formelle d'un domaine, plus précisément, « une spécification *formelle et explicite d'une conceptualisation partagée* » (Gruber, 1993)[437], permettant de fournir un vocabulaire formalisé de concepts et de leurs relations afin d'obtenir une meilleure extraction des informations recherchées.

Comme les principaux langages de représentation des connaissances issus des sciences cognitives et de l'Intelligence Artificielle, l'ontologie n'a pas une vocation exclusivement documentaire au sens de l'indexation et de la recherche d'information ; mais, elle vise également à participer à l'ingénierie des connaissances d'un domaine et, en particulier, à « spécifier explicitement une conceptualisation » (Gruber, 1993)[438], recensant les catégories d'objets et/ou les concepts du domaine considéré et éventuellement représentant leurs propriétés ainsi que les relations qu'ils entretiennent entre eux, obtenant des hiérarchies ou des réseaux de concepts (Habert et al., 1997)[439].

Cependant, leur élaboration est coûteuse, nécessitant de nombreuses interventions manuelles, puisque les techniques de construction d'ontologies de la littérature ne basent l'élaboration de l'ontologie sur aucune connaissance préalable du domaine mais sur un corpus de référence qui est analysé (Hernandez et al., 2006)[440].

Malgré cela, l'ontologie possède un niveau d'abstraction conceptuelle, jouant un rôle primordial dans la communication homme-machine. De plus, le domaine de la chimie peut être appréhendé comme une représentation mentale plus ou moins universelle ou comme des catégories partagées par une communauté (Zacklad, 2007)[441]. Par conséquent, le développement d'une ontologie est pertinent, notamment afin de :

---

[436] Le rôle de l'ingénierie ontologique dans le domaine des EIAH

[437] A Translation Approach to Portable Ontology Specification

[438] Idem

[439] Les linguistiques de corpus

[440] TtoO : une méthodologie de construction d'ontologie de domaine à partir d'un thésaurus et d'un corpus de référence

[441] Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI)

- *Partager la compréhension commune de la structure de l'information entre les personnes ou les fabricants de logiciels,*
- *Permettre la réutilisation du savoir sur un domaine,*
- *Expliciter ce qui est considéré comme implicite sur un domaine,*
- *Distinguer le savoir sur un domaine du savoir opérationnel*
- *Analyser le savoir sur un domaine (Noy et al., 2000)*[442]

Pour cela, nous choisissons l'ontologie pour modéliser les connaissances du domaine de chimie en arabe, reflétant une conceptualisation du monde, recensant les catégories d'objets et/ou les concepts du domaine considéré et éventuellement représentant leurs propriétés ainsi que les relations qu'ils entretiennent entre eux, obtenant des hiérarchies ou des réseaux de concepts (Habert et al., 1997)[443].

7.2 Ontologie terminologique

L'ontologie terminologique consiste à la représentation des termes d'un domaine et de leurs relations. L'idée de base de cette modélisation est que la signification d'un terme dépend des autres termes qui cooccurrent avec ce terme, c'est-à-dire que la signification d'un concept est liée au réseau sémantique auquel il fait partie et de ses relations avec les autres concepts du réseau. Le but est d'améliorer les résultats de l'extraction d'information en se basant sur une classification conceptuelle des informations recherchées (Despres et al., 2008[444]).

Cette tâche de modélisation s'effectue en trois étapes, correspondant à trois engagements :

- *« un engagement sémantique, fixant le sens linguistique des concepts,*
- *un engagement ontologique fixant leur sens formel,*
- *un engagement computationnel déterminant leur exploitation effective » (Bachimont, 2000)*[445]

7.2.1 Conceptualisation

La conceptualisation consiste à déterminer les concepts du domaine étudié, dénotés par des termes simples ou complexes, dans les corpus. A partir de l'étude terminologique et en

[442] Développement d'une ontologie 101 : Guide pour la création de votre première ontologie, p. 1
[443] Les linguistiques de corpus
[444] Réseau terminologique versus Ontologie
[445] Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en Ingénierie des connaissances

nous basant sur la classification de la chimie que nous avons construite, nous effectuons la correspondance entre les termes et les concepts qui leur sont associés.

7.2.1.1 Concept

Un concept peut être considéré comme une représentation abstraite d'une idée et/ou d'un objet conçu par l'esprit et défini par ce à quoi il renvoie dans le monde, permettant d'organiser les perceptions et les connaissances (cf. 1.1.26.2.2.1 Concept2.1.3.2).

Par exemple, le terme « ماء = m*â'* = eau » signifie liquide incolore et inodore.

La distinction entre terme et concept est que si les termes sont en nombre potentiellement infini, les concepts correspondent à un ensemble restreint de notions associées aux ressources d'une collectivité et dépendant notamment de sa langue (Maniez, 1999)[446].

En science en général, en chimie en particulier, les concepts sont appréhendés comme des représentations mentales plus ou moins universelles ou comme des catégories a priori largement partagées dans la communauté scientifique (Guarino, 1998)[447]. Par conséquent, les concepts de la chimie correspondent aux catégories de ce domaine, que nous avons déterminées dans notre classification (cf. 5.35.3 Classification adoptée).

7.2.1.2 Classification

La classification correspond à l'organisation des connaissances d'un domaine afin d'en faciliter l'accès et l'étude (cf. 5.2.26.2.2.1 Classification de la chimie2.1.3.2) ; elle nécessite une compréhension et une maitrise des notions du domaine étudié ainsi qu'une habileté à organiser ces connaissances afin d'obtenir une classification pertinente.

Par exemple, le terme « ماء = m*â'* = eau » peut être :

- Réactif d'une réaction,
- Produit d'une réaction,
- Solvant d'une réaction,
- Molécule d'eau composée d'un atome d'oxygène et de deux atomes d'hydrogène,
- Liquide utilisé pour nettoyer la verrerie.

[446] Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI)

[447] Formal Ontology and Information Systems

Ce terme possède plusieurs caractéristiques (cf. 2.2.2.2.26.2.2.1 Métonymie2.1.3.2) et sera relié par plusieurs relations afin de fournir à l'utilisateur toutes ces informations.

Ainsi, ce travail est effectué pour tous les termes identifiés ; il permet de déterminer les concepts du domaine étudié, considérés comme les classes dans une ontologie, ainsi que les propriétés de chaque concept, décrivant leurs caractéristiques et correspondant aux individus des classes dans une ontologie. Les classes et les individus sont des termes nominaux alors que les relations sont des termes verbaux.

7.2.2 Relations sémantiques

Dans le cadre de construction des ontologies, les relations sémantiques s'intéressent à deux types de relation (Harrathi, 2009)[448] : les relations hiérarchiques et les relations non-hiérarchiques, comprenant les relations d'opposition et les relations d'équivalence (Albeiriss, 2017)[449].

7.2.2.1 Relations hiérarchiques

Deux relations hiérarchiques existent : les relations d'hyperonymie et de méronymie (Aussenac-Gilles et al., 2000)[450].

D'une part, il s'agit des relations hyperonymes - hyponymes, allant du terme générique au terme spécifique, c'est-à-dire, de la classe supérieure à la classe inférieure, ou un rapport d'inclusion, c'est-à-dire, de l'hyponyme à l'hyperonyme. Cela implique une relation avec le verbe être, comme « l'azote est un élément chimique » ; il s'agit ici d'une définition. Et d'autre part, les relations holonymes – méronymes, allant du tout à la partie, impliquant une relation avec le verbe avoir, comme « la molécule d'eau a un atome d'oxygène et deux atomes d'hydrogène » ; il s'agit d'une relation partitive hiérarchisée, plus précisément de la catégorie composant - assemblage.

Ces relations hiérarchiques concernent les unités terminologiques complexes (UTC), celles contenant une relation de coordination, relation égalitaire reliant deux ou plusieurs éléments, présentant un patron morphosyntaxique dont la tête est l'hyperonyme et l'expansion de coordination l'hyponyme, comme « ألدهيدات أليفاتيـة و عطرية = *'aldahîd*ât *'alîf*âtiyyat wa *ʿiṭrî*yyat

[448] Extraction de concepts et de relations entre concepts à partir des documents multilingues : Approche statistique et ontologique, p. 84

[449] Conception d'une ontologie à partir d'une étude terminologique basée sur corpus : le cas de la chimie en arabe

[450] Les relations sémantiques : du linguistique au formel

= aldéhydes aliphatiques et aromatiques ». Cette relation de coordination, écrite en langage naturel, signifie qu'il existe des aldéhydes aliphatiques et des aldéhydes aromatiques, un aldéhyde étant un compose chimique ; dans le langage formel, cela se traduit par l'aldéhyde est une classe possédant deux individus, les aldéhydes aliphatiques et les aldéhydes aromatiques.

D'autres UTC, contenant cette fois-ci une relation de subordination, relation non égalitaire reliant les éléments entre eux par une relation de hiérarchisation, sont également concernées par ces relations hiérarchiques, notamment celles contenant une expansion d'annexion, comme « ذرة الكلور = *ḏ*arrat alklûr = atome de chlore » ou « ذرة الأزوت = *ḏ*arrat *'azût* = atome d'azote ». Cela indique que le chlore et l'azote sont composés d'atomes et puisque le chlore et l'azote sont des éléments chimiques, cela implique que tout élément chimique est constitué d'un atome, c'est-à-dire, l'élément chimique est une classe, son attribut est un atome et ses individus sont le chlore et l'azote. Par conséquent, les individus héritent de cet attribut.

Les raisonnements de composition de relations montrent que si les concepts C1 et C2 sont reliés par la relation R1, et si les concepts C2 et C3 sont liés par la relation R2, alors C1 et C3 sont liés implicitement par la relation R3 ; l'expression de ces relations permet aussi de dire que deux relations ne sont pas composables et toute composition non explicitée n'existe pas (Habrant et al., 1999)[451].

7.2.2.2 Relations non-hiérarchiques

Pour les relations non-hiérarchiques, nous nous sommes limités à trois relations : les relations d'opposition, les relations d'équivalence et les relations de causalité. En général, l'identification de ces relations consiste à trouver les paires ou les couples de termes qui forment les arguments d'une relation et à déterminer l'étiquette pour la relation sémantique qui relie les termes arguments de la relation.

7.2.2.2.1 Relation d'opposition

La relation d'opposition correspond à l'antonymie qui peut se former soit par des mots simples, opposés par leur sens, comme 'acide - base', soit par des mots préfixés, que le sens de leur préfixe oppose, comme 'saturé - insaturé', soit par des mots suffixés, que le sens de leur suffixe oppose, comme 'hydrophile - hydrophobe'.

[451] Utilisation des réseaux sémantiques pour la navigation dans l'hypertexte, p. 2

De plus, l'antonymie peut être absolue, comme cation - anion, ou relative, signifiant que les unités lexicales ne s'opposent que dans certains contextes, comme composé transparent - composé coloré. Ainsi, le sens de l'antonymie est différent. En effet, pour des termes contraires (ou complémentaires ou non gradables), l'un exclut l'autre comme 'organique - inorganique'. Pour des antonymes réciproques, il s'agit de deux faits complémentaires mais inversés, comme 'oxyder - réduire'. Quant aux antonymes polaires (ou gradables), ils comportent deux termes avec un ou plusieurs termes intermédiaires comme 'solide - liquide - gaz'.

7.2.2.2.2 Relation d'équivalence

La relation d'équivalence correspond à la synonymie, indiquant une relation de sens entre des unités lexicales, c'est-à-dire que les synonymes ont un même signifié et des signifiants différents et se distinguant des homonymes avec un même signifiant mais des signifiés différents.

D'une part, la synonymie peut être totale ou absolue, impliquant que le contexte n'intervient pas ; très rare, elle concerne généralement les nomenclatures scientifiques, notamment la nomenclature chimique, comme 'alcoolique = éthylique', mais également les collocations, relation privilégiée entre les mots, permettant d'exprimer l'intensité, le choix de l'adjectif, de l'adverbe ou de la locution adverbiale dépendant du verbe ou du substantif qui y est associé, comme 'solution fortement concentrée = saturée'. D'autre part, la synonymie la plus courante est partielle ou contextuelle, impliquant que c'est le contexte qui permet de savoir quelle acception est à prendre en considération dans le cas où les synonymes concernent des unités polysémiques, c'est-à-dire que la synonymie ne porte que sur une acception, puisque les synonymes peuvent varier en fonction des différents sens des mots polysémiques.

7.2.2.2.3 Relation de causalité

Dans les corpus techniques et scientifiques, la présence de relation de causalité, plus précisément, les relations causales efficientes et les relations causales formelles, permet d'expliquer la technique et la science, même si les scientifiques abandonnent l'expression langagière de la relation de causalité au bénéfice de formules, de théories et de concepts de loi (Sabra, 2012)[452].

Certains patrons morphosyntaxiques nominaux et/ou verbaux de la chimie révèlent dans les textes l'expression de causalité, notamment les relations d'influence ou d'interaction entre

[452] Les relations de causalité en arabe et en français avec une constitution de ressources linguistiques utilisables par l'informatique

actions ; traduites en langage formel, ces relations permettent de répondre explicitement à des requêtes, appelées questions de compétence (Bendaoud, 2007)[453], telles que « Avec quel composé chimique est séché le produit? » ou « Quelle spectroscopie est utilisée pour l'analyse du produit? » ou encore « Quelle réaction permet d'obtenir ce composé chimique? ».

7.2.3 Quelques pistes pour l'exploitation informatique

La conceptualisation d'un domaine de connaissance permet de recenser ses concepts et éventuellement de représenter leurs propriétés ainsi que les relations qu'ils entretiennent entre eux, organisées en hiérarchie ou réseaux de concepts (Habert et al., 1997)[454]. Les objectifs principaux pour la construction d'une ontologie sont de partager la compréhension commune de la structure de l'information d'un domaine donné pour les applications les plus diverses, de permettre la réutilisation du savoir sur un domaine, d'expliciter ce qui est considéré comme implicite sur un domaine et d'analyser le savoir sur un domaine (Noy et al., 2001)[455].

Afin de construire notre ontologie de la chimie en arabe, nous nous basons sur les travaux de Jazzar (2009)[456] qui construit une ontologie de l'économie pétrolière à partir d'une analyse terminologique ; les concepts sont modélisés par des nœuds, les relations entre les concepts par des arcs et nous choisissons d'implanter notre ontologie dans un modèle simple et répandu, le logiciel Protégé, en utilisant des langages OWL (Ontology Web language), afin de décrire les classes, les propriétés, les instances de classes et les relations entre ces instances (Albeiriss, 2017)[457].

7.2.3.1 Langages OWL (Ontology Web language)

Standard proposé par le W3C (World Wide Web Consortium), OWL (Ontology Web Language) est un langage formel utilisé pour représenter les ontologies (Zacklad, 2007)[458]. Cela permet de décrire avec précision les propriétés et les classes, puisqu'il s'appuie sur une stricte séparation classe/instance, l'héritage de propriétés, l'expression de contraintes de cardinalité et de contraintes logiques sur les relations entre propriétés, etc. Exprimé à partir du langage RDF (Ressource Description Framework), lui-même exprimé à l'aide de balises XML

[453] Construction d'une ontologie à partir d'un corpus de textes avec l'ACF
[454] Les linguistiques de corpus
[455] Développement d'une ontologie 101 : Guide pour la création de votre première ontologie
[456] L'ontologie de l'économie pétrolière en Arabie Saoudite et analyse terminologique anglais-français-arabe
[457] Conception d'une ontologie à partir d'une étude terminologique basée sur corpus : le cas de la chimie en arabe
[458] Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI)

comme tous les langages du web sémantique, le langage OWL vise à exporter et importer des informations autant en provenance de ressources Web que d'autres ontologies.

Il existe trois sous-langages d'OWL principaux : OWL Lite, OWL DL et OWL Full ; OWL DL (OWL Description Logics), fondé sur la logique descriptive, est utilisé pour représenter la connaissance terminologique d'un domaine d'application d'une manière formelle et structurée et permet notamment d'utiliser tous les entiers positifs dans les contraintes de cardinalité.

L'idée de base d'une ontologie est la cohérence de la conceptualisation du domaine étudié, nécessitant de décrire les classes auxquelles les individus appartiennent, ainsi que les propriétés héritées d'autres classes ; une classe peut alors être un objet et un ensemble d'objets, comme le terme « كحول = ku*ḥ*ûl = alcool » qui est un objet de la classe 'Composé chimique' mais également l'ensemble des alcools tels que « إيثانول = *'îṯânûl* = éthanol », « بوتانول = bûtânûl = butanol » et l'ensemble de ces alcools sont des composés chimiques puisque les instances d'une sous-classe héritent des instances d'une classe. À cela, il faut ajouter des propriétés, permettant d'affirmer des faits généraux sur les membres des classes et des faits particuliers sur les individus, sachant que les classes et les individus ne doivent pas forcément être adjacents l'un par rapport à l'autre, comme les alcools peuvent servir de solvants, de combustibles ou de réactifs.

7.2.3.2 Logiciel Protégé

Éditeur d'ontologies le plus connu et le plus utilisé, Protégé a été développé par l'université de Stanford et de Manchester en 2000 et a bien évolué puisqu'il a intégré les standards du web sémantique et OWL avec des interfaces graphiques et un moteur d'inférence pour raisonner. Ce logiciel permet de construire une ontologie pour un domaine précis en définissant des formulaires d'entrée des données afin d'acquérir des données à partir de ces formulaires. Nous préconisons l'utilisation de l'environnement Protégé pour la construction des ontologies du domaine en raison de son accès libre et de la facilité d'utilisation.

Pour l'implémentation informatique de l'ontologie de la chimie, nous avons utilisé la version 4.1.0 de Protégé et nous présentons comment éditer l'ontologie.

Dans un premier temps, nous lancons le logiciel Protégé (cf. Figure 49 : Lancement du logiciel Protégé).

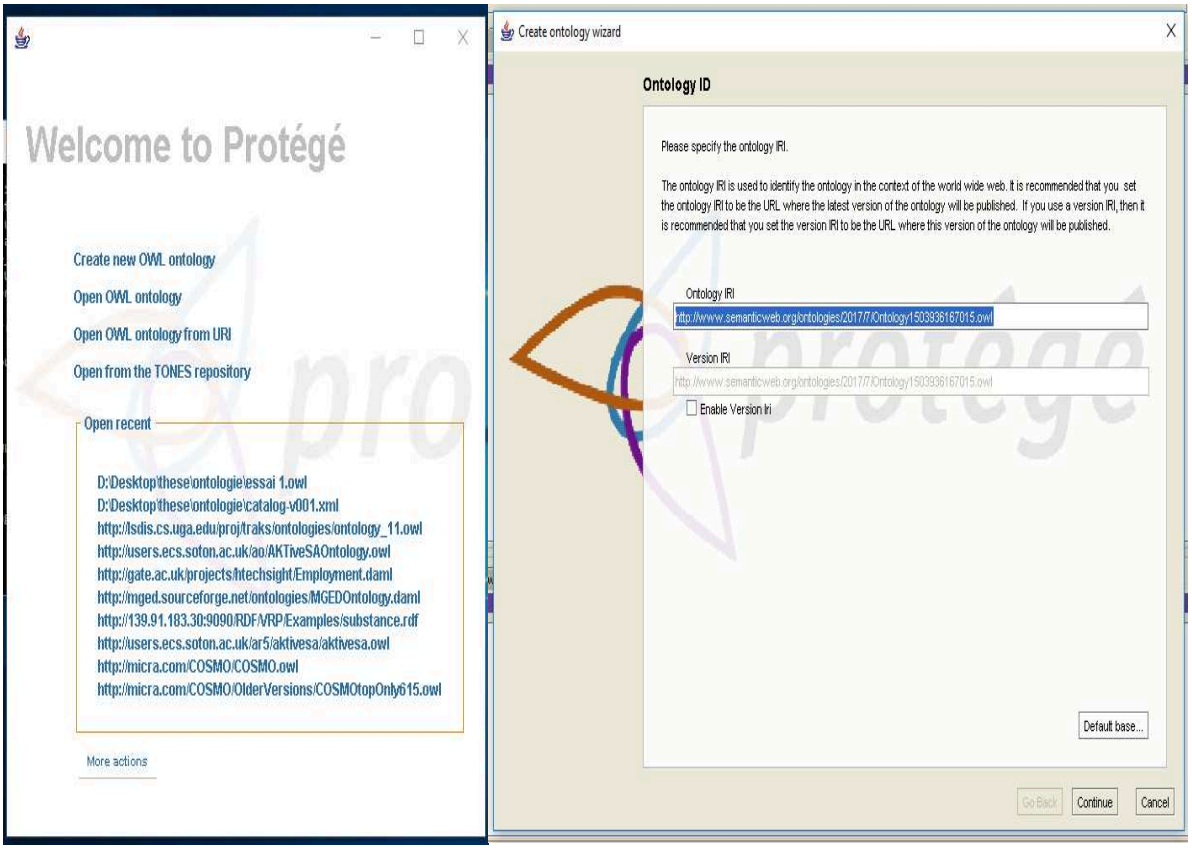


Figure 49 : Lancement du logiciel Protégé

Dans un second temps, nous créons une nouvelle ontologie en choisissant son nom, son emplacement et le langage de spécification (cf. Figure 51 : Interface de l'ontologie sur Protégé).

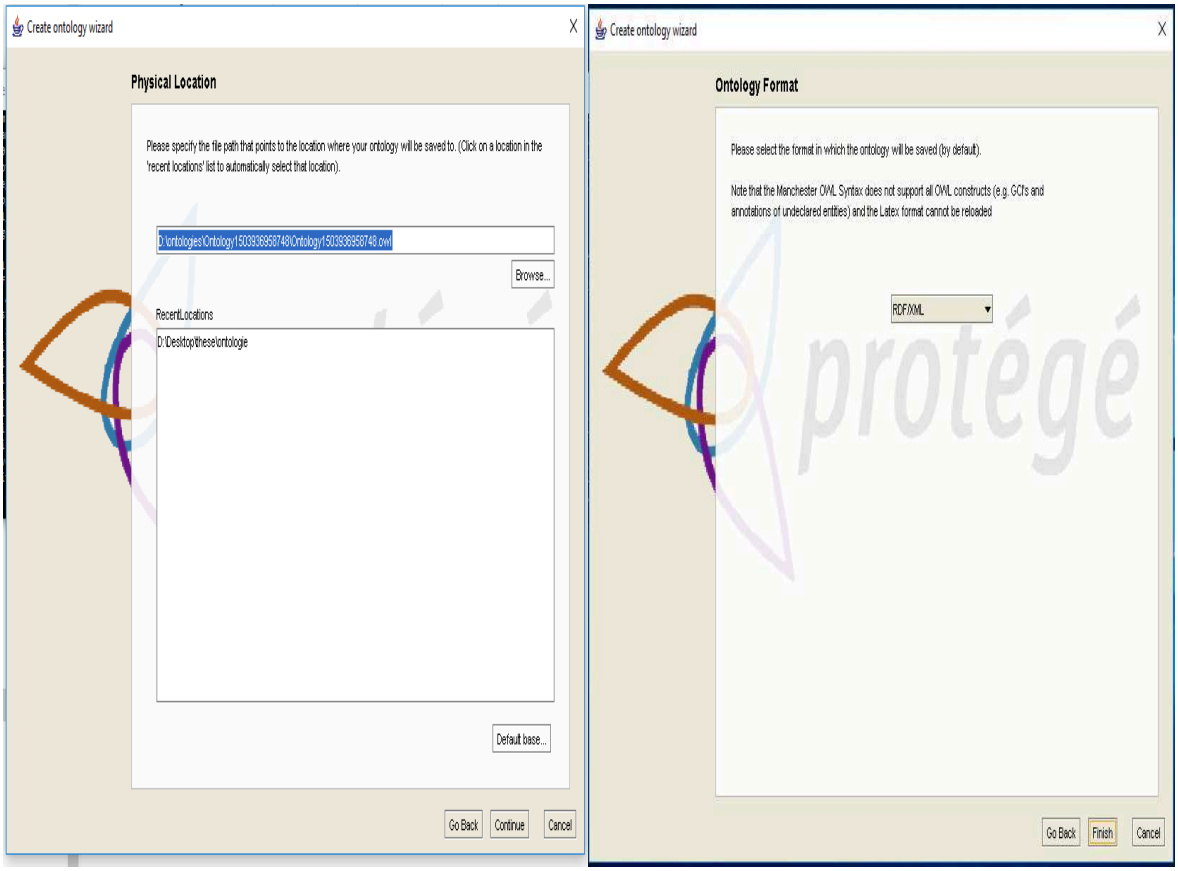


Figure 50 *: Création d'une nouvelle ontologie*

Nous obtenons une interface dans laquelle les classes de l'ontologie, ses propriétés et les instances des classes (les individus) sont ajoutées et sont respectivement représentées par les onglets 'Classes', 'Object Properties' et 'Individuals'.

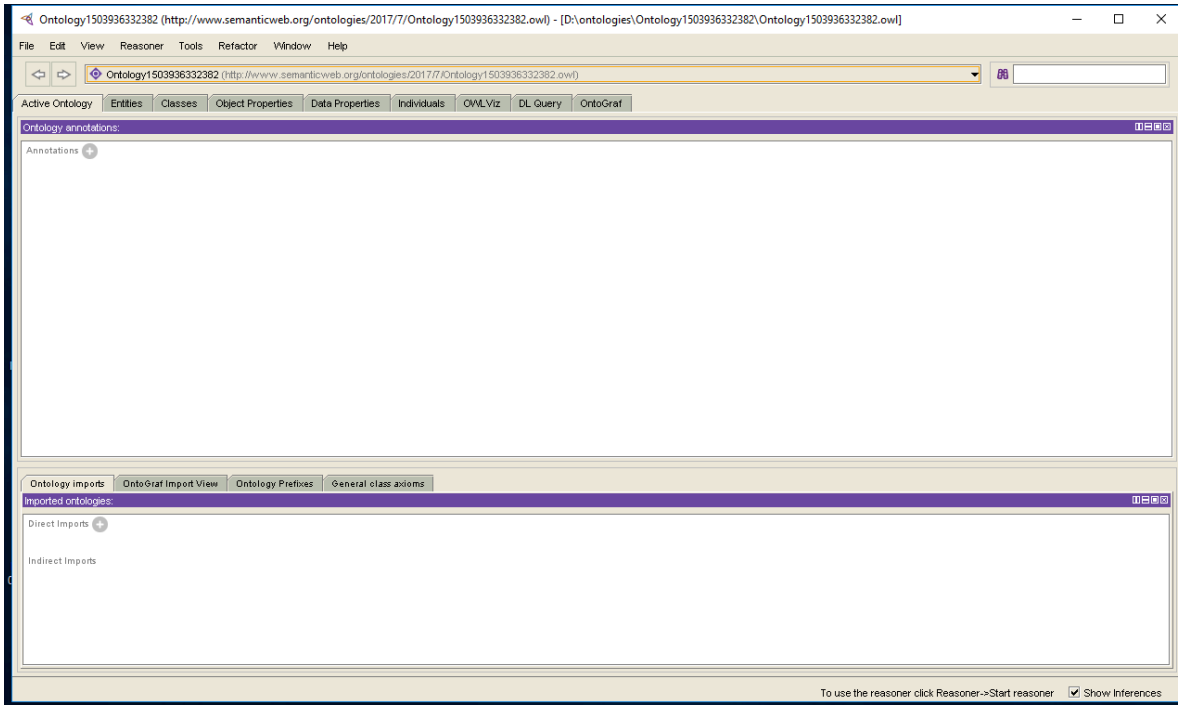


Figure 51 *: Interface de l'ontologie sur* Protégé

À présent, nous allons implémenter l'ontologie.

7.2.3.2.1 Classe

Les fonctionnalités de la fenêtre « Classes » permettent principalement à définir la hiérarchie des classes à créer ainsi que les commentaires pour chaque classe définie et à spécifier les contraintes relationnelles entre les classes.

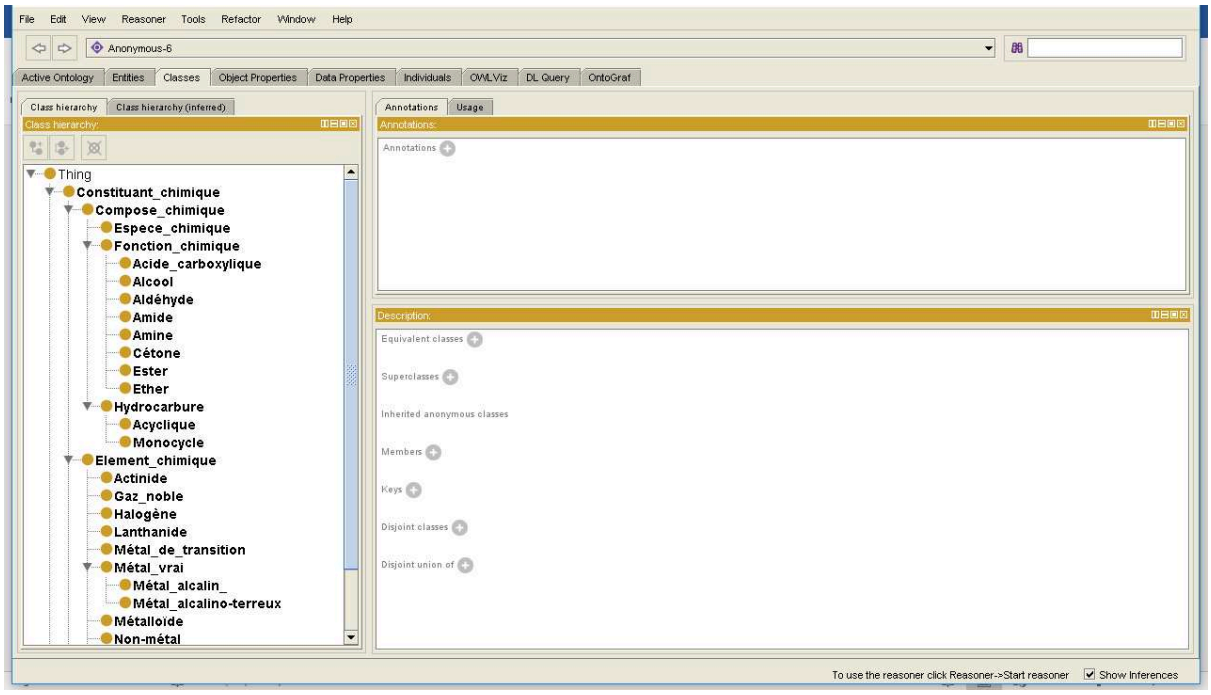


Figure 52 : Fenêtre « Classes » sur Protégé

7.2.3.2.2 Individu

La fenêtre « Individuals » regroupe les fonctions permettant de définir les instances objets des classes mais aussi de translater les propriétés relationnelles définies entre leurs types (classes) pour relier les individus, c'est-à-dire que les relations définies entre les classes sont les mêmes relations entre les instances de ces classes. Par exemple, pour la classe « composé chimique », nous définissons la relation « est constitué de » avec la classe « élément chimique » et la relation « est analysée par » avec la classe « méthode d'analyse », impliquant que ces mêmes relations seront définies entre les instances de la classe « composé chimique » et les instances des classes « élément chimique » et « méthode d'analyse ».

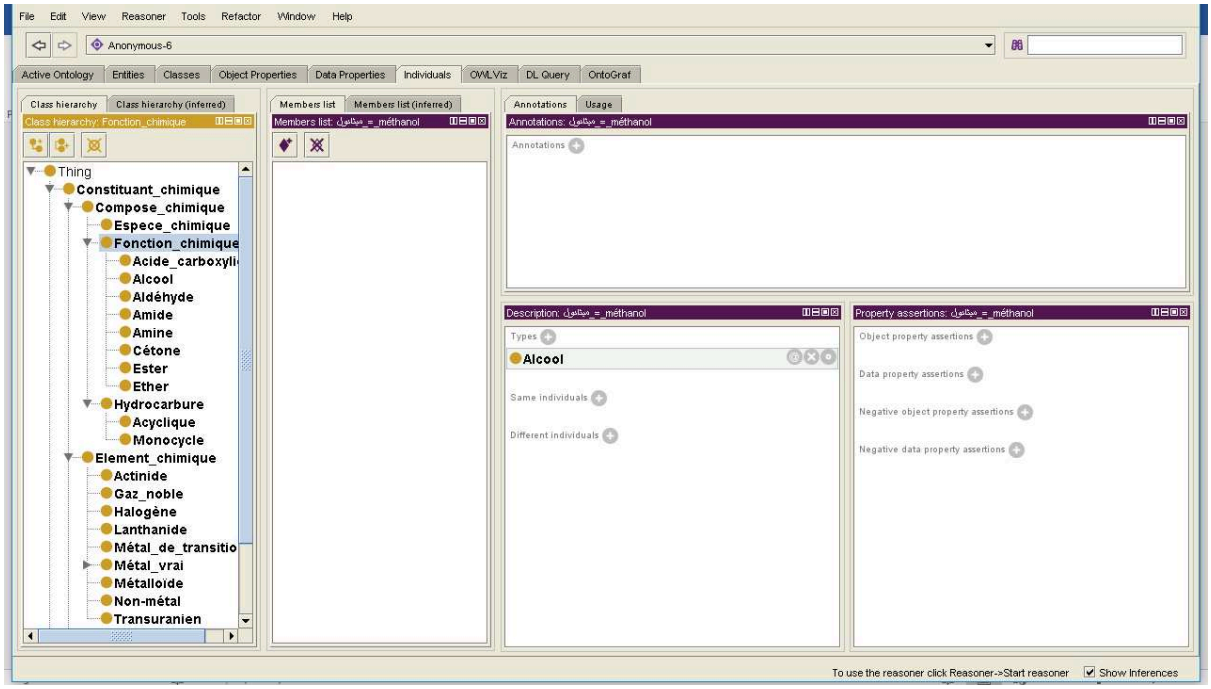


Figure 53 : Fenêtre « Individuals » sur Protégé

7.2.3.2.3 Propriété

La fenêtre « Properties » regroupe les fonctionnalités servant à définir les propriétés et les sous-propriétés sous forme hiérarchique et indique les options possibles de chaque propriété selon sa sémantique dans l'ontologie. Il s'agit des relations de type synonymie, 'est un', 'est une partie de' (Mallak, 2011)[459] et sont désignées par exemple par 'inversible', 'symétrique', 'transitive'… et leurs relations inverses si elles existent.

[459] De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information

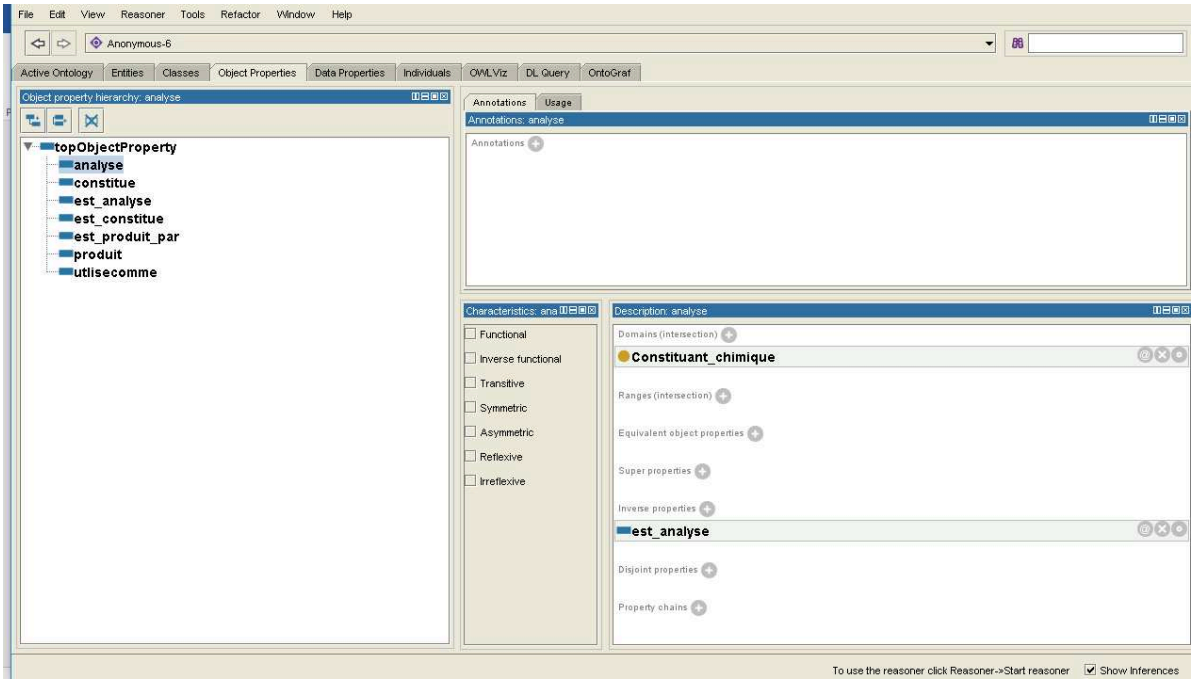


Figure 54 : Fenêtre « Properties » sur Protégé

7.2.3.2.4 Représentation de l'ontologie

À partir des classes, des individus et des propriétés, nous représentons l'ontologie à l'aide de la fonction « OntoGraf » afin de visualiser les connaissances du domaine.

Nous distinguons deux types de nœuds : les nœuds concepts et les nœuds relations, reliés entre eux par un arc, traduisant que le concept associé au nœud concept est un argument de la relation représentée par le nœud relation ; chaque nœud du graphe est étiqueté par un type dénotant un concept et un marqueur correspondant à une instance du concept. De la même manière, les nœuds relations sont étiquetés par un type qui correspond au nom de la relation. Par exemple, dans le couple (« eau », « réaction chimique »), l'étiquette de la relation peut être « produire » ou « utiliser ».

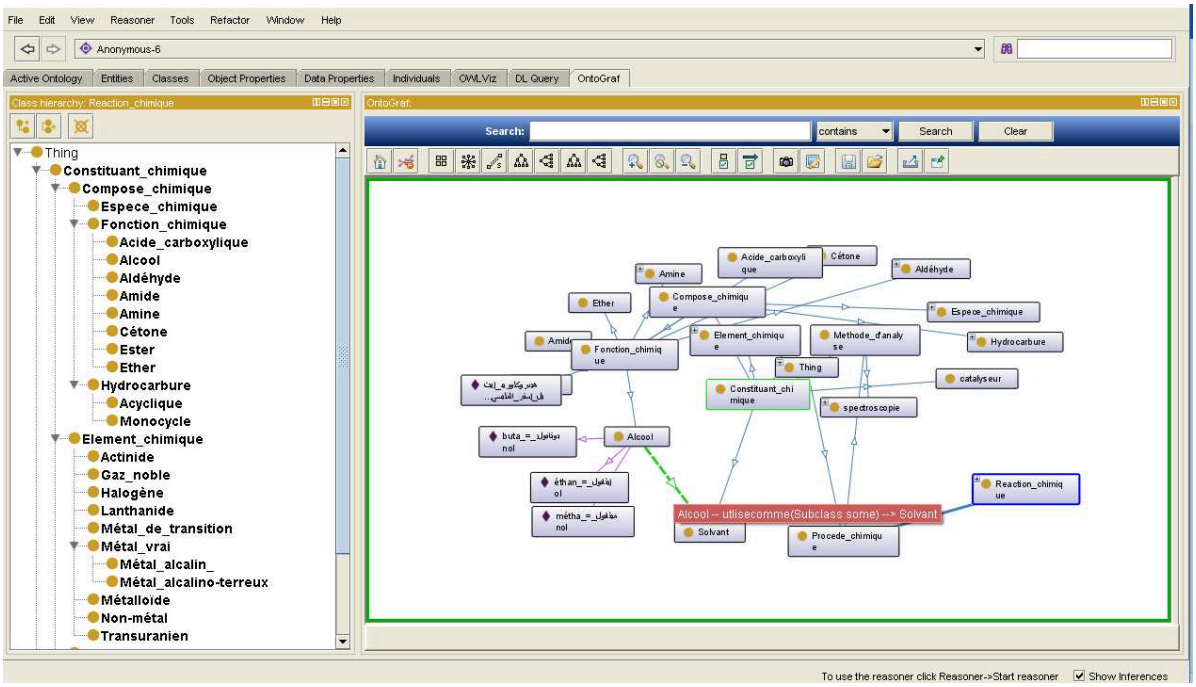


Figure 55 : Fenêtre « OntoGraf » sur Protégé

Ce graphe met en evidence les relations entre les classes et/ou les individus ; par exemple, l'alcool, un composé chimique contenant une fonction chimique, est utilisé comme solvant, un constitutant chimique. Cette figure représente les classes des « constituants chimiques », notamment la « fonction chimique », leurs individus « méthanol, éthanol et butanol » identifiés dans le corpus, et leurs propriétés, à savoir, l'alcool peut être utilisé comme un solvant.

Dans ce chapitre, nous avons présenté la construction d'une ontologie de la chimie en arabe. Cela a permis de représenter les termes de la chimie en arabe et leurs relations. Cette ontologie pourra être exploitée par des programmes informatiques interagissant avec l'utilisateur à l'aide d'un formulaire ou d'un autre type de langage de requête afin de fournir des réponses logiquement fondées, c'est-à-dire « vraies ». Elle permet notamment de faciliter l'accès à l'information du web, puisque les connaissances du domaine sont représentées à travers des concepts et des relations entre ces concepts.

De plus, construire cette ontologie nous a permis de tester et d'imaginer l'impact de différentes organisations des connaissances, mais surtout de vérifier la cohérence de notre conception du domaine de la chimie en arabe et ainsi de valider ce travail de recherche.

Dans cette partie, nous avons présenté une méthode d'extraction des termes de la chimie en arabe ainsi que leurs relations morphosyntaxiques. Cette approche devrait permettre à terme

d’extraire les informations recherchées à partir d’un texte original (corpus brut non organisé, composé de mots), d’obtenir un texte enrichi (corpus structuré, contenant des termes et des concepts), et de les organiser dans une ontologie selon des classes et des propriétés.

Conclusion générale

Notre travail de recherche nous a permis d'atteindre notre objectif de la conception d'un outil informatique qui permettra d'extraire les termes de la chimie en arabe et leurs relations. Réaliser cet outil d'extraction nécessitait une analyse linguistique approfondie basée sur corpus pour identifier les patrons morphosyntaxiques des termes et la mise en place d'une classification du domaine étudié afin de construire une grammaire d'identification.

Pour cela, nous avons dans un premier temps présenté l'étude, les procédés, les spécificités et les frontières de la terminologie de la chimie en arabe. À la lumière de notre étude, nous avons établi que la chimie présente une spécificité unique, avec un vocabulaire se composant de termes issus de la nomenclature chimique, normalisée par l'UICPA, ainsi que d'autres termes, l'ensemble étant tributaire des terminologies établies en anglais et/ou en français. Pour ces autres termes (majoritairement des noms et quelques verbes), nous avons identifié leurs modes de création lexicale (morphosyntaxique, transfert sémantique et emprunt). Cette identification a mis en évidence leurs structures morphosyntaxiques (expansion d'identification, expansion annective, expansion complétive et extension de coordination), mais également la combinaison de ces structures ainsi que de nouvelles constructions (spécifiques aux composés chimiques). Pour l'analyse de ces structures morphosyntaxiques, nous avons pris en compte la description lexicale et sémantique du terme (axe paradigmatique) et la description syntaxique du syntagme (axe syntagmatique).

Dans un second temps, nous avons constitué un corpus d'environ 100 000 mots non voyellés, corpus se composant d'un ensemble de textes (spécialisés, vulgarisés, fiables, authentiques) jugés pertinents pour la conception de l'extracteur. En plus de respecter les critères linguistiques, ces textes au format TXT autorisent une exploitation informatique des données, d'autres formats butant devant les contraintes techniques (peu de textes numérisés et certifiés en arabe, reconnaissance optique des caractères (OCR) arabes en cours de développement). Notre exploitation du corpus se base sur une méthode semi-automatique, réalisée à l'aide des logiciels AntConc, Xerox et Kawâkib. Comparé à un travail manuel, cette méthode présente

l'avantage d'alléger, d'accélérer et surtout d'enrichir le travail de fichage contextuel, ne sélectionnant plus des contextes arbitrairement, mais en prenant connaissance de toutes les occurrences du terme étudié et en définissant préalablement un certain nombre de concepts. À partir du corpus recueilli et des quelques outils informatiques dont nous disposons, nous avons identifié puis classé tous les termes de la chimie de notre corpus en fonction de leur sens contextuel, propre au domaine. Cette classification prend en compte tous les termes de notre corpus ; et par conséquent, peut être considérée comme une classification de la chimie, classification conçue pour être complétée par d'autres termes de la chimie non présents dans ce corpus en créant éventuellement de nouvelles classes et sous-classes.

Enfin, nous avons construit les règles d'une grammaire d'identification qui pourront être implémentées dans un système d'analyse morphosyntaxique de l'arabe afin d'extraire les termes du domaine et leurs relations. Les constructions syntaxiques et la classification que nous avons définie précédemment déterminent ainsi les différents patrons morphosyntaxiques (nominaux, adverbiaux et verbaux). Ces patrons morphosyntaxiques sont traduits informatiquement par des règles de grammaire, intégrant donc l'ensemble de la description lexicale, sémantique et syntaxique des termes simples et complexes. Cette approche par patrons, robuste, pourrait devenir une approche incontournable dans le cas de phrases complexes pour lesquelles les risques d'erreur et d'ambiguïté sont trop élevés. Cependant, définir un ensemble de règles stables et exhaustives par rapport à un objectif donné se révèle laborieux et fastidieux, avec une difficulté à garantir la cohérence de l'ensemble des règles puisque les patrons définis sont très dépendants du domaine, des types de texte, et même de l'application visée. De ce fait, cette approche est appelée à évoluer et à être enrichie par d'autres contributions. Une des applications de notre travail consiste à représenter les termes de la chimie en arabe et leurs relations par une ontologie. Cette ontologie pourra être exploitée par des programmes informatiques, et notamment faciliter l'accès de l'information du web, puisque les connaissances du domaine sont représentées à travers des concepts et des relations entre ces concepts. De plus, construire cette ontologie nous a permis de tester et d'imaginer l'impact de différentes organisations des connaissances, mais surtout de vérifier la cohérence de notre conception du domaine de la chimie en arabe et ainsi de valider ce travail de recherche.

Durant notre recherche, nous avons pu présenter nos travaux (poster, communications et publications) dans différentes manifestations scientifiques en France et à l'étranger avec un public composé de chercheurs de différentes spécialités (lexicologues, terminologues, linguistes et informaticiens). Ces travaux portaient sur les unités lexicales complexes dans une langue spécialisée (VIIe colloque international de lexicologie à Tunis), les verbes de la chimie en arabe (Conférence de Terminologie & Ontologie : Théories et Application à Chambéry), la nomenclature de la chimie en arabe (colloque international de l'Institut Supérieur des Langues de Gabès) et la construction d'une ontologie du domaine de la chimie (Colloque International sur le Document Électronique). Ces échanges et rencontres ont permis d'enrichir nos connaissances, de confronter nos idées, et de valider notre méthodologie.

Que ce soit lors de ces présentations et/ou durant notre travail de recherche, notre connaissance de la chimie nous a assuré la maîtrise de ses concepts et de ses besoins, ce qui nous a aidée à appréhender nos problématiques de recherche et d'y répondre efficacement.

De plus, une approche de fouille de textes nous a permis de structurer la construction des connaissances, représentées dans un langage formel, nous autorisant ainsi à raisonner sur le contenu des textes et d'avoir une vision synthétique du domaine.

En conciliant les deux aspects de notre double compétence en linguistique et en chimie, nous avons étudié la terminologie de la chimie en arabe dans une approche de fouille de textes. Cela nous a permis en fin de compte, la conception d'un outil d'extraction morphosyntaxique des termes de chimie en arabe. Cependant, il ne s'agit pas là de la fin de notre travail de recherche, loin de là, mais bien de son début. En effet, nous poursuivrons nos recherches pour élaborer cet outil afin de répondre aux problématiques du traitement automatique des langues en général et à l'arabe en particulier, avec la mise en place de coordinations de méthodologies et de standards entre équipes pour le traitement informatique de l'arabe. La réalisation de cet outil validera ce présent travail ; l'utilisation d'autres corpus confirmera les règles de la grammaire d'identification proposées. Une fois l'outil réalisé et validé, nous pourrions alors étudier d'autres domaines de spécialité. Cet outil serait alors utile pour le traitement informatique de l'arabe et des domaines de spécialité.

Index des notions

Annexes

Annexe 1 : Corpus

Notre corpus est composé d'une dizaine de textes et l'ensemble est constitué de 100000 mots environs. Nous présentons chaque référence en indiquant son auteur, le nombre de chapitres et le nombre de mots.

- Référence 1 : il s'agit du mémoire de magistère de Hayate Aliouche, soutenu le 15/09/2001 à l'université de Constantine 1. Il est composé de cinq chapitres, contenant 12845 mots exactement.
- Référence 2 : il s'agit de la thèse de doctorat d'état de Nourreddine Beghidja en Phytochimie, soutenue en 2008 à l'Université de Constantine 1. Elle est composée de trois chapitres et contient 24868 mots exactement.
- Référence 3 : il s'agit d'une dizaine de cours de chimie sur le site de l'éducation algérien regroupant une dizaine de millier de mots (Consulté 13/06/2013). Voici les liens des cours :

http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/acides_bases/INDEX.HTM
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/les%20colorants/home.html
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/html_potentiel/index.htm
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/potentiel1/home.html
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/el%20aminiyate/index.htm
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/TabClass/index.htm
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/hydrocabonate1/index.htm
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/reaction%20chim/home.html
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/couholate/Home.htm
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/secirite_labo/index.htm
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/carbo/index.htm
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/ald%20kitounate/index.htm
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/kimia_mahalil/Index.htm
http://www.infpe.edu.dz/cours/Enseignants/secondaire/Chimie/chimie1/index.htm

Cependant, ces cours ne sont plus accessibles actuellement.

- Référence 4 : il s'agit de trois comptes rendus d'organismes internationaux contenant 55368 mots exactement. Ils sont téléchargeables sur les liens suivants :
    - http://archive.basel.int/meetings/cop/cop7/docs/08a2a.doc
    - http://chm.pops.int/Portals/0/download.aspx?d=UNEP-POPS-POPRC.5-3.Arabic.doc
    - http://www.pops.int/documents/meetings/poprc/POPRC3/POPRC3_Report_a/POPRC3_Report_add7_a.doc

Annexe 2 : Liste des termes du corpus

Voici la liste des termes de notre corpus. Chaque terme est indiqué en arabe, suivi de son équivalent français et de la référence du corpus.

Terme arabe	Équivalent français	Référence
اتحاد دولي للكيمياء البحتة والتطبيقية	Union Internationale de Chimie Pure et Appliquée (UICPA)	Réf. 4.b
اترازين	atrazine	Réf. 4.b
احتراق	combustion	Réf. 4.b
اختزال	réduction	Réf. 4.a
اختزال قلوي	réduction alcaline	Réf. 4.a
اختزال كيميائي	réduction chimique	Réf. 4.a
اختزال كيميائي في المرحلة الغازية	Réduction Chimique en Phase Gazeuse (GPCR)	Réf. 4.a
ارجاع	réduction	Réf 1.a
ارغون	argon	Réf. 1.c
ازاحة كيميـائيـة	déplacement chimique	Réf. 1.b
ايزومير	isomère	Réf. 4.c
ازوت	azote	Réf. 1.b
اس هيدروجيني	Potentiel Hydrogène (PH)	Réf. 4.c
استر	ester	Réf. 1.a
استخلص	extraire	Réf. 1.b
اسم كيميائي	nom chimique	Réf. 4.b
اسيتات الايثيل	acétate d'éthyle	Réf. 1.c
اسيتون	acétone	Réf. 1.b
اشارة احادية	singulet	Réf. 2.c
اشارة ثنائية	doublet	Réf. 2.c
اصطناع	synthèse	Réf. 1.a
اكسد	oxyder	Réf. 1.b
اكسـدة	oxydation	Réf. 1.a
اكسيد	oxyde	Réf. 1.b
اكسيد التيتان	oxyde de titane	Réf. 1.c
اكسيد الحديد	oxyde de fer	Réf. 4.a
اكسيد الكبريت	oxyde de soufre	Réf. 4.a
اكسيد الكربون	monoxyde de carbone	Réf. 4.a
اكسيد المنغنيز	oxyde de manganèse	Réf. 1.b
اكسيد الهيدروجين	peroxyde d’hydrogène	Réf. 2.b
اكسجين	oxygène	Réf. 4.b
الدرين	aldrine	Réf. 4.c
الدول	aldol	Réf. 1.a
الدهيد	aldéhyde	Réf. 1.a
الدهيد اليـفاتي	aldéhyde aliphatique	Réf. 1.a
الدهيد عطري	aldéhyde aromatique	Réf. 1.a
المونيوم	aluminium	Réf. 4.b

الومنيا	alumine	Réf. 4.a
امتزاز	adsorption	Réf. 4.b
امتص	absorber	Réf. 4.b
امتصاص	absorption	Réf. 1.b
امين	amine	Réf. 1.a
انبوب اختبار	tube à essai	Réf. 3.a
اندوسولفان	endosulfan	Réf. 4.b
انصهار	fusion	Réf. 1.c
اوكتانول	octanol	Réf. 4.c
اوزون	ozone	Réf. 2.b
ايبوكسيد	époxyde	Réf. 1.d
ايثانول	éthanol	Réf. 1.b
ايثر	éther	Réf. 1.b
ايثر اندوسلفان	endosulfan éther	Réf. 4.c
ايثـر البتـرول	éther de pétrole	Réf. 1.b
ايزوبريـن	isoprène	Réf. 3.b
أيضـة	métabolite	Réf. 4.c
ايمين	imine	Réf. 1.c
اينولات	énolate	Réf. 1.a
اينولات كيرالية	énolate chiral	Réf. 1.a
ايون	ion	Réf. 4.a
بارا ميتوكسي بنزالدهيد	paraméthoxybenzaldéhyde	Réf. 1.b
بخّر	vaporiser	Réf. 1.b
بروتون	proton	Réf. 3.a
بروكسيد	peroxyde	Réf. 4.a
بروكسيد الهيدروجين	peroxyde d'hydrogène	Réf. 4.a
بروم	brome	Réf. 4.a
برومة	Bromé	Réf. 4.a
بلور	cristalliser	Réf. 1.b
بلور	Cristal	Réf. 1.b
بنزالدهيد	benzaldéhyde	Réf. 1.b
بنزين	benzène	Réf. 1.d
بنية كيميائية	structure chimique	Réf. 2.b
بوتاسيوم	potassium	Réf. 1.b
بوتان	Butane	Réf. 1.a
بوليمر	polymère	Réf. 4.a

بوليفينيل الكلوريد	chlorure de polyvinyle	Réf. 4.b
بوليبريثن	polyuréthane	Réf. 4.a
بور	bore	Réf. 1.b
ببرمنغنـات	permanganate	Réf. 1.a
ببرمنغنـات البوتاسيـوم	permanganate de potassium	Réf. 1.a
بيكربونات	bicarbonate	Réf. 4.a
بيكربونات الصوديوم	bicarbonate de sodium	Réf. 4.a
بيكلورام	piclorame	Réf. 4.b
تأيّن	s'ioniser	Réf. 4.a
تأكسد	s’oxyder	Réf. 4.b
تانتالوم	tantale	Réf. 4.b
تبريد	refroidissement	Réf. 1.b
تجربة	expérience	Réf. 1.a
تحت الكلوريت	hypochlorite	Réf. 4.a
تحليل	analyse	Réf. 2.c
تحليل بنيوي	analyse structurelle	Réf. 2.c
تحريك	agitation	Réf. 1.b
تحريك مغناطيسي	agitation magnétique	Réf. 1.b
تحضير	préparation	Réf. 1.c
تربـين احادي	monoterpène	Réf. 2.b
تربين احادي أحادية الحلقة كحولي	alcool monoterpènique monocyclique	Réf. 2.b
تربين احادي احادية الحلقة هيدروكربوني	hydrocarbure monoterpènique monocyclique	Réf. 2.b
تربين احادي الدهيدي	aldéhyde monoterpènique	Réf. 2.b
تربين احادي ثنائية الحلقة كحولي	alcool monoterpènique bicyclique	Réf. 2.b
تربين احادي ثنائية الحلقةهيدروكربوني	hydrocarbure monoterpènique bicyclique	Réf. 2.b
تربين احادي غير حلقي هدروكربوني	hydrocarbure monoterpènique acyclique	Réf. 2.b
تربين احادي كيتوني	cétone monoterpènique	Réf. 2.b
تربين ثنائـي	diterpène	Réf. 2.a
تربين ثلاثي	triterpène	Réf. 2.a
ترسيب	précipitation	Réf. 4.b
ترشيح	filtration	Réf. 1.b
تركيب	composition	Réf. 4.b
تركيبة كيميائية	composition chimique	Réf. 2.a
تركيز	concentration	Réf. 4.b
تركيز جوي	concentration atmosphérique	Réf. 4.b
ترميد	incinération	Réf. 4.b

تسخين	chauffage	Réf. 1.c
تصبن	saponification	Réf. 2.b
تطاير	volatilité	Réf. 4.b
تفاعل	réaction	Réf. 1.a
تفاعل ألدولي	réaction aldolique	Réf. 1.c
تفاعل سمّي	réaction toxique	Réf. 4.b
تفاعل كيميائي	réaction chimique	Réf. 3.a
تقطير	distillation	Réf. 1.b
تقطير مكثف	hydrodistillation	Réf. 2.b
تقطير مرتد	chauffage à reflux	Réf. 1.c
تكـاثف	condensation	Réf. 1.a
تكـاثف الدولي	condensation aldolique	Réf. 1.a
تكثّف	se condenser	Réf. 2.b
تكرير	raffinage	Réf. 4.b
تكرير النفط	raffinage du pétrole	Réf. 4.b
تلوث	pollution	Réf. 4.b
تلوث جوي	pollution atmosphérique	Réf. 4.b
تنقية	purification	Réf. 1.b
توازن	équilibre	Réf. 3.a
توازن الكتلة	bilan de matière	Réf. 4.b
توازن كيميائي	équilibre chimique	Réf. 3.a
تيتانيوم	titane	Réf. 1.b
تيرفينول متعدد الفلورة	terphénol polychloré	Réf. 4.a
ثابت التوازن	constante d'équilibre	Réf. 3.a
ثابت قانون هنري	constante de Henry	Réf. 1.b
ثلاثي إيثيل أمين	triéthylamine	Réf. 1.b
ثلاثي عنق	tricol	Réf. 1.c
ثلاثي فليور	trifluorure	Réf. 1.c
ثلاثي فليور البور	trifluorure de bore	Réf. 1.c
ثاني أكسيد التيتانيوم	dioxyde de titane	Réf. 4.b
ثنائي بنزو باراديوكسين متعدد الكلور	polychlorodibenzoparadioxine	Réf. 4.a
ثنائي بنزوفوران متعدد الكلور	polychlorodibenzofurane	Réf. 4.a
ثنائي الفينيل	biphényle	Réf. 4.a
ثنائي الفينيل سباعي البروم	heptabromobiphényle	Réf. 4.a
ثنائي الفينيل سداسي البروم	hexabromobiphényle	Réf. 4.a
ثنائي الفينيل متعدد البروم	polybromobiphényle	Réf. 4.a

ثنائي الفينيل متعدد الكلور	polychlorobiphényle	Réf. 4.b
ثنائي إيثيل الإيثر	éther diéthylique	Réf. 1.b
ثنـائي عنـق	bicol	Réf. 1.c
ثاني أكسيد الكربون	dioxyde de carbone	Réf. 4.a
ثنائي فينيل الكلور	chlorobiphényle	Réf. 4.a
ثلاثي الفينيل المتعدد الكلور	polychlorotriphényle	Réf. 4.a
ثلاثي كلورو البنزين	trichlorobenzène	Réf. 4.b
ثلاثي كلورو الفينول	trichlorophénol	Réf. 4.b
ثنائي كلوروميثان	dichlorométhane	Réf. 1.c
ثنائي كلوروميثان جاف	dichlorométhane anhydre	Réf. 1.c
جامد	solide	Réf. 4.a
جذر الهيدروكسيل	radical hydroxyle	Réf. 4.b
جفّف	sécher	Réf. 1.b
جو من الأزوت	atmosphère d’azote	Réf. 1.c
حزمة الامتصاص	faisceau d’absorption	Réf. 2.c
حلقة	cycle	Réf. 2.a
حلقة بنزين	cycle benzénique	Réf. 4.b
حضّر	préparer	Réf. 1.d
حفظ	conserver	Réf. 1.b
حمام ثلجي	bain de glace	Réf. 1.b
حمض	acide	Réf. 1.a
حمض اميني	acide aminé	Réf. 1.a
حمض الخل	acide acétique	Réf. 3.a
حمض الشكيميك	acide shikimique	Réf. 2.b
حمض المفالونيك	acide mévalonique	Réf. 2.b
حمض النيتريك	acide nitrique	Réf. 4.a
حمض برونستد	acide de Brönsted	Réf. 3.a
حمض ضعيف	acide faible	Réf. 3.a
حمض قوي	acide fort	Réf. 3.a
حمض كربوسيل اندوسلفان	endosulfan acide carboxylique	Réf. 4.c
حمض –2 أمينو –3 هيدروكسي بوتان ثنائي أويك	acide 2-amino-3-hydroxybutanedioïque	Réf. 1.a
حموضة	acidité	Réf. 4.a
خاصية كيرالية	propriété chirale	Réf. 1.a
خاصية كيميائية	propriété chimique	Réf. 4.b
خليط	mélange	Réf. 4.b

خماسي كلورو البنزين	pentachlorobenzène	Réf. 4.b
خماسي كلورونترو البنزين	pentachloronitrobenzène	Réf. 4.b
دائرة المستخلصات الكيميائية	Chemical Abstract Service (CAS)	Réf. 4.b
درجة الإنصهار	température de fusion	Réf. 1.c
دلدرين	dieldrine	Réf. 4.b
دورق أحادي العنق	erlenmeyer monocol	Réf. 1.c
دي.دي.تي	DDT (dichloro-diphényle-trichloro-éthane)	Réf. 4.b
ديول اندوسلفان	Endosulfan diol	Réf. 4.c
ديوكسين	dioxine	Réf. 4.a
ذرة	atome	Réf. 3.a
ذرة الآزوت	atome d’azote	Réf. 3.a
ذرة الهيدروجين	atome d’hydrogène	Réf. 4.b
ذوبان	dissolution	Réf. 4.b
رابطة	liaison	Réf. 2.c
رابطة مضاعفة	liaison double	Réf. 2.c
راتنج	résine	Réf. 4.a
راسب	précipité	Réf. 1.b
رباعي إيتوكسيد التيتانيوم	tétraéthoxyde de titane	Réf. 1.c
رباعي كلورو البنزين	tétrachlorobenzène	Réf. 4.b
رباعي هيدروفوران	tétrahydrofurane	Réf. 1.b
رشّح	filtrer	Réf. 1.b
رصاص	plomb	Réf. 4.a
رقم التسجيل في سجل دائرة المستخلصات الكيميائية	numéro CAS (Chemical Abstract Service)	Réf. 4.b
زئبق	mercure	Réf. 4.a
زرنيخ	arsenic	Réf. 4.a
زنك	zinc	Réf. 4.a
زيولايت	zéolite	Réf. 4.a
سائل	liquide	Réf. 4.b
سائل هيدرولكي	liquide hydraulique	Réf. 4.a
سباعي الكلور	heptachlore	Réf. 4.b
سخّن	chauffer	Réf. 4.a
سداسي كلورو البنزين	hexachlorobenzène	Réf. 4.b
سداسي كلورو بنزوفينون	hexachlorobenzophénone	Réf. 4.b
سداسي كلورو حلقي الهكسان	hexachlorocyclohexane	Réf. 4.c
سرعة التفاعل	vitesse de réaction	Réf. 3.a

سعـة	capacité	Réf. 1.c
سكب	verser	Réf. 1.c
سلسلة كربونية	chaîne carbonée	Réf. 2.b
سلفونات	sulfonate	Réf. 4.c
سلفونيل	sulfonyle	Réf. 2.b
سلفونيل كلوريد	chlorure de sulfonyle	Réf. 2.b
سمية	toxicité	Réf. 4.b
سيريوم	cérium	Réf. 4.a
سيسكويتربين	sesquiterpène	Réf. 2.a
سيسكويتربين لاكتونيـة	lactone sesquiterpenique	Réf. 2.a
سيسيترتربيـن	sesterterpène	Réf. 2.a
سيـليـس	silice	Réf. 1.c
سيناريزن	cinnarizine	Réf. 4.b
شاردة	ion	Réf. 3.a
شاردة الأمونيوم	ion ammonium	Réf. 3.a
شكّل	constituer	Réf. 1.a
صوديوم	sodium	Réf. 1.b
صيغة اجمالية	formule brute	Réf. 2.c
صيغة مفصلة	formule développée	Réf. 2.c
طبقة عضوية	phase organique	Réf. 1.b
طبقة مائية	phase aqueuse	Réf. 1.b
طرد مركزي	centrifugation	Réf. 4.a
طريقة اصطناع	méthode de synthèse	Réf. 1.a
طريقة تحليلية	méthode d’analyse	Réf. 1.b
طريق التفاعل	méthode de réaction	Réf. 1.a
طريقة العمل	mode d’opératoire	Réf. 1.b
طـريق الاكسـدة	méthode d’oxydation	Réf. 1.a
طريقة ستورك	méthode de Stork	Réf. 1.a
طريقة للتفاعل الألدولي	méthode de la réaction aldolique	Réf. 1.a
ضـاف	ajouter	Réf. 1.b
ضـغط	pression	Réf. 1.b
ضـغط مرجع	pression de référence	Réf. 1.c
طولوين	toluène	Réf. 1.b
طيف	spectre	Réf. 1.b
طيف الاشعة تحت الحمراء	spectre des rayons infrarouges	Réf. 1.b
طيف الرنين النووي المغناطيسي	spectre de résonance magnétique nucléaire	Réf. 1.b

طيف الكتلة	Spectre de masse	Réf. 2.c
طيف الامتصاص	spectre d’absorption	Réf 1.b
طيف امتصاص الاشعة فوق البنفسجية	spectre d’absorption des rayons ultraviolets	Réf. 2.c
طيف امتصاص الاشعة تحت الحمراء	spectre d’absorption des rayons infrarouges	Réf 1.b
عمل مخبري	travail expérimental	Réf. 1.a
عملية التنقية	procédé de purification	Réf. 2.c
عملية الفصل	procédé de séparation	Réf. 2.c
عينة	échantillon	Réf. 1.a
غاز	gaz	Réf. 3.a
غاز النشادر	gaz ammoniac	Réf. 3.a
غسل	rincer	Réf. 1.b
غليسين	glycine	Réf. 1.b
غليسينات	glycinate	Réf. 1.a
غليسينات إلكتروفيلية	glycinate électrophile	Réf. 1.a
فسفرة	phosphorylation	Réf. 2.b
فسفور	phosphore	Réf. 4.c
فضة	argent	Réf. 4.a
فلافونيد	flavonoïde	Réf. 2.a
فلوريد	fluorure	Réf. 4.a
فلوريد الصوديوم	fluorure de sodium	Réf. 4.a
فلوريد متعدد الفينيليدين	Polyfluorure de vinylidène	Réf. 4.a
فوسفات	phosphate	Réf. 4.a
فينول	phénol	Réf. 4.a
فينول كلورية	chlorophénol	Réf. 4.a
فينيل	phényle	Réf. 4.a
قاس	mesurer	Réf. 1.b
قاعدة	base	Réf. 1.c
قاعدة شيف	base de Schiff	Réf. 1.a
قدّر	estimer	Réf. 1.a
قطّر	distiller	Réf. 1.b
قطرة	goutte	Réf. 1.c
قطرة بقطرة	goutte à goutte	Réf. 1.c
قمع	entonnoir	Réf. 1.c
قمع بروم	ampoule à brome	Réf. 1.c
قمع فصل	ampoule à décanter	Réf. 1.c
قانون هنري	loi de Henry	Réf. 4.b

كثافة	densité	Réf. 2.b
كادميوم	cadium	Réf. 4.a
كالسيوم	calcium	Réf. 1.b
كبريت	soufre	Réf. 4.a
كبريتات	sulfate	Réf. 1.b
كبريت الاندوسلفان	sulfate d'endosulfan	Réf. 4.c
كبريتات الصوديوم	sulfate de sodium	Réf. 1.b
كبريتات المغنسيوم	sulfate de magnésium	Réf. 1.c
كبريتيد	sulfure	Réf. 4.b
كحول	alcool	Réf. 1.c
كحول الاليلي	alcool allylique	Réf. 2.b
كحول ثالثي	alcool tertiare	Réf. 1.b
كربون	carbone	Réf. 1.c
كربون عضوي	carbone organique	Réf. 4.b
كربون منشط	charbon actif	Réf. 4.a
كربونات	carbonate	Réf. 4.a
كروماتوغرافيا	chromatographie	Réf. 1.b
كروماتوغرافيا الطبقة الرقيقة	Chromatographie sur Couche mince (CCM)	Réf. 2.c
كروماتوغرافيا الطبقة الرقيقة التحليلية	chromatographie sur couche mince analytique	Réf. 2.c
كروماتوغرافيا العمود	chromatographie sur colonne	Réf. 1.b
كلوبيريليد	clopyralide	Réf. 4.b
كلور	chlore	Réf. 1.a
كلور التالونيل	chlorothalonil	Réf. 4.b
كلور البنزين	chlorobenzène	Réf. 4.b
كلورة	chloration	Réf. 4.b
كلورة النتروبنزين	chloration du nitrobenzène	Réf. 4.b
كلوردان	chlordane	Réf. 4.b
كلوربيريفوس- ميثيل	chlorpyriphos-méthyl	Réf. 4.b
كلوروفينول	chlorophénol	Réf. 4.a
كلوروفورم	chloroforme	Réf. 1.b
كلوريد	chlorure	Réf. 1.c
كلوريد الثيونيل	chlorure de thionyle	Réf. 1.c
كلوريد الصوديوم	chlorure de sodium	Réf. 1.c
كلوريد الفضة	chlorure d'argent	Réf. 4.a
كلوريد الهيدروجين	chlorure d'hydrogène	Réf. 4.a

كيروسين	kérosène	Réf. 4.a
كينتوزين	quintozène	Réf. 4.b
كينون	quinone	Réf. 2.a
كينولين	quinoléine	Réf. 1.b
لاكتون	lactone	Réf. 4.c
لاكتون اندوسلفان	endosulfan lactone	Réf. 4.c
لزوجة	viscosité	Réf. 4.b
ماء	eau	Réf. 1.b
ماء مقطر	eau distillée	Réf. 1.b
مادة	matière	Réf. 1.d
مادة خام	matière première	Réf. 4.a
مادة خطرة	matière dangeureuse	Réf. 4.b
مادة سمية	matière toxique	Réf. 4.b
مادة عضوية	matière organique	Réf. 4.b
مادة عضوية مكلورة	matière organique chlorée	Réf. 4.b
مادة كيميائية	matière chimique	Réf. 4.b
مادة ماصة	matière absorbante	Réf. 4.a
مبيد	biocide	Réf. 4.b
مبيد الأعشاب	herbicide,	Réf. 4.b
مبيد الآفات	pesticide	Réf. 4.b
مبيد الفطريات	fongicide	Réf. 4.b
متعدد	multiplet	Réf. 2.c
متعدد الأستيلين	polyacetylène	Réf. 2.a
متفاعل	réactif	Réf. 1.b
مجموعة وظيفية	groupe fonctionnel	Réf. 1.a
مجموعة وظيفية كيميائية	groupe fonctionnel chimique	Réf. 1.a
مجموعة الاستر	groupe ester	Réf. 1.b
مجموعة الامين	groupe amine	Réf. 1.a
مجموعة الايمين	groupe imine	Réf. 1.c
مجموعة الكحول	groupe alcool	Réf. 1.c
مجموعة المثيل	groupe méthyl	Réf. 2.b
مجموعة الهيدروكسيل	groupe hydroxyle	Réf. 1.a
محلول	solution	Réf. 1.b
محلول مائي	solution aqueuse	Réf. 2.a
محلول مشبع	solution saturée	Réf. 1.c
مخبر	laboratoire	Réf. 1.a

مدة التفاعل	temps de réaction	Réf. 1.c
مذيب	solvant	Réf. 1.b
مذيب عضوي	solvant organiques	Réf. 4.a
مذيب عضوي مهلجن	solvant organique halogéné	Réf. 4.a
مذيب مكلور	solvant chloré	Réf. 4.b
مردود	rendement	Réf. 1.a
مركب	composé	Réf. 1.a
مركب ارجانوهلوجيني	composé organohalogéné	Réf. 4.a
مركب اروماتي	composé aromatique	Réf. 2.a
مركب متعدد البرومة	composé polybromé	Réf. 4.a
مركب عضوي	composé organique	Réf. 1.a
مركب كيرالـي	composé chiral	Réf. 1.a
مركب كحولي	composé alcoolique	Réf. 1.d
مركب محايد	composé neutre	Réf. 4.b
مزيج	mélange	Réf. 1.b
مستحلب	émulsion	Réf. 4.a
مشتق	dérivé	Réf. 1.a
مشتق بنزيلي	dérivé benzylique	Réf. 1.d
مشتق الاحماض الامينية	dérivé des acides aminés	Réf. 1.a
معادلة كيميائية	équation chimique	Réf. 2.a
معالجة المياه المستعملة	traitement des eaux usées	Réf. 4.b
معامل تفريق	coefficient de diffusion	Réf. 4.b
معدن	métal	Réf. 4.b
معقد	complexe	Réf. 1.c
معقد التيتانيوم	complexe de titane	Réf. 1.c
مغنسيوم	magnésium	Réf. 1.b
مفروق هكسادسنال	trans hexadécénal	Réf. 1.a
ملح	sel	Réf. 1.c
ملوث	polluant	Réf. 4.b
ملوث عضوي	polluant organique	Réf. 4.b
ملوث عضوي ثابت	polluant organique persistant	Réf. 4.b
مول	mole	Réf. 1.b
ميتوكسي	méthoxy	Réf. 1.c
ميتوكسيل	méthoxyle	Réf. 1.a
ميثان	méthane	Réf. 1.c
ميثانول	méthanol	Réf. 1.b

نترو	nitro	Réf. 1.a
نتروزامين ثنائي الإيثيل	diéthylnitrosamine	Réf. 4.b
ناتج	produit	Réf. 1.c
ناتج التفاعل	produit de la réaction	Réf. 1.c
نافتوكينون	naphthoquinone	Réf. 2.a
نترات	nitrate	Réf. 4.a
النتريت	nitrite	Réf. 4.a
نتروبنزين	nitrobenzène	Réf. 4.b
نتروجين	azote	Réf. 4.a
نحاس	cuivre	Réf. 4.b
نزع	éliminer	Réf. 4.a
نزع الماء	déshydratation	Réf. 4.a
نشادر	ammoniac	Réf. 3.a
نفثالين	naphtalène	Réf. 4.a
نفثالين متعدد الكلورة	polychloronaphtalène	Réf. 4.a
نفثالين مكلور	chloronaphtalène	Réf. 4.a
نفط	pétrole	Réf. 4.b
نوربورنين	norbornène	Réf. 4.c
نواة	noyau	Réf. 2.c
نواة الكربون	Noyau de carbone	Réf. 2.c
نواة ذرة الهيدروجين	noyau de l'atome d'hydrogène	Réf. 2.c
نيوبيوم	niobium	Réf. 4.b
وزن	masse	Réf. 1.b
وزن جاف	masse anhydre	Réf. 4.b
وسط حمضي	milieu acide	Réf. 2.b
وظيفة حمضية	fonction acide	Réf. 1.a
هالوجين	halogène	Réf. 4.a
هدرجة	hydrogénation	Réf. 1.a
هدرجة لامتناظرة	hydrogénation asymétrique	Réf. 1.a
هكسان	hexane	Réf. 1.b
هوية كيميائية	identité chimique	Réf. 4.b
هيدرور	hydrure	Réf. 1.b
هيدرور الكالسيوم	hydrure de calcium	Réf. 1.b
هيدروجين	hydrogène	Réf. 1.d
هيدروكربون	hydrocarbure	Réf. 4.a
هيدروكربون مهلجن	hydrocarbure halogéné	Réf. 4.a

هيدروكربون مهلجن عطري	hydrocarbure halogéné aromatique	Réf. 4.a
هيدروكسي	hydroxy	Réf. 1.a
هيدروكسيد	hydroxyde	Réf. 1.b
هيدروكسي إثير اندوسلفان	endosulfan hydroxyéther	Réf. 4.c
هيدروكسيد البوتاسيوم	hydroxyde de potassium	Réf. 1.b
هيدروكسيل	hydroxyle	Réf. 1.a
هيدروكلوريد	chlorhydrate	Réf. 1.c
هيدروكلوريد ايثيل استر الغليسين	chlorhydrate d'ester éthylique de glycine	Réf. 1.c
هيكل كربوني	squelette carboné	Réf. 2.b
هيكل كيميائي	structure chimique	Réf. 4.c
يود	iode	Réf. 1.b

Annexe 3 : Fiches terminologiques

Voici un échantillon des fiches terminologiques des termes de notre corpus. Chaque fiche contient le terme arabe en entrée avec sa transcription et son équivalent français. Nous indiquons sa classe et/ou une sous-classe ainsi que sa catégorie lexicale (nom, adjectif, verbe). Une définition est proposée et le contexte avec la référence du corpus sont précisés afin d'attester des termes. Si des variantes terminologiques sont identifiées, nous les citons. Nous indiquons également la formation construction et la modélisation du terme. Pour les UTC, nous indiquons la règle de grammaire associée. Cet échantillon de fiches est classé en suivant la classification adoptée (cf. 5.3 Classification adoptée).

- Fiches terminologiques des termes désignant un constituant de la matière

إيثانول = ʾiṯânûl = éthanol	
Classe, sous-classe	Composé chimique, fonction chimique, alcool
Définition	Alcool primaire, sous forme liquide incolore, d'odeur agréable, miscible à l'eau en toutes proportions et miscible à de nombreux solvants organiques
Catégorie lexicale	Nom
Construction syntaxique	Ø
Modélisation	UTS
Contextes	يقطر كل من الإيثانول والميثانول فوق المغنيزيوم واليود. (Réf 1.b)
Termes reliés sémantiquement / associés	كحول إيثيلي

بنزين = banzîn = benzène	
Classe, sous-classe	Composé chimique, hydrocarbure monocyclique, aromatique
Définition	Composé organique, sous forme de cristaux incolores, infiniment soluble dans l'éthanol, l'éther, l'acide acétique, l'acétone et le toluène.
Catégorie lexicale	Nom
Construction syntaxique	Ø
Modélisation	UTS
Contextes	نضيف 0.83 مكافئ من هيدروكسي بينانون في 77 مل من البنزين (Réf 1.c)
Termes reliés sémantiquement / associés	بنزن

تيتانيوم = tîtâniyûm = titane	
Classe, sous-classe	Élément chimique, métal de transition, solide
Définition	Élément chimique, qui a l'état stable, est sous forme de métal blanc à éclat métallique, possédant une résistance mécanique assez faible associée à une grande ductilité et une bonne résistance à la corrosion.
Catégorie lexicale	Nom
Construction syntaxique	Ø
Modélisation	UTS
Contexte	نضع 0.78 غ، 3.59 ميلي مول من معقد التيتانيوم. (Réf 1.b)
Terme relié sémantiquement / associés	تيتان

ثلاثي إيثيل أمين = *ṯulâṯî ’îṯîl ’amîn* = triéthylamine	
Classe, sous-classe	Composé chimique, fonction chimique, amine
Définition	Composé chimique, constitué d'une amine tertiaire symétrique et liquide à température ambiante.
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Nouvelle construction
Modélisation	UTC = UL + UTS + UTS
Règle de grammaire	UL + Terme (Composé chimique) + Terme (Composé chimique) → Terme (Composé chimique)
Contextes	نضيف 4.59 مل، 32.95 ميلي مول من ثلاثي إيثيل أمين. (Réf 1.c)
Termes reliés sémantiquement / associés	Ø

ثنائي كلوروميثان = *ṯunâ’î klûrûmîṯ*ân = dichlorométhane	
Classe, sous-classe	Composé chimique, hydrocarbure acyclique, alcane
Définition	Composé chimique, sous forme liquide incolore à température ambiante et utilisé comme solvant des composés organiques.
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Nouvelle construction
Modélisation	UTC = UL + UTS
Règle de grammaire	UL + Terme (Composé chimique) → Terme (Composé chimique)
Contextes	يقطر ثنائي كلوروميثان فوق هيدرور الكالسيوم. (Réf 1.b)
Termes reliés sémantiquement / associés	Ø

صوديوم= ṣûdyûm = sodium	
Classe, sous-classe	Elément chimique, métal alcalin, solide
Définition	Élément chimique, sous forme de métal à l'état stable, très réactif et qui tend à perdre son électron périphérique.
Catégorie lexicale	Nom
Construction syntaxique	Ø
Modélisation	UTS
Contexte	يجفف الإيثر فوق CaCl2 مدة 24 ساعة ثم يقطر فوق الصوديوم. (Réf 1.c)
Terme relié sémantiquement / associés	Ø

كبريتات الصوديوم = kibrîtât alṣûdyûm = sulfate de sodium	
Classe, sous-classe	Espèce chimique, composé ionique
Définition	Composé ionique, formé d'un ion sulfate et de deux ions sodium. Utilisé principalement dans la fabrication de détergents et dans le procédé kraft ainsi que le traitement de la pâte à papier.
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Expansion d'annexion
Modélisation	UTC = UTS + UTS
Règle grammaire	Terme (Composé chimique) + Terme (Elément chimique)Ann → Terme (Composé chimique)Ann
Contextes	تجمع الطبقات العضوية و تجفف فوق كبريتات الصوديوم $Na2SO_4$. (Réf 1.b)
Termes reliés sémantiquement / associés	كبريتات

هيدروكلوريد إيثيل إستر الغليسين = chlorhydrate d'ester éthylique de glycine	
Classe, sous-classe	Espèce chimique
Définition	Composé chimique, sous forme de cristal blanc
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Nouvelle construction
Modélisation	UTC = UTS + UTS + UTS + UTS
Règle grammaire	Terme (Composé chimique) + Terme (Composé chimique) + Terme (Composé chimique) + Terme (Composé chimique)Ann → Terme (Composé chimique)Ann
Contextes	في ثنائي عنق ذو سعة 250 مل (4.6غ، 32.95 ميلي مول) من هيدروكلورور إيثيل إستر الغليسين. (Réf 1.c)
Termes reliés sémantiquement / associés	Ø

- Fiches terminologiques des termes désignant des réactions chimiques

ارجاع = ʾirjâʿ = réduction	
Classe, sous-classe	Réaction chimique, oxydo-réduction
Définition	Réaction chimique par laquelle un (ou plusieurs) atome(s) d'une molécule ou d'un ion gagne(nt) un (ou plusieurs) électron(s).
Catégorie lexicale	Nom
Construction syntaxique	Ø
Modélisation	UTS
Contexte	و التي تؤدي بعد إرجاع الوظيفة الحمضية إلى آمينات ثنائية الهيدروكسيل. (Réf 1.a)
Termes reliés sémantiquement / associés	اختزال

تأكسد = ʾaksada = s'oxyder			
Classe, sous-classe	Verbe spécialisé		
Définition	Être transformé (un corps simple ou composé) en oxyde, par combinaison avec l'oxygène.		
Catégorie lexicale	Verbe		
Structure argumentale	UTS + UTC		
Réalisations linguistiques des arguments	المواد العضوية عالية الذوبانية في الماء خماسي كلور البنزين	بالضوء	في الجو
Contextes	تصبح المواد العضوية عالية الذوبانية في الماء وتتأكسد لتنتج ثاني أكسيد الكربون والماء وأحماض أو أملاح غير عضوية. (Réf 4.a)		
	قد يتأكسد خماسي كلور البنزين بالضوء في الجو . (Réf 4.b)		
Termes reliés sémantiquement / associés	اكسدة, اكسد		
Modélisation	UTS (Verbe spécialisé) + UTC (Composé chimique)		
Règle grammaire	Terme (Verbe spécialisé) + UL + Terme (Composé chimique) + Terme (Composé chimique)Ann → Terme (Verbe spécialisé)		

تكـاثف الدولي = takâ*ṯuf ʾaldûlî* = condensation aldolique	
Classe, sous-classe	Réaction chimique, condensation
Définition	Réaction chimique, permettant la formation de liaisons carbone-carbone.
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Expansion d'identification
Modélisation	UTC = UTS + UTS
Règle grammaire	Terme (Réaction chimique) + Terme (Composé chimique)Ann → Terme (Réaction chimique)Ann
Contextes	لقد تم إصطناع عدة مركبات كيميائية (أحماض آمينية و مشتقاتها) طبيعية و غير طبيعية بالإعتماد على إستراتيجيـة التكـاثف الألدولي لقواعد شيف. (Réf 1.b)
Termes reliés sémantiquement / associés	تكـاثف

- Fiches terminologiques des termes désignant des méthodes d'analyse

كروماتوغرافيا الطبقة الرقيقة التحليلية = *krûmâtûġâfiyâ alṭ*abaqat alraqîqat alta*ḥ*lîliyat = chromatographie sur couche mince analytique	
Classe, sous-classe	Méthode d'analyse, chromatographie
Définition	Technique de chromatographie dont la phase mobile est liquide, couramment utilisé pour séparer des composants dans un but d'analyse
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Expansion d'annexion + expansion d'identification
Modélisation	UTC = UTS + UTS + UTS + UTS
Règle grammaire	Terme (Méthode d'analyse) + UL+ UL + Terme (Méthode d'analyse) → Terme (Méthode d'analyse)
Contextes	جرت عملية التمليص بواسطة خليط من (50:50) toluène/ acétate d'éthyle مع مراقبة مستمرة للكسور بواسطة كروماتوغرافيا الطبقة الرقيقة التحليلية. (Réf 2.c)
Termes reliés sémantiquement / associés	كروماتوغرافيا

كروماتوغرافيا العمود = *krûmâtûġâfiyâ al*ʿamûd = chromatographie sur colonne	
Classe, sous-classe	Méthode d'analyse, chromatographie
Définition	Technique de chromatographie où la silice se trouve dans une colonne, très utilisée dans la purification en chimie organique, dans le but de séparer les différents composants d'un produit mais aussi de purifier le produit d'une réaction.
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Expansion d'annexion
Modélisation	UTC = UTS + UTS
Règle grammaire	Terme (Méthode d'analyse) + UL → Terme (Méthode d'analyse)
Contextes	قمنا بتنقية المركب الناتج باستخدام طريقتين: التقطير عند ضغط مرجع و كروماتوغرافيا العمود. (Réf 1.b)
Termes reliés sémantiquement / associés	كروماتوغرافيا

- Fiches terminologiques des termes désignant des procédures chimiques

ترشيح = *taršîḥ* = filtration	
Classe, sous-classe	Procédure chimique
Définition	Procédé de séparation permettant de séparer les constituants d'un mélange qui possède une phase liquide et une phase solide au travers d'un milieu poreux.
Catégorie lexicale	Nom
Construction syntaxique	Ø
Modélisation	UTS
Contextes	نقوم بترشيح الناتج باستعمال ثنائي إيثيل الإيثر. (Réf 1.b)
Termes reliés sémantiquement / associés	رشّح

غسّل= ġassala = rincer			
Classe, sous-classe	Verbe polysémique dont au moins un sens est spécialisé		
Définition	Passer quelque chose à l'eau ou dans un autre liquide, pour le nettoyer rapidement ou pour le débarrasser d'un produit spécifique de lavage.		
Catégorie lexicale	Verbe		
Structure argumentale	UTS + UTC + Prép + UTS		
Réalisations linguistiques des arguments	الراسب المتحصل عليه	بعد الترشيح جيدا	بالإيثر
Contextes	يغسل الراسب المتحصل عليه بعد الترشيح جيدا بالإيثر (Réf 1.b)		
Termes reliés sémantiquement / associés	تغسيل		
Modélisation	UTS (Verbe polysémique) + UTC (Composé chimique) + prép + UTS (Composé chimique)		
Règle grammaire	Terme (Verbe polysémique) + Terme (Composé chimique) + prép + Terme (Composé chimique) → Terme (Verbe polysémique)		

قطّر= qaṭṭara = distiller			
Classe, sous-classe	Verbe très spécialisé		
Définition	Une Entité Chimique est filtrée		
Catégorie lexicale	Verbe		
Structure argumentale	UTS + Prép + UTS UTS + UTC + Prép + UTC UTS + UTC + Prép + UTS + Prép + UL		
Réalisations linguistiques des arguments	ثنائي كلوروميثان كلورور الأسيل	فوق: الصوديوم فوق : هيدرور الكالسيوم فوق : الكينولين	بنسبة 1:10
Contextes	ثم يقطر فوق الصوديوم (Réf 1.b)		
	يقطر ثنائي كلوروميثان فوق هيدرور الكالسيوم (Réf 1.b)		
	يقطر كلورور الأسيل فوق الكينولين بنسبة 1:10 (Réf 1.b)		
Termes reliés sémantiquement / associés	تقطير		
Modélisation	UTS (Verbe spécialisé) + Prép + UTS (Élément chimique) UTS (Verbe spécialisé) + UTC (Composé chimique) + Prép + UTC (Composé chimique) UTS (Verbe spécialisé) + UTC (Composé chimique) + Prép + UTS (Composé chimique) + Prép + UL		
Règle grammaire	Terme (Verbe spécialisé) + prép + Terme (Élément chimique) → Terme (Verbe spécialisé) Terme (Verbe spécialisé) + UL + Terme (Composé chimique) + prép + Terme (Composé chimique) + Terme (Élément chimique)Ann → Terme (Verbe spécialisé) Terme (Verbe spécialisé) + Terme (Composé chimique) + Terme (Composé chimique) + prép + Terme (Composé chimique) + prép + UL → Terme (Verbe spécialisé)		

قطرة بقطرة = qaṭrat biqaṭrat = goutte à goutte	
Classe, sous-classe	Procédure chimique
Définition	Procédé de distribution de produits chimiques permettant sa régulation.
Catégorie lexicale	Syntagme adverbial
Construction syntaxique	Expansion complétive
Modélisation	UTC = UTS + prép + UTS
Règle grammaire	Terme (Composé chimique) + préposition + Terme (Composé chimique) → Terme (Procédure chimique)
Contextes	ثم تتم إضافة مكافئين من ن-بوتانال (0.63 مل، 0.51 غ، 7.1 ميلي مول) قطرة بقطرة. (Réf 1.c)
Termes reliés sémantiquement / associés	قطرة

- Fiches terminologiques des termes désignant de la verrerie

ثلاثي عنق = ṯulâṯî ʿunq = tricol	
Classe, sous-classe	Equipement, verrerie
Définition	Récipient largement utilisé en verrerie de laboratoire, constitué de trois cols.
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Expansion d'annexion
Modélisation	UTC = UL + UTS
Règle grammaire	UL + Terme (Verrerie) → Terme (Verrerie)
Contextes	في ثلاثي عنق ذو سعة 100 مل، نضع 0.78 غ. (Réf 1.c)
Termes reliés sémantiquement / associés	عنق - ثنائي عنق

ثنائي عنق = ṯunâʾî ʿunq = bicol	
Classe, sous-classe	Equipement, verrerie
Définition	Récipient largement utilisé en verrerie de laboratoire, constitué de deux cols.
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Expansion d'annexion
Modélisation	UTC = UL + UTS
Règle grammaire	UL + Terme (Verrerie) → Terme (Verrerie)
Contextes	في ثنائي عنق ذو سعة 500 مل، نضع 20.96 غ. (Réf 1.b)
Termes reliés sémantiquement / associés	عنق - ثلاثي عنق

دورق أحادي العنق = dawraq ʾuḥâdî alʿunq = erlenmeyer monocol	
Classe, sous-classe	Equipement, verrerie
Définition	Récipient largement utilisé en verrerie de laboratoire, constitué d'une base conique et d'un seul col cylindrique.
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Expansion d'identification + Expansion d'annexion
Modélisation	UTC = UTS + UL + UTS
Règle grammaire	Terme (Verrerie) + UL + Terme (Verrerie)Ann → Terme (Verrerie)Ann
Contextes	بعد انتهاء مدة التفاعل نسكب الناتج في دورق أحادي العنق . (Réf 1.c)
Termes reliés sémantiquement / associés	دورق

قمع فصل = qimʿ faṣl = ampoule à décanter	
Classe, sous-classe	Équipement, Verrerie
Définition	Élément de verrerie de laboratoire, utilisé pour séparer par décantation deux liquides non-miscibles pour effectuer une extraction liquide-liquide.
Catégorie lexicale	Syntagme nominal
Construction syntaxique	Expansion d'identification
Modélisation	Terme (Verrerie) + Terme (Procédure chimique) → Terme (Verrerie)
Règle grammaire	Terme (Méthode d'analyse) + UL+ UL + Terme (Méthode d'analyse) → Terme (Méthode d'analyse)
Contextes	نقوم بتحويل الناتج إلى قمع فصل. (Réf 1.c)
Termes reliés sémantiquement / associés	قمع

Annexe 4 : Programme de chimie

Voici quelques exemples de sommaire de programme de chimie de l'enseignement secondaire et universitaire, que nous avons consulté pour concevoir notre classification :

- « العلوم الفيزيائية والتكنولوجية السنة الرابعة من التعليم المتوسط », manuel scolaire de l'enseignement secondaire en Algérie de 2016-2017. Voici le programme du livre avec la partie concernant la chimie :

الفهرس

الظواهر الميكانيكية

الظواهر الكهربائية

المادة و تحولاتها

الظواهر الضوئية

التكنولوجيا و الإعلام و الإتصال

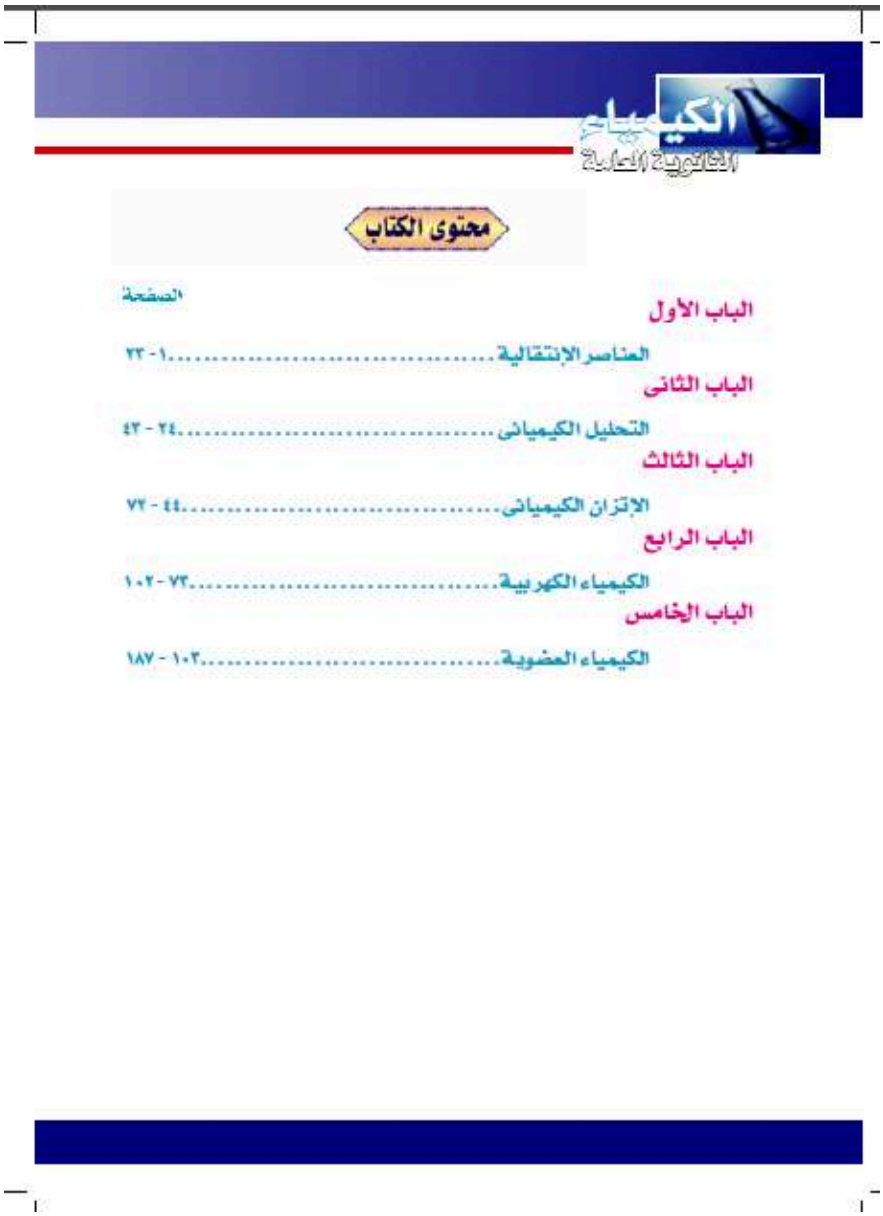
- « الكاتب العلوم الفيزيائية للسنة الثالني من التعليم الثانوي », manuel scolaire de l'enseignement secondaire en Algérie de 2017-2018. Voici le programme du livre concernant la chimie :

مجال الظواهر الضوئية	
الوحدة1 : العدسات عناصر لعدة أجهزة بصرية	
1- نشاطات أولية	196
2- خلاصة الدراسة	197
الوحدة 2 : الصورة المعطاة من طرف عدسة	
1- العدسات المقربة	202
2- العدسات المبعدة	205
الوحدة 3 : نمذجة العدسة المقربة	
1- شكل الحزمة الضوئية النافذة من عدسة مقربة	211
2- نموذج العدسة الرقيقة و التمثيل البياني للأشعة	212
3- تحديد بياني لنقطة صورة موافقة لنقطة جسم	213
4- مفهومي الجسم الحقيقي و الجسم الوهمي	215
الوحدة 4 : الضوء و الحياة اليومية	
1- مفهوم التضخيم في الأجهزة البصرية	224
2- المكبرة	225
3- المجهر	226
4- المنظار الفلكي	227
5- الرؤية و عيوب البصر	228

مجال المادة وتحولاتها	
الوحدة1 : نموذج الغاز المثالي	
- طريقة لتعيين كمية المادة في الحالة الغازية	
1- مفهوم ضغط الغاز و قياسه	238
2- مفهوم درجة الحرارة و قياسها	243
3- دراسة العوامل المؤثرة في الغاز	245
4- نموذج الغاز المثالي	250
الوحدة 2 : قياس الناقلية	
- طريقة لتعيين كمية المادة في المحاليل الشاردية	
1- المحاليل المائية	261
2- النقل الكهربائي للمحاليل الشاردية	267
الوحدة 3 : تعيين كمية المادة بواسطة المعايرة	
1- التفاعل بين المحاليل الحمضية و الأساسية	285
2- تفاعلات الأكسدة و الإرجاع	296
الوحدة 4 : مدخل إلى الكيمياء العضوية	
1- مدخل للكيمياء العضوية	312
2- الفحوم الهيدروجينية	316
3- التسمية حسب توصيات IUPAC	322
4- تأثير السلسلة الفحمية على الخصائص الفيزيائية	326
5 - المرور من مجموعة مميزة إلى أخرى	334
6- البترول	337

- « الكيمياء للصف الثالث ثانوي »[460], manuel scolaire de l'enseignement secondaire en Egypte de 2017-2018. Voici le programme du livre :

[460] https://egyyfast.blogspot.fr/2017/08/chemistry-tbook-third-secondary-2018.html



- « أسس الكيمياء العضوية », ouvrage universitaire libyen (Wael Ghaleb Mohamed & Walid Mohamed Alsayti, 2008). Voici le sommaire du livre :

المحتويات

الفصل الثالث

الفصل الرابع / الألدهيدات والكيتونات

الفصل الخامس / الأحماض الكربوكسيلية

الفصل السادس / مشتقات الأحماض الكربوكسيلية

الفصل السابع / الأمينات ومشتقاتها

Annexe 5 : Règles de grammaire

Voici les règles de grammaire des termes de notre corpus.

UL + préposition + Terme (Composé Chimique) → Terme (Procédure chimique)

UL + Terme (Composé chimique) + Terme (Composé chimique) → Terme (Composé chimique)

UL + Terme (Composé chimique) → Terme (Composé chimique)

UL + Terme (Physique)Ann + Terme (Réaction Chimique)Adj + Terme (Physique)Adj → Terme (Méthode d'analyse)Ann

UL + Terme (Méthode d'analyse)Adj → Terme (Méthode d'analyse)Adj

UL + Terme (Méthode d'analyse)Ann → Terme (Méthode d'analyse)Ann

UL + Terme (Procédure chimique)Adj → Terme (Méthode d'analyse)Adj

UL + Terme (Procédure chimique)Ann → Terme (Méthode d'analyse)Ann

UL + Terme (Réaction Chimique)Adj → Terme (Méthode d'analyse)Adj

UL + Terme (Réaction Chimique)Ann → Terme (Méthode d'analyse)Ann

UL + Terme (Verrerie)Adj → Terme (Verrerie)Adj

UL + Terme (Verrerie)Ann → Terme (Verrerie)Ann

Terme (Elément chimique) + Terme (Elément chimique)Adj → Terme (Composé chimique)Adj

Terme (Elément chimique) + Terme (Elément chimique)Ann → Terme (Composé chimique)Ann

Terme (Composé chimique) + préposition + Terme (Composé chimique) → Terme (Procédure chimique)

Terme (Composé chimique) + préposition + Terme (Composé chimique) + Terme (Composé Chimique)Adj → Terme (Composé chimique)

Terme (Composé chimique) + préposition + Terme (Composé chimique) + Terme (Composé Chimique)Ann → Terme (Composé chimique)

Terme (Composé chimique) + Terme (Elément chimique)Adj → Terme (Composé chimique)Adj

Terme (Composé chimique) + Terme (Elément chimique)Ann → Terme (Composé chimique)Ann

Terme (Composé chimique) + Terme (Composé chimique)Adj → Terme (Composé chimique)Adj

Terme (Composé chimique) + Terme (Composé chimique)Ann → Terme (Composé chimique)Ann

Terme (Composé chimique) + Terme (Composé chimique) + Terme (Composé chimique) + Terme (Composé chimique)Adj → Terme (Composé chimique)Adj

Terme (Composé chimique) + Terme (Composé chimique) + Terme (Composé chimique) + Terme (Composé chimique)Ann → Terme (Composé chimique)Ann

Terme (Composé chimique) + Terme (Réaction chimique)Adj → Terme (Composé chimique)Adj

Terme (Composé chimique) + Terme (Réaction chimique)Ann → Terme (Composé chimique)Ann

Terme (Méthode d'analyse) + UL → Terme (Méthode d'analyse)

Terme (Méthode d'analyse) + UL + UL + Terme (Méthode d'analyse) → Terme (Méthode d'analyse)

Terme (Méthode d'analyse) + Terme (Composé chimique)Adj → Terme (Méthode d'analyse)Adj

Terme (Méthode d'analyse) + Terme (Composé chimique)Ann → Terme (Méthode d'analyse)Ann

Terme (Méthode d'analyse) + Terme (Composé chimique)Ann + Terme (Composé chimique)Adj → Terme (Méthode d'analyse)Ann

Terme (Procédure chimique) + Terme (Composé chimique)Adj → Terme (Procédure chimique)Adj

Terme (Procédure chimique) + Terme (Composé chimique)Ann → Terme (Procédure chimique)Ann

Terme (Procédure chimique) + UL + Terme (Procédure chimique)Ann + Terme (Physique)Ann → Terme (Procédure chimique)Ann

Terme (Procédure chimique) + Terme (Physique)Adj → Terme (Procédure chimique)Adj

Terme (Procédure chimique) + Terme (Physique)Ann → Terme (Procédure chimique)Ann

Terme (Procédure chimique) + Terme (Composé chimique)Ann + Terme (Composé chimique)Adj → Terme (Procédure chimique)Ann

Terme (Réaction chimique) + Terme (Composé chimique)Adj → Terme (Réaction chimique)Adj

Terme (Réaction chimique) + Terme (Composé chimique)Ann → Terme (Réaction chimique)Ann

Terme (Verbe énonciateur) + UL + Terme (Composé chimique)Ann → Terme (Verbe énonciateur

Terme (Verbe énonciateur) + UL +Terme (Réaction chimique) + Terme (Composé Chimique) → Terme (Verbe énonciateur)

Terme (Verbe polysémique) + UL + Terme (Composé chimique) + Terme (Composé chimique)Adj → Terme (Verbe polysémique)

Terme (Verbe polysémique) + UL + Terme (Composé chimique) + Terme (Composé chimique)Ann → Terme (Verbe polysémique)

Terme (Verbe polysémique) + UL + Terme (Composé chimique) + prép + Terme (Composé chimique) + Terme (Composé chimique)Adj → Terme (Verbe polysémique)

Terme (Verbe polysémique) + UL + Terme (Composé chimique) + prép + Terme (Composé chimique) + Terme (Composé chimique)Ann → Terme (Verbe polysémique)

Terme (Verbe polysémique) + prép + Terme (Élément chimique) → Terme (Verbe polysémique)

Terme (Verbe polysémique) + Terme (Élément chimique) + prép + Terme (Élément chimique) → Terme (Verbe polysémique)

Terme (Verbe spécialisé) + UL + Terme (Composé chimique) + Terme (Composé chimique)Adj → Terme (Verbe spécialisé)

Terme (Verbe spécialisé) + UL + Terme (Composé chimique) + Terme (Composé chimique)Ann → Terme (Verbe spécialisé)

Terme (Verbe spécialisé) + UL + Terme (Composé chimique) + Terme (Réaction chimique) + Terme (Composé chimique)Adj → Terme (Verbe spécialisé)

Terme (Verbe spécialisé) + UL + Terme (Composé chimique) + Terme (Réaction chimique) + Terme (Composé chimique)Ann → Terme (Verbe spécialisé)

Terme (Verbe spécialisé) + prép + Terme (Élément chimique) → Terme (Verbe spécialisé)

Terme (Verbe spécialisé) + prép + Terme (Composé chimique) + Terme (Composé chimique)Adj → Terme (Verbe spécialisé)

Terme (Verbe spécialisé) + prép + Terme (Composé chimique) + Terme (Composé chimique)Ann → Terme (Verbe spécialisé)

Terme (Verbe spécialisé) + UL + Terme (Composé chimique) + prép + Terme (Composé chimique) + Terme (Élément chimique)Ann → Terme (Verbe spécialisé)

Terme (Verbe spécialisé) + Terme (Composé chimique) + Terme (Composé chimique) + prép + Terme (Composé chimique) + prép + UL → Terme (Verbe spécialisé)

Terme (Verrerie) + Terme (Procédure chimique)Adj → Terme (Verrerie)Adj

Terme (Verrerie) + Terme (Procédure chimique)Ann → Terme (Verrerie)Ann

Terme (Verrerie) + UL + Terme (Verrerie)Adj → Terme (Verrerie)Ann

Terme (Verrerie) + UL + Terme (Verrerie)Ann → Terme (Verrerie)Ann

Bibliographie

Abbes, R. 2004. *La conception et la réalisation d'un concordancier électronique pour l'arabe,* Thèse, Université de Lyon ENSSIB/INSA.

Abbès, R. et Dichy, J. 2008. « Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1, JADT 2008, Actes des 9èmes *Journées internationales d'Analyse statistique des* Données Textuelles, Presses Universitaires de Lyon, Lyon, p. 31-44.

Abdul Hay, A. 2012. *Constitution d'une ressource sémantique arabe à partir de corpus* multilingue aligné, Thèse, Université de Grenoble.

Abdul Hay, A. et Kraif, O. 2013. « Constitution d'une ressource sémantique arabe à partir de corpus multilingue aligné », TALN-RÉCITAL 2013 les S*ables d'Olonne,* 299-312.

Abi Ghanem-Chadarevian, C. 2016. « Socioterminologie et interactions langagières en arabe », Repères DoRiF n.10 - Le terme : un produit social ? DoRiF Università, Rome.

Albeiriss, B. 2016. « Étude Terminologique de la Nomenclature de la Chimie en Arabe dans une Approche de Fouille de Textes », Les discours spécialisés : enjeux, descriptions et pratiques, Actes du colloque international de l'Institut Supérieur des Langues de Gabès (ISLG), Gabès (à paraître).

Albeiriss, B. 2017. « Modélisation des unités lexicales complexes d'une langue spécialisée : le cas de la chimie en arabe », Le Lexique entre Langue et Discours, Actes du VIIe colloque international de lexicologie, Association de la Lexicologie Arabe en Tunisie, Tunis (à paraître).

Albeiriss, B. 2017. « Terminologie basée sur corpus : les verbes de la chimie en arabe », TOTh 2017 : Terminologie & Ontologie : Théories et Applications, Chambéry. 95-105 (à paraître).

Albeiriss, B. 2017. « Conception d'une ontologie à partir d'une étude terminologique basée sur corpus : le cas de la chimie en arabe », CiDE.20 : Le Document ? Europia, Lyon. 241-250.

Alrahabi, M. et Dichy, J. 2009. « Levée d'ambigüité par la méthode d'exploration contextuelle : la séquence 'alif-nûn (ان) en arabe », 2ème Conférence Internationale (SIIE 2009), Ghenima, M., Ouksel, A. et Sidhom, S. (eds.), *Systèmes d'Information et Intelligence Economique*, unis, Hammamet, IHE éditions, 573-585.

Ammar S. et Dichy J. 1999. Les verbes arabes, Paris, Hatier (coll. Bescherelle).

Anawati, G-C. 1997. « L'alchimie arabe », Histoire des sciences arabes : Technologie, alchimie et sciences de la vie, Volume 3, sous la direction de Roshdi Rashed, 111-140, Seuil.

Arbach, A. et Ali, S. 2013. « Aspects théoriques et méthodologiques de la représentativité des corpus », Corela [En ligne], HS-13 | 2013.

Assal, A. 1994. « La métaphorisation terminologique », Terminologie et Traduction, Vol 2, 235-242.

Atwell, E., Al-Sulaiti, L., Al-Osaimi, S. et Abu Shawar B. 2004. « A Review of Arabic Corpus Analysis Tools-Un Examen d'Outils pour l'Analyse de Corpus Arabes », JEP-TALN 2004, Arabic Language Processing, Fez, 19-22.

Aussenac-Gilles, N. 2005. *Méthodes ascendantes pour l'ingénierie des connaissances,* Habilitation à diriger les recherches, Université de Toulouse.

Aussenac-Gilles, N. et Séguéla, P. 2002. « Les relations sémantiques : du linguistique au formel », *Cahiers de Grammaire 25 'Sémantique et Corpus',* 175-198.

Aussenac-Gilles N., Biébow B. & Szulman S. « D'une méthode à un guide pratique de modélisation des connaissances à partir de textes », Actes des 5es journées Terminologie et Intelligence Artificielle (TIA 2003), Strasbourg. Ed. F. Rousselot, ENSAIS, 41-53.

Awad, D. 2013. La ponctuation arabe : Histoire et règles - Étude contrastive avec le français et l'anglais, Thèse, Université Lumière-Lyon 2.

Bachimont, B. 2000. « Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances », Ingénierie des connaissances : évolutions récentes et nouveaux défis, Eyrolles, 305-323.

Bagge, C. 1999. « Analyse sémantique comparative des vocabulaires scientifiques anglais et français », Meta XXVIII, n°4, 391-407.

Barna, C. 2014. Divergences et convergences dans la terminologie médicale vétérinaire pour les vertébrés domestiques entre le roumain et le français, Thèse, Université Paris 3.

Beesley, K. 1996. « Arabic Finite-State Morphological Analysis and Generation », Proceedings of COLING, Copenhagen, 89-94.

Beesley, K. 2001. « Finate-State Morphological Analysis and Generation of Arabic at Xerox Research : Status and Plans in 2001 », Actes du Arabic NLP Workshop at ACL/EACL.

Belguith Hadrich, L., Baccour, L et Mourad, G. 2005. « Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules », Actes de TALN 2005, 451-456.

Belguith Hadrich, L. et Chaâben, N. 2006. « Analyse et désambiguïsation morphologiques de textes arabes non voyellés », Actes de la 13ème édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006), 493-501.

Bendaoud, R., Hacene, M-H., Toussaint, Y., Delecroix, B. et Napoli, A. 2007. Construction d'une ontologie à partir d'un corpus de textes avec l'ACF

Benveniste, E. 1989. Problèmes de linguistique générale, Paris, Gallimard.

Berland, S. et Grabar, N. 2002. « Assistance automatique pour l'homogénéisation d'un corpus Web de spécialité », JADT 2002 : 6es *Journées internationales d'Analyse statistique des* données textuelles.

Berthelot, M. 1885. Les origines de l'alchimie, G. Steinheil, Paris, 445p.

Berthelot, M. 1889. Introduction à l'étude de la chimie des anciens et du moyen âge, Culture et civilisation, Bruxelles, 342p.

Bertels, A. 2009. « Etudier la sémantique des termes techniques : des théories à la pratique », Actes de la 8ème conférence internationale Terminologie et Intelligence Artificielle (TIA 2009). Terminologie et Intelligence Artificielle, Toulouse.

Bianchini, L., Rossi, M. et Mabrour, A. 2008. « Les mots de l'eau : entre terminologie spécialisée et analyse interculturelle », Synergies, n° 4, 123-132.

Biber, D. 1994. « Representativeness in corpus design », Linguistica Computazionale, IX-X, 377-408.

Binon, J. et Verlinde, S. 2000. « Les langue(s) de spécialité(s) : mythe ou réalité ? Lexicographie et langue(s) de spécialité(s) », Des mots aux dictionnaires : travaux de la section Lexicologie, lexicographie, onomastique, toponymie, Actes du XXIIe Congrès international de linguistique et philologie romanes, Bruxelles, Tübingen : Niemeyer, 616-628.

Bonnet, V. 2001. La construction d'une langue savante en Europe du Ve au XIXe siècle : le latin et le grec dans les sciences, Thèse, Université de Lyon 2.

Bourigault, D. 1996. « Conception et exploitation d'un logiciel d'extraction de termes : problèmes théoriques et méthodologiques », Clas, Thoiron et Béjoint (dir.), Lexicomatique et dictionnairiques, Actes des IVèmes journées du réseau thématique 'Lexicologie, Terminologie et Traduction', Aupelf-Uref, Montréal, 137-145.

Bourigault, D., Slodzian, M. 1999. « Pour une terminologie textuelle », Terminologies nouvelles, 19, 29-32.

Bourigault D., Aussenac-Gilles, N., Charlet, J. 2004, « Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas », *Revue d'Intelligence Artificielle (RIA) : 'Techniques Informatiques et structuration de terminologies'*, PIERREL J.-M. et SLODZIAN M. (Ed.), Paris : Hermès, 18 (1), 87-110.

Bourigault, D., Fabre, C., Frerot, C., Jacques, M-P. et Ozdowska, S. 2005. « Syntex, analyseur syntaxique de corpus », Actes de TALN 2005, Dourdan, France.

Bourigault, D. 2007. Un analyseur syntaxique opérationnel : SYNTEX ». Mémoire d'habilitation à diriger des recherches, Université de Toulouse 2.

Braham, A. 1998. « Remarques sur la constitution de la base de données lexicale de l'arabe DIINAR.1 », Journée de l'ATALA : corpus textuels étendus et base de données lexicales en arabe.

Braham, A., Dichy, J., Ghazali, S. et Hassoun, M. 2002. « La base de connaissances linguistiques DIINAR.1 (DIctionnaire INformatisé de l'Arabe, version 1) », Braham A. (ed), Actes de la conférence internationale sur le Traitement automatique de l'arabe, Proceedings of the International Symposium on The Processing of Arabic, Tunis : Université de La Manouba, 45-56.

Budin, G. 2007. « L'apport de la philosophie autrichienne au développement de la théorie de la terminologie : ontologie, théories de la connaissance et de l'objet », 11-23.

Cabré, M-T. 1998. La terminologie - Théorie, méthode et applications, Ottawa : Les Presses de l'Université d'Ottawa.

Cabré, M-T. 2000. « Terminologie et linguistique : la théorie des portes », Terminologies nouvelles, 2, 10-15.

Cabré, M-T. 2007. « La terminologie, une discipline en évolution : le passé, le présent et quelques éléments prospectifs », L'Homme, M-C. ; Vandaele, S. (dir.). Lexicographie et terminologie : compatibilité des modèles et des méthodes, Ottawa : Les Presses de l'Université d'Ottawa, 79-109.

Cabré, M-T. 2008. « Constituer un corpus de textes de spécialité », Les Cahiers du Cel. Paris : UFR d'Études Interculturelles de Langues Appliquées, 37-56.

Candel D. 1995. « Locutions et langues de spécialité », Martins-Baltar ed.

Cappeau, P. et Gadet, F. 2007. « L'exploitation sociolinguistique des grands corpus », Revue française de linguistique appliquée, 12 (1), 99-110.

Catach, N. 1994. La ponctuation. Histoire et système, Paris, PUF, 128 p.

Champclaux, Y. 2009. *Un modèle de recherche d'information basé sur les graphes et l*es *similarités structurelles pour l'amélioration du processus de recherche d'information*, Thèse, Université de Toulouse.

Chukwu, U. 1993. Le repérage des termes dans un corpus bilingue anglais/français, Thèse, Université de Lyon 2.

Chukwu, U. 1998. « Dépouillement de corpus à des fins terminologiques dans un univers dépendant du temps », Meta 433, 411-425.

Cohen, D. 1970. Etudes de linguistique sémitique et arabe, Paris : Mouton, 49-78.

Condamines, A. 1999. « Alternance nom/verbe : explorations en corpus spécialisé. B. Victorri et J. François (eds) : Sémantique du lexique verbal, Actes de l'atelier de Caen », Cahiers de *l'Elsap*, 41-48.

Condamines, A. 2005. « Linguistique de corpus et terminologie », Langages, n°157, 36-47.

Collet, T. 2000. La réduction des unités terminologiques complexes de type syntagmatique, Thèse, Université de Montréal.

Corbeil, J-C. 1999. La terminologie : une discipline au service d'objectifs multiples.

Corbeil, J-C. 2007. « Le rôle de la terminologie en aménagement linguistique : genèse et description de l'approche québécoise », Gerhard Budin, Jean-Claude Corbeil, Loïc Depecker [et al] : Langages 168, Genèses de la terminologie contemporaine (sources et réception), p94.

Costa, R. et Silva, R. 2004. « The verb in the terminological collocations. Contribution to the development of a morphological analyser MorphoComp », Proceedings of Language Resources and Evaluation – LREC 2004, ELRA, Lisbon.

Cottez, H. 1994. « Les bases épistémologiques et linguistiques de la nomenclature chimique de 1787 », Meta : journal des traducteurs / Meta : Translators' Journal, 39 (4), 676-691.

Daille, B. 1994. Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques, Thèse, Université Paris 7.

Daille, B. 2002. Découvertes linguistiques en corpus, Mémoire d'Habilitation à Diriger des Recherches en Informatique, Université de Nantes.

Darmesteter, A. 1877. De la création actuelle de mots nouveaux dans la langue française et des lois qui la régissent, Université de Paris.

Darwish, K. 2002. « Building a Shallow Arabic Morphological Analyzer in One Day », Actes du workshop Computational approaches to Semitic languages, 47-54.

De Bessé, B. 1990. « La définition terminologique », Chaurand, Jacques/Mazière, Francine (eds.) : La définition, 252-261.

De Bessé, B. 2000. « Le domaine », Bejoint, Henri et Philippe Thoiron (dir.), (2000) : Le sens en terminologie, Lyon, Presses universitaires de Lyon, coll. 'Travaux du C.R.T.T.', 381.

De Morveau, G.et al. 1787, Méthode de nomenclature chimique, proposé par MM. De Morveau, Lavoisier, Berthollet et de Fourcroy.

De Saussure, F. 1968. Cours de linguistique générale, Première partie, Principes généraux, ch. Ier, Nature du signe linguistique, § 1, Signe, signifié, signifiant. Éd. Payot, Paris, 97-103.

Debili, F. et Souissi, E. 1998. « Etiquetage grammatical de l'arabe voyellé ou non », Proceeding Semitic '98 Proceedings of the Workshop on Computational Approaches to Semitic Languages, 16-25, Montreal, Quebec, Canada.

Descles, J-P. 1987. « Réseaux sémantiques : la nature logique et linguistique des relateurs », Langages, Sémantique et intelligence artificielle, sous la direction de François Rastier, 22e année, n°87, 55-78.

Descles, J-P. 1990. Langages Applicatifs langues naturelles et cognition, Hermès. Paris.

Depecker, L. 1994. « L'aménagement terminologique : de l'usage à la décision », Terminologies nouvelles, n°12, 9-13.

Depecker, L. 2002. Entre signe et concept. - Eléments de terminologie générale, Presses de la Sorbonne Nouvelle, 198 p.

Depecker, L. 2005. « Contribution de la terminologie à la linguistique », Langages, n°157, La terminologie : nature et enjeux, sous la direction de Loïc Depecker, 6-13.

Depecker, L. 2005. « Présentation », Langages, n° 157, 3-5.

Despres, S. et Szulman, S. 2008. « Réseau terminologique versus Ontologie », TOTh 2008, France.

Dichy, J. 1990. *L'écriture dans la représentation de la langue : la lettre et le mot en arabe,* Thèse, Université Lumière-Lyon 2.

Dichy, J. 1997. « Pour une lexicomatique de l'arabe : L'unité simple et l'inventaire fini des spécificateurs du domaine du mot », Erudi, 42 (2), 291-306.

Dichy, J. 2001. « La néologie stratifiée de l'arabe, fidélité contradictoire du lexique à sa mémoire, à partir du champ sémantique peuple-patrie-nation », *L'Arabisant* n°35, Paris, Association Française des Arabisants, 30-41.

Dichy, J. 2010. « La polyglossie de l'arabe, illustrée par deux corpus d'époques et de natures différentes : un échange radiophonique syrien et un conte des Mille et Une Nuits », Bozdemir, M. et Clavet, L.-J. (éds), Les politiques linguistiques en Méditerranée, Paris, Honoré Champion, 219-245.

Dichy, J. 2017. « Polyglossie de l'arabe et subsidiarité : au-delà des confusions entrainées par la notion de 'diglossie' », Arabe standard et variations régionales : Quelle(s) politique(s) linguistique ...Par Héba Medhat-Lecocq, Collectif.

Dichy, J., Braham, A., Ghazali S. et Hassoun, M. 2002. « La Base de connaissances linguistique DIINAR.1, (DIctionnaire INformatisé de l'ARabe - version 1) », A. Braham (éd.), Actes du colloque international sur Le Traitement *automatique de l'arabe,* Tunis, 45-56.

Diab, M., Hacioglu, K. and Jurafsky, D. 2004. « Automatic Tagging of Arabic text: from raw text to base phrase chunks », Proceedings of HLT/NAACL-200, 149-152.

Diab-Duranton, S. 2015. Substitution et créativité lexicales en arabe : Compilation, théorisation, restructuration, Geuthner, Paris.

Debili, F. 2001. « Traitement automatique de l'arabe voyellé ou non », Correspondances n°46, IRMC, Tunis.

Diki-Kidiri, M. 2007. « Éléments de terminologie culturelle », Actes du Colloque Terminologie : approches transdisciplinaires, Ottawa, 2-4.

Dubois, J. 1969. « Lexicologie et analyse d'énoncé », Cahiers de lexicologie 2, 115-126.

Dubois, J. 1994. Dictionnaire de linguistique et des sciences du langage, Larousse, Paris, 576p.

Dubuc, R. 2002. Manuel pratique de terminologie, 4e éd., Brossard, Linguatech, 194 p.

Edeline, F. 2009. « Les fonctions sémiotique et heuristique des symboles chimiques : ou de l'icône au symbole et retour », Protée, 37 (3), 45-56.

El Jihad, A. et Yousfi, A. 2005. « Étiquetage morpho-syntaxique des textes arabes par modèle de Markov caché », Actes de RECITAL 2005, 649-654.

Elkateb, S., Black, W., Rodriguez, H., Alkhalifa, M. Vossen, P., Pease, A. et Fellbaum, C. 2006. « Building a WordNet for Arabic », Proceedings of The fifth international conference on Language Ressourcesand Evaluation (LREC 2006).

Faber, P. 2014: « Frames as a Framework for Terminology », Handbook of Terminology, John Benjamins, Amsterdam, London, 14-33.

Frérot, C. 2000. Vitamines, Carotène et Polyphénols, DESS Industrie de la langue et traduction spécialisée.

Friedl, J-E-F. 2003. Maîtrise des expressions régulières. Paris, O'Reilly, 337 p.

Friedli, C-K-W. 2002. Chimie générale pour ingénieur, Presses polytechniques et universitaires romandes.

Gaiffe, B., Jacquey, E. et Kister, L., 2009. « Approche lexico-sémantique de l'extraction terminologique : utilisation de ressources lexicographiques et validation sur corpus », *Toth'09*, Annecy.

Galisson, R. et Coste, D. 1976. Dictionnaire de didactique des langues. Paris, Hachette.

Gambier, Y. 1991. « Travail et vocabulaire spécialisé : prolégomènes à une socioterminologie », Meta, 36 (1), 8-15.

Gandon, F. 2008. Graphes RDF et leur Manipulation pour la Gestion de Connaissances, Mémoire d'habilitation à diriger des recherches, Université de Nice – Sophia Antipolis.

Gaubert, C. 2010. « Kawâkib, une application web pour le traitement automatique de textes arabes », Anisl, 43.

Gaudin, F. 1993. Pour une socioterminologie : des problèmes sémantiques aux pratiques institutionnelles, éd. Duculot, coll. « Champs linguistiques », Bruxelles, 286 p.

Gentilhomme, Y. 2000. « Termes et textes mathématiques. Réflexions linguistiques non standard », Cahier de la lexicologie, 76, 57-89.

Ghazi, J. 1987. « Propositions pour une typologie nouvelle de la création lexicale », Arabica, T. 34, Fasc. 2, 147-163.

Ghénima, M. 1998. Un système de voyellation de textes arabes, Lyon.

Ghoul, D. 2011. *Outils génériques pour l'étiquetage* morphosyntaxique de la langue arabe : *segmentation et corpus d'entrainement*, Rapport du master, Université Grenoble 3.

Ghoul, D. 2013. « Développement de ressources pour l'entraînement et l'utilisation de l'étiqueteur morphosyntaxique TreeTagger sur l'arabe », TALN-RECITAL 2013 les Sables *d'Olonnes*, 69-82.

Gouadec, D. 1990. Terminologie. Constitution des données, Paris-La Défense : Afnor, 218 p.

Gross G. 1988. « Degré de figement des noms composés », Langages, n°90, Larousse.

Gross, G. 2008. « Les classes d'objets », Lalies, 28, 111-165.

Guarino, N. 1998. « Formal Ontology and Information Systems », Formal Ontology in Information Systems, N. Guarino (ed), Amsterdam, IOS Press, 3-15.

Guilbert, L. 1973. « La spécificité du terme scientifique et technique », Langue française : Les vocabulaires techniques et scientifiques, 17 (1), 5-17.

Guilbert, L. 1975. La créativité lexicale, Larousse, Paris, 285 p.

Gruber, T-R. 1993. « A Translation Approach to Portable Ontology Specification », Knowledge Acquisition, 5, 199-220.

Habash, N. 2010. « Introduction to Arabic Natural Language Processing », Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

Habert, B. 2000. « Des corpus représentatifs : de quoi, pour quoi, comment ? », Bilger, M. (éd.), Linguistique sur corpus - Etudes et réflexions, Presses Universitaires de Perpignan.

Habert, B., Nazarenko, A. et Salem, A. 1997. Les linguistiques de corpus, Paris :Armand Colin, 240 p.

Habert, B. et Zweigenbaum, P. 2002. « Régler les règles », TAL, 13 (2), 83-105

Habrant, J., Corbel, A., Girardot, J.J. et Savoy, J. 1999. « Utilisation des réseaux sémantiques pour la navigation dans l'hypertexte », Colloque Multimédia et Construction des Savoirs, Besançon.

Halleux, R. 1997. « La réception de l'alchimie arabe en occident », Histoire des sciences arabes : Technologie, alchimie et sciences de la vie, Volume 3, sous la direction de Roshdi Rashed, 141-154.

Hamzaoui, R. 1965. L'Académie arabe de Damas et les problèmes de la modernisation de la langue arabe, Leiden, E.J. Brill, 84 p.

Hamzé, H. 1998. « De la racine au mot ou du mot à la racine : problématique de la création d'une nouvelle mémoire de l'emprunt en arabe », Revue tunisienne de sciences sociales n o spécial : actes du colloque de linguistique, Tunis, 117, 35 ème année, p62.

Haralambous, Y. et Lavagnino, E. 2011. « La réduction de termes complexes dans les langues de spécialité », TAL : traitement automatique des langues, 52 (1), 37-68.

Hardane, J. 1994. *Rôle du français dans l'élaboration terminolo*gique arabe, 482 p.

Harrathi, F. 2009. Extraction de concepts et de relations entre concepts à partir des documents multilingues : Approche statistique et ontologique, Thèse de doctorat, Institut Nationale des Sciences Appliquées de Lyon.

Hassoun, M. 1987. Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application, Thèse pour le doctorat d'état (ès Lettres), Université de Lyon 1.

Hermans, A. 1989. « La définition des termes scientifiques », Meta, 34(3), 529-532.

Hernandez, N. et Mothe, J. 2006. « TtoO : une méthodologie de construction d'ontologie de domaine à partir d'un thésaurus et d'un corpus de référence », rapport interne IRIT/RR.

Hudrisier, H. et Ben Henda, M. 2017. « Normalisation de la langue et de l'écriture arabe : enjeux culturels régionaux et mondiaux ».

Humbley, J. 1996. « La légitimation en terminologie », Sémiotiques, Didier Erudition, n°11, Paris, 119-136.

Humbley, J. 2001. « Quelques enjeux de la dénomination en terminologie », Cahiers de praxématique, n°36, Montpellier, Pulm, 93-115.

Humphreys, K., Demetriou, G. and Gaizauskas, R. 2002 « Two applications of information extraction to biological science journal articles : enzyme inter-actions and protein structures », Proceedings of the Pacific Symposium on Biocomputing (PSB-2000), Honolulu, Hawaii, USA, 505-516.

Hunt, A. trad. Depovere, P. 2006. La chimie de A à Z, Dunod, Paris, 466 p.

Ibekwe-Sanjuan, F. 2007. Fouille de textes : méthodes, outils et applications, éditions Hermès-Lavoisier, 352p.

ISO - Organisation internationale de normalisation : ISO 1087-1 : 2000. Travaux terminologiques -- Vocabulaire -- Partie 1 : Théorie et application, Genève.

ISO - Organisation internationale de normalisation : ISO 860 : 2007. Travaux terminologiques -- Harmonisation des concepts et des termes, Genève.

ISO - Organisation internationale de normalisation : ISO 704 : 2009. Principes et méthodes de la terminologie, Genève.

Jacques, M-P. 2003. Approche en discours de la réduction des termes complexes dans les textes spécialisés, Thèse, Université de Toulouse.

Jaber, F. 2012. Les manuels scolaires arabes de chimie : analyse de la terminologie et du discours, Thèse, Université de Paris, INALCO.

Jazzar, S. 2009. L'ontologie de l'économie pétrolière en Arabie Saoudite et analyse terminologique anglais-français-arabe, Thèse, Université de Franche-Comté, Besançon.

Kaczmarek, K. 2011. Nomenclatures française et polonaise de la chimie organique. Analyse comparative, Thèse, Université Mickiewicz de Poznań.

Kettani-Idrissi, A. 1987. La néologie arabe : problèmes et perspectives, Thèse, Université de Paris 3.

Khaioutine, A. et Tiulnina, V. 1996. Tenants et aboutissants des études sur les notions de terminologie et nomenclature dans la linguistique russe et soviétique.

Kister, L., Jacquey, E. et Gaiffe, B. 2009. « Fusion d'un thesaurus et d'une terminologie : utilisation de ressources existantes pour amorcer une onto-terminologie »,

Khoja, S. 2001. « APT: Arabic part-of-speech tagger », Proceedings of The Student Workshop at the second meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001), 20-26.

Kiraz, G-A. 1996. « Analysis of the Arabic Broken Plural and Diminutive », Proceedings of the 5th international conference and Exhibition on Multi-Lingual Computing.

Kleiber G. 1997. « Massif / comptable et partie / tout », Verbum, Vol XIX n°3, La relation partie-tout, Presses universitaires de Nancy

Kocourek, R. 1982. La Langue française de la technique et de la science, Oscar Brandstetter, Wiesbaden.

Kocourek, R. 1991. [1982]. La langue française de la technique et de la science. Vers une linguistique de la langue savante, 2e éd., Wiesbaden/Paris, Brandstetter Verlag, 327 p.

Kouloughli, D-E. 1994. Grammaire de l'arabe d'aujourd'hui, Pocket-Langues pour tous, 349p.

Ladenburg, A. 1909. Histoire du développement de la chimie depuis Lavoisier jusqu'à nos jours, A. Hermann et fils, Paris, 388 p.

Lachheb, K. 2008. Lexique du Commerce Electronique. Anglais – Français – Arabe, Cefal.

Lachkar, A. 2014, Langues, cultures et médias en Méditerranée, l'Harmattan : Paris, 238 p.

Larcher, P. 2003. Le système verbal de l'arabe classique, Aix-en-Provence : Publications de l'Université de Provence, Collection Didactilangue, 194 p.

Larcher, P. 2008. « Dérivation arabisante et ishtiqâq arabe : histoire d'un malentendu », Barbara Kaltz (éd), Regards croisés sur les mots non simples, collection Langages, Lyon : ENS-Editions, 85-94.

Larcher, P. 2012. « Un cas de tératologie dérivationnelle en arabe ? Le verbe istakâna », Romano-Arabica New Series n°12, 55 Years of Arab Studies in Romania, 159-168.

Labbé, D. 1990. « Normes de saisie et de dépouillement des textes politiques », Cahier du CERAT, 1-135.

Lavoisier, A-L, De Morveau, Berthollet et De Fourcroy. 1789. Méthode nomenclature chimique, Paris, Cuchet.

Le Guern, M. 1973. Sémantique de la métaphore et de la métonymie, Paris : Larousse, 126p.

Le Guern, M. 1989. « Sur les relations entre terminologie et lexique », Meta : journal des traducteurs / Meta : Translators' Journal, 34, (3), 340-343.

Lehmann, A. et Martin-Berthet, F. 2005. Introduction à la lexicologie, Paris : Armand Colin, 214 p.

Leigh, G. J., Favre, H. A. et Metanomski, W. V. 2001, Principes de nomenclature de la chimie. Introduction aux recommandati*ons de l'IUPAC*. De Boeck Université, Paris.

Lelubre, X. 1992. La terminologie arabe contemporaine de l'optique : faits - théories – évaluation, Thèse, Université Lumière-Lyon 2.

Lelubre X. 2005. « Le statut de l'adjectif en langue de spécialité », De la mesure dans les termes, Lyon, Presses Universitaires de Lyon, 249-268.

Lerat, P. 1995. Les langues spécialisées, coll. "Linguistique nouvelle", Paris, PUF.

Lerat, P. 2002. « Vocabulaire juridique en schémas d'arguments juridique », Meta 47, 2, 155-162.

Leroy, S. 2004. « Extraire sur patrons : allers et retours entre analyse linguistique et repérage automatique ». Revue Française de linguistique appliquée, 9 (1), 25-43.

L'Homme, M.-C. 1995. « Définition d'une méthode de recensement et de codage des verbes en langue technique : applications en traduction », TTR, 2 (8), 67-88.

L'Homme, M.C.. 1998. « Le statut du verbe en langue de spécialité et sa description lexicographique », Cahiers de lexicologie, 73 (2), 61-84.

L'Homme, M-C. 2004. La terminologie : principes et techniques, Les Presses de l'Université de Montréal : Montréal, 280 p.

L'Homme, M-C. 2012. « Le verbe terminologique : un portrait de travaux récents », Neveu, F. et al. (éd.), Actes du 3e Congrès mondial de linguistique française, Lyon, France, EDP Sciences.

Loffler-Laurian, A-M. 1984. « Vulgarisation scientifique : formulation, reformulation, traduction », Langue française, n°64, Français technique et scientifique : reformulation, enseignement, sous la direction de Jean Peytard, Daniel Jacobi et André Pétroff, 109-125.

Lorente, M. 2007. « Les unitats lèxiques verbals dels textos especialitzats. Redefinició d'una proposta de classificació », Lorente, M. et al. (ed.) Estudis de lingüístics i de lingüística aplicada en honor de M. Teresa Cabré Catellví, Volum II, De deixebles, Barcelona : Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra, Sèrie Monografies 11-12, 365-380.

Lorente, M. et Bevilacqua, C. 2000. « Los verbos en las aplicaciones terminograficas », Actas del VII Simposio Iberoamericano de TErminologia RITerm 2000, Lisboa : ILTEC.

Maamouri, M., Bies, A., Buckwalter, T. and Mekki, W. 2004. « The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus », Paper template for Coling, Geneva - Semantic Scholar

Maldonado A., 2006. Diversité Moléculaire : Application au Criblage Virtuel, Corrélation avec des Propriétés physico-Chimiques, Thèse, Université Paris 7 – Diderot.

Mallak, I. 2011 *De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en* Rec*herche d'Information*, Thèse, Université de Toulouse.

Meftouh K., Smaïli K., Laskri M. T. 2007. « Constitution d'un corpus de la langue Arabe à partir du Web », Colloque International sur le Traitement Automatique de la Langue Arabe - CITALA'07, Rabat, Maroc.

Mesfar, S. 2008. Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard, Thèse, Université Franche Compté.

Messaoudi, A. 2007. « La transitivité en Arabe classique entre interprétation et syntaxe », Cahier du CRISCO n 23.

Meyer, C-F. 2002. English corpus linguistics : An introduction. Cambridge : Cambridge University Press.

Mizoguchi, R. 2004. « Le rôle de l'ingénierie ontologique dans le domaine des EIAH », *Sciences et Technologies de l'Information et de la Communication pour l'Education et la* Formation, Vol. 11.

Monteil, V. 1960. *L'arabe moderne*, Klincksieck, Paris.

Morgenroth, K. 1994. Le terme technique, Max Niemeyer, Tûbingen.

Mc Enery T. et Wilson A. 2001 (1996). Corpus linguistics, an introduction, Edinburgh University Press.

Mouelhi, Z. 2008. *Essai de lexicométrie d'une œuvre arabe classique : Al-'Imtâ' wa*-l-*Mu'ânasa de Tawhîdî.* Thèse, Université Lumière-Lyon 2.

Mouelhi, Z. 2008 AraSeg* : un segmenteur semi-automatique des textes arabes ».

Mourad, G. 2001. Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des Applications informatiques : SegATex et CitaRE, Thèse, Université de Paris-Sorbonne.

Mourad, M., Antoniadis, G. et Zrigui, M. 2008. « Nouvelles ressources et nouvelles pratiques pédagogiques avec les outils TAL », TICEMED 08, Journal Information Sciences for Decision Making (Journal ISDM), ISDM32, n°571.

Muller, C. 1969. « La statistique lexical », Langue française, n°2, Le lexique, sous la direction de Louis Guilbert, 30-43.

Namer, F. 2003. « Le modèle Lstat : ou comment se constituer une base de données morphologique à partir du Web », Revue québécoise de linguistique, 32 (1), 85-109.

Nédellec, C., Ould Abdel Vetah, M. et Bessières, P. 2001 'Sentence filtering for information extraction in genomics, a classification problem', Actes de Conference on Practical Knowledge Discovery in Databases, PKDD'2001, 326-338, Freiburg.

Neifar, W et Ben Ltaief, A. 2016. « Acquisition terminologique en arabe : État de l'art », Actes de la conférence conjointe JEP-TALN-RECITAL, vol 3 : RECITAL.

Noy, N. F. et McGuinness, D. L. 2000. « Développement d'une ontologie 101 : Guide pour la création de votre première ontologie », Stanford, CA 94305.

Ouersighni, R. 2002 La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe : utilisation pour la détection et le diagnostic des fautes d'accord, Thèse, Université Lumière-Lyon 2.

Paquin, A. 2006. Étude de la néologie dans la terminologie du terrorisme avant et après septembre 2001 : une approche lexicométrique. Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de Maître en linguistique, Université de Montréal.

Pavel, S. et Nolet, D. 2001. Précis de terminologie, Canada, ISBN 0-660-61616-5, No de cat. S53-28.

Petit, G. 2003. « Lemmatisation et figement lexical : les locutions de type SV », Cahiers de Lexiologie, Centre national de la recherche Scientifique, 82 (1), 30-57.

Peraldi, S. 2011. Indétermination terminologique et multidimensionnalité dans le domaine de la chimie organique : analyse à partir d'un corpus spécialisé de langue anglaise, Thèse, Université Paris 7.

Pillet, V. 2000. *Méthodologie d'extraction automatique d'information* à partir de la littérature *scientifique en vue d'alimenter un nouveau système d'information*, Thèse, Université de droit, d'économie et des sciences d'Aix-Marseille.

Pimentel, J. 2007. O comportamento do verbo constituir em contexto de especialidade, Mémoire de maitrise. Faculdade de Ciências Sociais e Humanas. Universidade Nova de Lisboa.

Pimentel, J. et L'Homme, M-C. 2011. « Annotation syntaxico-sémantique de contextes spécialisés : application à la terminographie bilingue », van Campenhoudt, M., T. Lino et R. Costa (éd.). *Passeurs de mots, passeurs d'espoir : lexicologie, terminologie et traduction face* au défi de la diversité, Paris : Édition des archives contemporaines/Agence universitaire de la francophonie, 651-670.

Pimentel, J. 2012. « Description de verbes juridiques au moyen de la sémantique des cadres », Terminologie & Ontologie : Théories et applications (Toth 2011), Annecy 2011.

Prince, V. et Kodratoff, Y. 2007. « Le défi Fouille de Textes : Quels paradigmes pour la reconnaissance automatique d'auteurs ? », *Revue des Nouvelles Technologies de l'Information*, Hermann, E (10), 1-14.

Raheel, S. 2010. *L'Apprentissage Artificiel pour la Fouille de Données Multilingues :* Application à la Classification Automatique des Documents Arabes, Thèse, Université Lumière-Lyon 2.

Rastier F. 2005. « Enjeux épistémologiques de la linguistique de corpus », G. Williams (éd.), La linguistique de corpus, Rennes : PUR, 31-45.

Recommandations relatives à la terminologie, 1990. éd. Conférence des services de traduction des États de l'Europe occidentale, Groupe de travail terminologie et documentation, Chancellerie de la Confédération suisse.

Reguigui, A. 2002. Anatomie des syntagmes terminologiques arabes : analyse formelle et quantitative, Université Laurentienne, Sudbury.

Rey, A. 1976. « Néologisme, un pseudo concept? », Cahiers de lexicologie, n°28, 3-17.

Rey, A. 1979. La terminologie : noms et notions, Que sais-je ? Paris, PUF.

Richter, J.-Cl., Panico R., Powell, W. H., 1993, A Guide to IUPAC Nomenclature of Organic Compounds, Recommendations; Eds; Blackwell.

Rigaudy, S.P. et Kleaney J. 1979. « IUPAC, Organic Chemistry Division, Commission on Nomenclature of Organic Chemicals », Nomenclature of Organic Chemicals, J, Eds, Pergamon Press.

Roche, M. 2004. Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes. Thèse, Université Paris 11.

Roche, C. 2007. « Le terme et le concept : fondements d'une ontoterminologie », TOTh 2007 : Terminologie & Ontologie : Théories et Applications, Annecy.

Roche, C. 2008. « Quelle terminologie pour les sociétés de l'information ? », Lexique, dictionnaire et connaissance dans une société multilingue, Cahiers de linguistique, 43/1, 139.

Roman, A. 1987. « Les formes infinitives de l'arabe dans l'organisation générale de la langue arabe », L'infinitif, Presses universitaires de Lyon, 1987, 211-242.

Roman, A. 1990. *Grammaire de l'arabe,* Que sais-je ? Paris, P.U.F.

Roman, A. 1999. La création lexicale en arabe, ressources et limites du système de nomination dans une langue humaine naturelle, Presses Universitaires de Lyon.

Rondeau, G. 1991. Introduction à la terminologie, Québec, Gaétan Morin.

Rossi, M. 2015. « In rure alieno. Métaphores et termes nomades dans les langues de spécialité », Aspects linguistiques et culturels des discours spécialisés, Berne, Berlin, Bruxelles, Francfort-sur-le-Main, New York, Oxford, Vienne : Peter Lang, 114-120.

Sablayrolles. J-F. 1996-1997. « Néologismes : une typologie des typologies », Cahier du C.I.E.L, 10-48.

Sablayrolles. J-F. 2000. La néologie en français contemporain : examen du concept et analyse de productions néologiques récentes, Champion, Paris, 588 p.

Sabra, A. 2012. Les relations de causalité en arabe et en français avec une constitution de ress*ources linguistiques utilisables par l'informatique,* Thèse, Université Paris-Sorbonne.

Sadik, A. 2010. Néologie scientifique amazighe corpus lexical de chimie, Mémoire de Master, Université Ibn Zohr Agadir.

Séguéla, P. 2001. Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques, Thèse, Université de Toulouse III.

Serrano, L., Grilheres, B., Bouzid, M. et Charnois, T. 2011. « Extraction de connaissances pour le renseignement en sources ouvertes », *SOS'2011 at EGC'2011*.

Silva, R., Costa, R. et Ferreira, F. 2004. « Entre langue générale et langue de spécialité une question de collocations », Ela. Études de linguistique appliquée, n° 135, (3), 347-359.

Sinclair, J. 1991. Corpus, concordance, collocations, Oxford, Oxford University Press.

Sinclair, J. 1996. « Preliminary recommandations on Corpus Typology », EAG--TCWG--CTYP/P.

Sinclair, J. McH. (éd.) 2004. « How to use corpora in language teaching », Philadelphia : J. Benjamins, Studies in corpus linguistics, 308 p.

Slodzian, M. 2000. « L'émergence d'une terminologie textuelle », Le sens en terminologie, Béjoint &Thoiron (ed), Le sens en terminologie, Presses Universitaires de Lyon, 61-85.

Smadja, F. 1993. « Retrieving collocations from text : Xtract », Computational Linguistics, 19 (1), 143-177.

Studer, R., Benjamins V.R. et Fensel D. 1998. « Knowledge Engineering: Principles and Methods », IEEE Transactions on Data and Knowledge Engineering, 25 (162), 161-197.

Talafheh, A. 2003. La terminologie grammaticale complexe dans le Kitâb de Sîbawayhi, Thèse, Université Lumière-Lyon 2.

Tanguy, L. 2012. Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes, Habilitation à diriger des recherches, Université de Toulouse leMirail.

Tellier, I. 2007. Introduction au TALN et à l'ingénierie linguistique, cours d'informatique, Université de Lille 3.

Tellier, C. 2008. Verbes spécialisés en corpus médical : une méthode de description pour la *rédaction d'art*icles terminographiques, Travail dirigé présenté au Département de linguistique et de traduction, Université de Montréal.

Toussaint, Y. 2011. Fouille de textes : des méthodes symboliques pour la construction *d'ontologies et l'annotation sémantique guidée p*ar les connaissances, Habilitation à diriger les recherches, Université Henri Poincaré - Nancy 1.

Temmerman, R. 2000. « Une théorie réaliste de la terminologie : le sociocognitivisme », Terminologies nouvelles 21, 58-64.

Troupeau, G. 1994. « La formation du vocabulaire scientifique et intellectuel », Prepols, N°VII, 15-16.

Tuerlinckx, L. 2004. « La lemmatisation de l'arabe non classique », JADT 2004 : 7es Journées *internationales d'Analyse statistique des Données Textuelles.*

Vicente, C. G. 2009. « La didactique du concept de langue spécialisée : vers une approche traductologique de la question », Mutatis Mutandis, 2 (1), 38-49.

Wüster, E. 1981. « L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses », Textes choisis

de terminologie 1, Fondements théoriques de la terminologie, Presses de l'université de Laval, 55-114.

Zaafrani, R. 2002. Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère, Thèse, Université Lumière-Lyon 2.

Zacklad, M. 2007. « Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI) », 35e Congrès annuel de l'Association Canadienne des Sciences de l'Information, *Partage de l'information dans un* monde fragmenté : Franchir les frontières, sous la dir. de C. Arsenault et K. Dalkir. Montréal, Canada.

Zghibi, R. 2002/2003. « Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646 », Document numérique, Vol. 6, 155-182.

Zipf, G. K. 1949. Human Behavior and the Principle of the Least Effort : An Introduction to Human Ecology.