Nouvelle Méthode d'Extraction Automatique Bilingue des Syntagmes Terminologiques Nominaux à Base de leurs Noyaux et du Balisage Structurel XML du Corpus Aligné

Maha M. EL BACHA

Département de français, Faculté des Langues (Al-Alsun), Université d'Ain-Chams, Égypte
Maha.elbacha@alsun.asu.edu.eg

Résumé : ***La présente recherche s'insère dans le cadre de la terminologie computationnelle. Elle présente une nouvelle méthode d'extraction automatique bilingue des Syntagmes Terminologiques Nominaux (STN) se basant sur leurs noyaux qui sont obligatoirement des Termes simples nominaux (TSN) et exploitant le balisage structurel XML de leur corpus bilingue aligné entre le français et l'arabe. La recherche aborde en détails le principe fondateur ainsi que les phases de notre méthode dite statistique et structurelle, à commencer par l'extraction et la segmentation, passant par maintes phases de dépouillement automatique et d'alignement automatique de terminologie jusqu'à parvenir à la validation finale. Au terme de la recherche, nous effectuons un examen de la résolution de l'extraction automatique en vertu des résultats terminologiques assortis de statistiques. La recherche étale également l'originalité de la méthode d'extraction bilingue automatique figurant dans nombre de facteurs dont le plus important s'avère être la productivité et la génération terminologique des noyaux nominaux et de nouveaux STN.***

Mots-clés : ***syntagme terminologique nominal - extraction automatique de termes - terme simple - noyau terminologique - balisage structurel XML - corpus aligné - terminologie computationnelle.***

1 INTRODUCTION

La syntagmatique occupe une place prépondérante dans les études terminologiques étant donné que les syntagmes terminologiques sont les unités les plus récurrentes des corpus spécialisés. Les études axées sur la formation syntagmatique ont prouvé la prédominance des syntagmes dans les nomenclatures terminologiques dans presque tous les domaines de spécialisations. D'autre part, l'analyse structurelle des syntagmes terminologiques débouche sur un schéma prototypique où la structure de base est la syntagmatique nominale. D'où l'importance octroyée à l'étude des Syntagmes Terminologiques Nominaux (STN) qui intègrent deux caractéristiques types des unités terminologiques, à savoir : la syntagmatique et la nomination.

En effet, l'étude des STN renferme un défi majeur qui figure dans leur repérage. Il est indéniable que tout repérage des unités terminologiques s'applique automatiquement aux corpus spécialisés. On parle depuis des années des systèmes d'extraction automatique de termes qui varient selon leurs méthodes et se distinguent par leur efficacité dans les langues concernées. Notre problématique dans le présent article sera centrée sur la présentation d'une nouvelle méthode d'extraction automatique bilingue des Syntagmes Terminologique Nominaux, laquelle sera décrite en détails dans les points suivants.

2 LES SYNTAGMES TERMINOLOGIQUES NOMINAUX (STN)

Le syntagme terminologique nominal (STN) est une unité terminologique motrice et puissante dans les corpus spécialisés. A la différence du terme simple (TS), le syntagme terminologique nominal constitue une unité complexe aux niveaux syntaxique et conceptuel. Nombreuses sont les définitions qui expliquent ce que nous désignons par STN représentant ses diverses caractéristiques et spécificités qui le distinguent des autres unités linguistiques complexes.

Toute définition du STN doit porter sur trois aspects complémentaires, à savoir : syntagmatique, terminologique et nominal. L'aspect syntagmatique est saillant dans la structure complexe du STN qui se compose de deux mots pleins au moins. L'aspect terminologique reflète son appartenance à un domaine spécialisé. Et finalement l'aspect nominal puisque le STN est composé autour d'un noyau qui est essentiellement un nom ou un substantif. Nous pouvons définir donc le STN comme étant :

> Une unité terminologique complexe composé d'au moins deux mots pleins séparés par une espace typographique blanche dont le noyau est un substantif qui est nécessairement un terme simple et qui possède comme fonction dénominative de rendre la complexité notionnelle d'un concept spécialisé. [1 : p.110]

En effet, les syntagmes terminologiques nominaux occupent une position prioritaire par rapport aux autres types d'unités terminologiques, vu sa fonction substantive qui est de nature à dénommer les concepts spécialisés par opposition aux verbes qui désignent temporellement des actions, ainsi que les autres parties de discours qui décrivent soient des noms ou des actions, telles les adjectifs et les adverbes. En outre, la complexité structurelle du STN lui octroie un fort pouvoir dénominatif conceptuel qui permet de refléter la complexité notionnelle des domaines spécialisés. Et plus le STN se complexifie, plus il se dote d'une puissance conceptuelle au niveau terminologique.

C'est exactement cette puissance conceptuelle et dénominative qui a poussé les chercheurs à octroyer un poids considérable à l'étude des syntagmes terminologiques nominaux. Ce poids n'est pas l'apanage d'un domaine plus que l'autre. En effet, l'intérêt accordé au STN couvre presque tous les domaines spécialisés et les études portant sur l'abord des problématiques du STN font taches d'huile au niveau international. Parmi lesdites problématiques figure l'extraction automatique des STN.

3 EXTRACTION AUTOMATIQUE DE TERMES

L'extraction automatique des STN soulève deux problèmes intrinsèques. Il s'agit premièrement de la modalité automatique du repérage des unités terminologiques, et deuxièmement de la nature complexe des STN qui risque de compliquer les opérations d'extraction automatique. Avant d'aborder notre méthode d'extraction automatique des STN, il importe de passer en revue les méthodes communément admises dans l'extraction automatique de termes[1] et leurs différentes approches.

L'extraction terminologique, dont le repérage et la segmentation sont les opérations majeures, a subi une automatisation complète depuis des décennies et il est devenu absurde de traiter un corpus spécialisé sans recourir à l'extraction automatique de sa terminologie ou à un « *extracteur de termes* », comme le dénomme Marie-Claude L'Homme [2 : p.166]. Une extraction automatique des termes peut être expliquée en ces termes :

> Une reconnaissance automatique dans un corpus spécialisé des unités, simples ou complexes, susceptibles d'être des termes pour les faire ressortir en tant que candidats termes (CT) dans des listes qui doivent être écrémées et validées par un terminologue en vue de décider du statut terminologique desdites unités CT. Une fois la validation terminée, la liste des CT se transforme en liste de termes proprement dite. [1 : p.204]

La définition précédente nous offre une vue complète du processus automatique de l'extraction terminologique, lequel ajuste la performance terminologique, réduit à néant le temps et l'effort et maximise la qualité du produit terminologique escompté.

En effet, l'extraction automatique des termes n'est pas un système à voie unique : il existe plusieurs méthodes et plusieurs approches dans l'abord de l'extraction automatique de termes. En vue de pouvoir classifier lesdites méthodes et approches, nous devons prendre en considération trois critères conditionnant cette diversité : le type d'unité terminologique visée par l'extraction, la méthodologie adoptée par l'extraction et le nombre de langues impliquées dans l'extraction. Nous pouvons donc passer en revue les méthodes d'extraction automatique de termes selon les critères précités comme suit :

A. Méthodes d'extraction automatique de termes selon le type d'unité terminologique visée

Les méthodes d'extraction automatique de termes diffèrent selon que l'unité terminologique à extraire est une unité simple, autrement dit un terme simple (TS), ou une unité complexe, qui peut englober les composantes de la phraséologie terminologique ayant à leur tête les collocations et les syntagmes terminologiques (ST). La différence entre les deux types d'unités réside dans la nature de leur structure qui implique par voie de conséquence de changer la stratégie d'extraction. On a recours à des stratégies d'extraction automatique simple pour repérer les termes simples qui se composent d'un seul mot. Par contre les stratégies visant les unités complexes, telles les syntagmes terminologiques, se compliquent certainement pour parvenir à contourner leur complexité structurelle qui varie à son tour selon la variété des unités complexes à extraire. Dans la présente recherche, nous nous contentons des méthodes visant l'extraction des syntagmes terminologiques en général (ST).

[1] L'extraction automatique de termes est connue également selon l'appellation de Patrick Drouin [4] par « acquisition automatique de termes » de l'anglais « Automatic Term Recognition ». Nous optons dans la présente recherche pour l'appellation de « extraction automatique de termes ».

B. Méthodes d'extraction automatique de termes selon la méthodologie adoptée

La méthodologie adoptée désigne la stratégie avec laquelle sont conçus et élaborés les algorithmes d'extraction automatique des ST. Trois méthodes connues sont de mise en la matière[2], à savoir : la méthode linguistique, la méthode statistique et la méthode mixte ou hybride.

1) Méthode linguistique : elle compte principalement sur les connaissances linguistiques d'ordre lexical, morphologique, syntaxique et/ou morphosyntaxique. Cette méthode débouche sur nombre de stratégies d'extraction, à savoir : primo, l'identification des patrons typiques des ST ou des STN dont la formation syntaxique remplit des matrices prédéterminées et qui repose sur des étiqueteurs assignant à chaque mot d'un corpus spécialisé des étiquettes morphosyntaxiques pour que le système puisse les détecter automatiquement ; secundo, le repérage des frontières fondé sur « *des connaissances "en négatif" concernant les configurations grammaticales dont on sait qu'elles ne peuvent plus être des constituants de termes »* [3 : p.25]. Plusieurs systèmes et/ou outils sont développés sur la base de la méthode linguistique, tels NOMINO par David et Plante (1990), LEXTER par Bourigault (1994) et FASTER par Jacquemin (1997). Ces divers outils sont connus sous le nom de « *systèmes d'extraction automatique des candidats à terme (SEACAT)* ». Deux reproches sont à adresser à cette méthode : il s'agit premièrement de sa dépendance d'une langue déterminée et deuxièmement du taux de bruit élevé, vu la similitude confuse entre des unités qui partagent la complexité et la nomination sans être terminologiques.

2) Méthode statistique : elle se contente exclusivement des mesures quantitatives et des calculs des valeurs numériques et se passe, au contraire de la méthode précédente, des connaissances linguistiques. Nombreuses mesures statistiques s'imposent comme indices fiables sur l'identification des unités terminologiques, comme la « *fréquence* », que ce soit la fréquence absolue ou brute ou la fréquence relative, la « *répartition* » et la « *dispersion* ». Le calcul statistique appelé « *Information Mutuelle (IM)* » de l'anglais « *Mutual Information (MI)* », mis en place par les pionniers K.W. Church et P. Hanks (1989) pour identifier les collocations en mesurant le degré d'association entre ses composantes, est dument exploité dans l'extraction des ST. Il est question également du « coefficient d'étrangeté » de l'anglais « coefficient of weirdness », développé par Kurshid Ahmad (1996), se basant sur le calcul de la fréquence relative d'une forme dans deux corpus différents, l'un spécialisé et l'autre général. Par opposition à la méthode linguistique, la méthode statistique consacre l'indépendance de ces stratégies et calculs par rapport aux langues, ce qui permet leur extension à d'autres langues d'application. Toutefois, cette méthode quantitative manifeste des lacunes relatives à l'augmentation du taux de silence, d'autant plus que le seuil minimal et/ou maximal déterminé par le système élimine toute unité de nature terminologique mais se situant au-delà dudit seuil. En outre, l'extrapolation des résultats terminologiques issus d'un modèle statistique risque d'être douteuse si la taille du corpus ne dépasse pas certaines limites fixées par Drouin à titre indicatif à « *100 000 occurrences* » [4 : p.93].

3) Méthode hybride ou mixte : elle est générée comme résultat naturel de la fusion des deux méthodes précédentes pour en combler les lacunes et en exploiter les avantages. Les travaux de Daille (1993), débouchant sur la conception du système ACABIT, et de Drouin (2002) sur les « *pivots lexicaux spécialisés (PLS)* », débouchant sur la conception de TermoStat, s'inscrivent dans le cadre de la méthode hybride. Il faut noter que les systèmes basés sur la méthode hybride diffèrent dans l'ordre adopté entre les deux méthodes linguistique et statistique. Toujours, la méthode hybride n'est pas exempte d'inconvénients et reste malgré tout tributaire de la langue de travail.

C. Méthodes d'extraction automatique de termes selon le nombre de langues impliquées

Les méthodes abordées, jusqu'ici, concernent l'extraction automatique unilingue, qui s'applique sur un corpus monolingue. L'extraction automatique bilingue fonctionne selon des modèles plus compliqués étalés par Gaussier [5] et L'Homme [2 : p.209] dans 3 modèles distincts :

1) Modèle de la double extraction des candidats-termes puis alignement : ce modèle repose sur deux phases consécutives : la première phase consiste en une double extraction monolingue, primo au niveau du corpus source, et secundo au niveau du corpus cible ; la seconde phase concerne l'alignement entre les deux listes de candidats-termes (CT). Ce modèle vise principalement les STN binaires.

2) Modèle de l'alignement des mots puis identification simultanée des termes sources et cibles : ce modèle s'oppose au premier, puisqu'il procède premièrement à l'alignement entre corpus source et corpus cible au niveau des mots, et applique subséquemment une analyse syntaxique sur les phrases des deux corpus pour l'identification simultanée des CT sources et cibles.

[2] Patrick Drouin, quant à lui, y ajoute une méthode qu'il qualifie de « *mécanique* » [4 : p.54]. Selon Drouin toujours, ce modèle mécanique vise moins les termes que les collocations. C'est pourquoi, nous n'avons pas opté dans la présente recherche pour la classification de Drouin.

3) Modèle de l'extraction de la langue source puis identification des séquences de traduction en langue cible : ce modèle consiste en une extraction automatique monolingue des CT au niveau de la langue source uniquement. Après quoi, on procède à l'identification des équivalents de traduction lors de l'alignement au niveau des mots.

Il faut prendre en considération que, lors de l'adoption de l'un des trois modèles de l'extraction bilingue susmentionnés, les méthodes de l'extraction monolingue sont dument exploitables. Le système hybride de Daille et al. (1994) est exploité dans le premier modèle. Le deuxième modèle a eu recours à des systèmes tels que Termight, qui est un système linguistique conçu par Dagan et Church (1997), et TRINITY, système statistique conçu par Hull (2001). Quant au troisième modèle, le système SYNTEX proposé par Ozdowska et Bourigault (2004) est utilisé dans deux versions l'une consacrée à l'anglais et l'autre au français [6 : pp.47-53]. Une dernière réflexion à ce propos est digne de mention concernant l'importance du recours à des corpus parallèles et/ou alignés dans l'augmentation de la résolution de l'alignement.

A la suite de ce bref panorama, il importe de situer notre méthode d'extraction automatique de ST à la lumière de cette panoplie de stratégies et de méthodes, tout en prenant en considération la spécificité de notre paire linguistique représenté dans un corpus aligné entre le français et l'arabe.

4 METHODE D'EXTRACTION AUTOMATIQUE DES STN A PARTIR DE LEURS NOYAUX SIMPLES

A. Le noyau terme simple nominal (TSN) comme indice de repérage du syntagme terminologique nominal (STN)

Suivant le schéma syntagmatique de Bourigault, tout Syntagme Terminologique (ST) est composé « *d'une tête, qui est le terme (représentant la classe) générique, et d'une expansion, qui est un complément (un adjectif ou un groupe nominal) mentionnant un caractère spécifique* » [7 : p.3]. Les deux composantes basiques d'un ST sont donc « *la tête* », ou dans d'autres appellations[3] « *le noyau* » auquel nous optons dans la présente recherche, et « *l'expansion* »[4].

Selon El Bacha [1 : p.113], le « *noyau* » constitue « *l'unité centrale d'un ST* » qui doit indubitablement être un terme simple (TS), dont la catégorie détermine celle du ST en entier. Par voie de conséquence, le STN se compose d'un noyau qui est obligatoirement un terme simple nominal (TSN). D'où l'appellation de « noyau » auquel nous avons eu recours et qui représente à fond deux caractéristiques :

1) la position qu'il occupe au sein d'un syntagme terminologique, étant donné qu'il est ambulant par rapport à ses expansions et peut soit occuper la première place comme dans [commission consultative], soit occuper la seconde place comme pour [grande commission], soit occuper la place centrale par rapport à deux expansions comme dans [grande commission mixte – grandes commissions de l'Assemblée Générale].

2) le rôle qu'il exerce dans la reproduction terminologique syntagmatique que nous désignons par productivité terminologique et qui « *consiste dans l'aptitude d'un noyau nominal qui est évidemment un TSN à générer des STN, d'où l'appellation de 'génération syntagmatique'* » [1 : p.297].

C'est exactement cette génération syntagmatique qui est à l'origine de notre méthode d'extraction automatique des STN. Cette méthode se base principalement sur le noyau nominal du ST, qui est un terme simple nominal (TSN).

D'autre part, le mot, qui est une unité lexicale simple séparée par deux espaces blanches des unités avoisinantes, représente « *un élément de référence important dans les applications du traitement automatique de la langue* » [10 : p.30]. En terminologie, le terme simple (TS), étant l'équivalent du mot dans les textes spécialisés, représente quant à lui la matière rudimentaire de tout traitement automatique d'un corpus spécialisé. Et eu égard à la prédominance des formes nominales au niveau terminologique simple, le TSN se trouve donc au cœur de toute opération d'extraction automatique syntagmatique, d'autant plus qu'il est le noyau autour duquel s'articulent les diverses expansions d'un STN.

Ce rôle concentrique du TSN dans la génération syntagmatique des unités terminologiques phraséologiques plus complexes, à commencer par les collocations terminologiques, passant par les syntagmes terminologiques, allant aux compositions phraséologiques plus développées, reste malheureusement déconsidéré comme indice de repérage efficace dans les méthodes d'extraction automatique de termes au profit d'autres indices que nous jugeons superficiels, soit sur le plan structurel dans la formation des STN, soit sur le plan fonctionnel du statut des STN dans la phraséologie

[3] Le terme « tête » employé par plusieurs terminologues et chercheurs tels Marie-Claude L'Homme (2004), Agatha Savary (2000), Marie-Paule Jacques (2000) et Andy Lauriston (1993), est désigné tantôt par « base » comme c'est le cas dans les écrits de Claudine Bertrand (1998) et Isabelle Lapierre [16], tantôt par « noyau » par Martinet (1979) et Richard Bouché (1989).

[4] Le terme « expansion » est également désigné par « cooccurrence » chez Bertrand et par « modificateur » par L'Homme.

terminologique. C'est pourquoi, nous favorisons la démarche valorisant le TSN dans l'extraction automatique des STN, comme l'affirment Nakagawa et Mori dans leurs travaux portant le même intérêt au TS dans la citation suivante : « *By this experimental clarification, we could conclude that the single-noun term's power of generating compound noun terms is useful and essential in ATR* » [11 : p.2][5].

Sur quel corpus spécialisé s'appliquera l'extraction automatique des STN à partir des TSN ? La description détaillée de notre corpus d'analyse (CA) se fait dans la section suivante.

B. Le corpus d'analyse (CA)

Notre corpus d'analyse (CA) consiste en un corpus bilingue aligné. A l'état brut, notre corpus figure dans les textes des résolutions de l'Assemblée Générale de l'ONU. Il s'agit donc d'un corpus hyper spécialisé représentatif du domaine juridique et des domaines de compétence de l'Assemblée Générale (AG) de l'ONU. Les textes composant notre corpus comptent 363 résolutions dans chacune des langues française et arabe, soit au total 726 résolutions[6].

Le CA a subi une série d'opérations sophistiquées de préparation et de balisage structurel via le « langage à balises extensible (XML) » de l'anglais « eXtensible Markup Language ». Le balisage structurel, comme son nom l'indique, a pour rôle de sauvegarder deux types de données dans notre corpus :

- Les métadonnées d'archivage qui conservent « *les informations génétiques* » [9 : p.49] relatives à la description structurelle logique du corpus, y compris la spécialisation, le numéro de la résolution, la date, la session, la taille, etc.
- Les données textuelles linguistiques renfermant les différentes parties du texte des résolutions.

Le schéma XML de notre CA est conçu sur la base de 5 éléments textuels principaux et 124 éléments d'archivage subsidiaires, comme le montre la figure suivante :

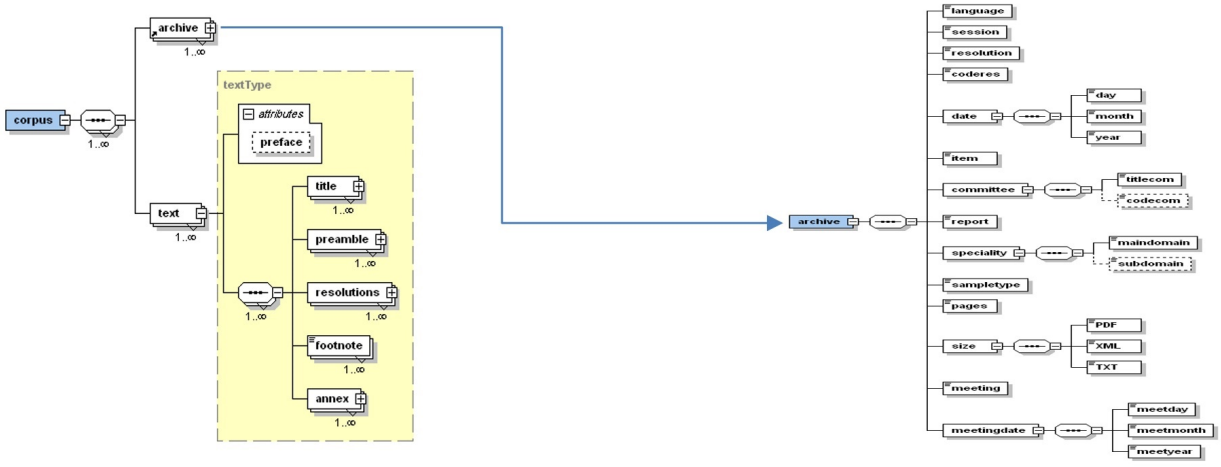


Figure 1 : Schéma binaire XML du balisage structurel de notre CA

A l'issue du balisage structurel, le CA passe par des phases de traitement automatique dont la plus importante figure dans le codage structurel reposant tout de même sur les balises structurelles du schéma XML. Le principe du codage structurel consiste à assigner à chaque élément textuel ses métadonnées d'archivage sans les fusionner avec les données textuelles

[5] La traduction française la citation est comme suit : « *Par cette clarification expérimentale, nous pourrions conclure que le pouvoir du terme simple de générer des termes complexes nominaux est utile et primordial dans la reconnaissance automatique des termes* ».

[6] Notre corpus constitue un sous-corpus généré d'un corpus volumineux et personnalisé aux fins de la présente étude, et ce dans le cadre des travaux de recherche dans la construction des corpus pour servir dans des projets de traitement automatique des langues naturelles, avec le concours d'une équipe de scientifiques compétents dans les domaines du TAL, de la linguistique computationnelle, de la terminologie computationnelle, de la linguistique de corpus et de la traductique. Cette équipe, dont je suis l'un des membres, vient de lancer une plateforme expérimentale sous le nom de « *Traduction et Technologie de la langue (TLT)* » de l'anglais « *Translation & Language Technology (TLT)* ». Le lien pour la plateforme expérimentale est comme suit : https://tlt-eg.com/. Le lien pour le portail dédié aux travaux dans la linguistique de corpus est : https://corpus.tlt-eg.com/

linguistiques, lesquelles seront objet d'extraction terminologique. Chaque élément porte donc un code commun pour les deux parties du corpus correspondant aux deux langues. Ce code structurel commun garantie pleinement l'alignement de notre CA à tous les niveaux possibles : les éléments, les paragraphes, les phrases, les propositions et les mots. Le CA aligné après ces maintes phases compte au total 1, 049, 069 mots et 86, 288 propositions.

C. Principe de l'extraction automatique bilingue des STN

En effet, l'adoption d'une méthode ou d'une autre pour l'extraction automatique des termes, qu'ils soient simples ou complexes, ne relève plus d'un choix forcé, bien au contraire, elle émane d'une réflexion raisonnée à propos de la problématique terminologique, alignant une gamme de variantes, y compris les unités terminologiques en jeu, les langues concernées par le travail terminologique, la taille du corpus et son mode de préparation, l'objectif du travail et les outils technologiques d'application. Toutes ces variantes oscillent entre le compromis et la contrainte comme l'affirme Drouin [4 : p.180] en ces termes :

> L'élaboration d'algorithmes d'acquisition automatique de termes réside ainsi dans l'art du compromis. Il faut savoir cerner l'ensemble des observations théoriques et empiriques faites au sujet des termes et les transformer en contraintes.

Partant de ce double principe du « compromis et contraintes », nous avons personnalisé notre méthode d'extraction automatique des STN selon nos objectifs terminologiques, notre potentiel informatique et le statut de notre corpus de travail déjà préparé et traité, tout en prenant en considération la spécificité de notre paire linguistique alignant la langue française avec la langue arabe qui pose maintes défis en matière du traitement automatique en général et de l'extraction automatique des termes en particulier. En outre, notre travail terminologique impose un double processus consécutif d'extraction automatique, le premier impliquant les Termes Noyaux Simples Nominaux (TNSN) qui sert de base pour le second visant les Syntagmes Terminologiques Nominaux (STN). Dans les deux processus, la méthode de repérage terminologique ainsi que le modèle d'extraction bilingue diffèrent tantôt dans la méthodologie adoptée, tantôt dans l'ordre du processus même. Dans les lignes suivantes, nous décrivons la méthodologie à laquelle nous avons eu recours dans les deux processus, et nous reléguons la description détaillée de l'ordre des phases de chacun des deux processus au point (5).

Dans le premier processus de l'extraction des TNSN, nous avons opté pour la méthode statistique dont la quantification parvient parfaitement à décrire et à interpréter les phénomènes terminologiques dans un temps record par rapport aux autres méthodes (linguistique et hybride). La méthode statistique convient également à l'abord commun de deux langues appartenant à des familles linguistiques différentes, telles le français (Langue A) et l'arabe (Langue B), étant donné que la langue arabe soulève des problèmes sérieux quant à l'automatisation de ses structures morphosyntaxiques. D'où vient l'importance de la méthode statistique pour notre étude. Quant au modèle d'extraction bilingue des TNSN, nous avons appliqué le premier modèle de la « double extraction des candidats-termes puis alignement », où l'extraction se fait au niveau de chaque langue séparément, à commencer par la langue A (le français), puis la langue B (l'arabe) ; après quoi nous opérons l'alignement entre les candidats-termes extraits via la méthode statistique.

Quant au second processus de l'extraction des STN, la méthode statistique est toujours de mise, mais elle est doublée par une nouvelle méthode que nous établissons sous l'intitulé de « *méthode structurelle* » afférant au balisage et au codage structurels XML de notre corpus aligné précédemment décrits. Ce balisage structurel assure parfaitement le repérage génétique des segments alignés de notre corpus à n'importe quel niveau de l'alignement : au niveau des textes du corpus, au niveau des paragraphes, au niveau des phrases, au niveau des propositions et au niveau des mots. Notre méthode s'avère être donc une « *méthode mixte statistique structurelle* », où l'extraction des candidats-termes STN se fait sur chaque langue du corpus aligné à part puis on procède à l'alignement ou la recherche des équivalents.

Avant de rentrer dans le vif de la description des phases de l'extraction automatique bilingue des STN, il faut avouer que tout travail terminologique partage un processus commun dans sa charpente globale, comprenant nombre de phases consécutives standard qui diffèrent intrinsèquement selon la méthode adopté (linguistique, statistique ou hybride) et le modèle suivi (monolingue ou bilingue). Ceci dit, nous pouvons résumer ce processus dans les phases suivantes :

1) Extraction automatique des candidats-termes (CT) : Toute extraction automatique implique une opération de segmentation en unités ou tokenisation (de l'anglais Tokenization), qui est un concept afférant au traitement automatique des langues (TAL) et désigne la décomposition d'« *une suite de caractères en « unités » : mots simples ou unités polylexicales* » [14 : p.162]. En terminologie, l'unité visée par la segmentation est une unité terminologique qui varie d'une étude à l'autre et par voie de conséquence modifie les algorithmes de segmentation. L'extraction automatique se base sur la segmentation et a pour rôle d'extraire les unités segmentées à l'aide de délimiteurs formels. Il faut noter que l'extraction automatique débouche ipso facto sur des listes des candidats-termes (CT) et non sur des

listes d'unités terminologiques définitives. Nous entendons par candidats-termes (CT) « *toute suite de caractères identifiée comme susceptible de constituer un terme spécialisé* » [12 : p.18]. Cette définition des CT dénote que tout système d'extraction automatique manque certes de perfection et implique une intervention humaine du terminologue pour décider du statut terminologique des unités extraites à travers une série de phases manuelles et/ou automatiques de pointe, telle le dépouillement et la validation.

2) Dépouillement des listes des CT : Par dépouillement nous désignons la « *distinction entre les unités qui ont un statut terminologique saillant et les autres unités appartenant au vocabulaire de la langue générale* » [8 : p.205]. Cette tache assumée par le terminologue consiste principalement, dans un premier temps, à distinguer entre terme et non-terme se basant sur l'intuition linguistique et l'expertise du terminologue[7], et repose, dans un second temps, sur des procédures systémiques de nature à raffiner les résultats de l'extraction automatique bilingue, à savoir : le nettoyage, l'élagage, et le tri, étapes effectuées automatiquement certes sous la direction du terminologue. Le nettoyage se situe au niveau de l'élimination automatique des coquilles qui persistent toujours dans les listes après extraction et qui ont échappé au maints processus de nettoyage du corpus même. L'élagage des listes CT renferme deux procédures : primo, l'élimination des unités qui n'affichent aucun statut terminologique ou qui n'appartiennent pas au type d'unité terminologique visée par l'étude, et secundo, la normalisation formelle des CT ayant un statut terminologique mais qui affichent une déformation par rapport à la forme standard de l'unité terminologique. Le Tri consiste à la mise en ordre des données de la liste CT suivant deux modèles : le tri alphabétique et le tri par ordre de fréquence. Ces deux modèles d'ordonnancement permettent au terminologue de concevoir les CT selon des optiques différentes pour juger de leur statut terminologique. Reste à signaler que le tri est une opération récurrente dont le modèle varie selon la phase du travail et l'objectif visé. Il est à noter que le dépouillement peut intégrer d'autres étapes selon la spécificité de l'étude ou des unités terminologiques à extraire comme nous verrons dans la description des phases d'extraction automatique que ce soit des TNSN, ou des STN. A l'issue de la phase du dépouillement, la liste des CT se transforme en une liste de termes.

3) Alignement des listes de termes : L'alignement désigné ici concerne l'appariement de terminologie et consiste à « *mettre en correspondance des unités terminologiques sources et leurs équivalents terminologiques cibles* » [1 : p.225]. En d'autres termes, l'alignement des listes de termes est une recherche d'équivalents des unités terminologiques de la langue source A dans la langue cible B. Deux méthodes sont mises au point pour l'alignement de terminologie : la première consiste dans l'application des algorithmes d'appariement entre les deux listes de termes extraits automatiquement de manière séparée d'un corpus parallèle, se basant comme dans l'alignement de textes au niveau des mots sur des indices typographiques, linguistiques et/ou statistiques ; la seconde consiste à extraire automatiquement les unités terminologiques du corpus source uniquement, puis on applique des algorithmes d'alignement sur le corpus cible pour en extraire les équivalents, reposant également sur des indices typographiques, linguistiques et/ou statistiques. Nous ajoutons à ces deux méthodes une troisième qui se base sur des indices structurels. La méthode de l'alignement dans notre recherche opte conjointement pour les première et troisième méthodes se basant se les indices statistiques et structurels à la fois. L'alignement se fait au niveau des termes simples nominaux, puis au niveau des syntagmes terminologiques nominaux.

4) Validation des listes des termes : Il s'agit de la vérification finale du statut terminologique des termes extraits automatiquement et issus du dépouillement selon une source fiable et crédible renforçant la détermination préliminaire par le terminologue du statut terminologique des CT. Selon El Bacha [1 : p.310], la validation constitue « *un travail de vérification et d'accréditation des résultats d'une étude terminologique* ». En effet, la validation peut se faire au fur et à mesure du travail du dépouillement, ou peut être reléguée vers la fin du travail terminologique. Les sources de validation manuelle ou automatique varient entre les corpus de référence, les banques de termes et les bases de données terminologiques (BDDT).

Les phases susmentionnées sont poursuivies littéralement dans notre recherche, mais varient légèrement entre les deux processus d'extraction des TNSN et des STN. Dans la section suivante, nous décrivons en détails les phases de notre méthode d'extraction automatique bilingues des STN.

5 PHASES DE L'EXTRACTION AUTOMATIQUE BILINGUE DES STN

Comme déjà mentionné à maintes reprises, notre méthode d'extraction automatique de termes s'articule autour de deux processus consécutifs et complémentaires, dont les phases et les étapes seront étalées en détails dans les points suivants :

[7] La distinction entre terme et non-terme ainsi que la détermination du statut terminologique des CT se basant sur l'intuition du terminologue n'est jamais aléatoire. Elle se fait selon des critères scientifiques comme ceux repérés par El Bacha [13] dans son article « *Plans de distinction entre terme et non-terme comme indices de repérage automatique des termes* ». Ces critères sont donc des indices concrets pour la détermination du statut terminologique des CT au niveau manuel ainsi que pour l'identification des unités terminologiques dans les corpus spécialisés au niveau automatique [8 : pp.207-211].

A. *Extraction automatique bilingue des termes noyaux simples nominaux (TNSN)*

1) Extraction bilingue des candidats-termes simples : La première étape de l'extraction figure dans la segmentation en unités. Les unités concernées par la segmentation dans cette phase sont les termes simples (TS) en général et les termes simples nominaux (TSN) en particulier. Nous avons certes exclu les termes simples verbaux et adverbiaux. Par contre, nous avons retenu les termes simples adjectivaux, étant donné que les adjectifs servent d'expansions pour les noyaux nominaux dans les STN comme nous verrons au point (6).

La deuxième étape de l'extraction applique deux mesures statistiques sur le CA, à savoir : la fréquence brute[8] et la répartition. Par fréquence brute, nous désignons le nombre total d'occurrences d'une forme graphique dans le CA. Le comptage de la fréquence des formes graphiques se fait selon leur forme brute ou initiale, c'est-à-dire lemmatisée[9].

En effet, pour être significative, la fréquence ne suffit pas à elle seule comme mesure statistique fiable, puisqu'une forme graphique (une unité terminologique en l'occurrence) peut être considérablement récurrente dans un corpus, mais dans un nombre restreint de textes. Cela dit, son potentiel terminologique (PT) ainsi que son statut terminologique restent fort douteux. En outre, dans un corpus comme le nôtre composé de textes des résolutions de l'AG, la fréquence manque de fiabilité et nécessite un moyen de calculer le poids des unités terminologiques par rapport au corpus en entier. C'est pourquoi, nous avons doublé la fréquence par la mesure de la répartition, pour ainsi pouvoir calculer le nombre de textes dans lesquels une forme graphique apparait. La concurrence des deux mesures statistiques comme la fréquence et la répartition consolide les indices d'identification du statut terminologique des TNSN.

Le comptage statistique de la fréquence et la répartition est exécuté par des algorithmes XQuery, langage de programmation dédié à la manipulation et le traitement des données XML, telles celles figurant dans notre CA. Le comptage statistique s'applique sur les deux parties du CA séparément. La figure 2 ci-dessous illustre le résultat du comptage statistique XQuery où chaque CT est assorti de deux mesures, l'une pour la fréquence et l'autre pour la répartition :

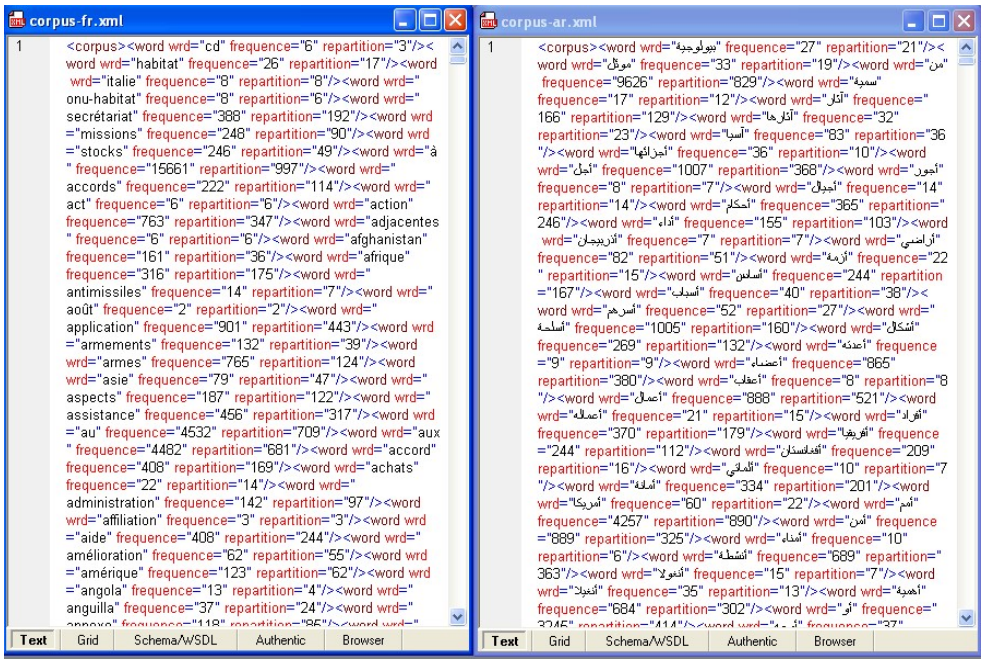


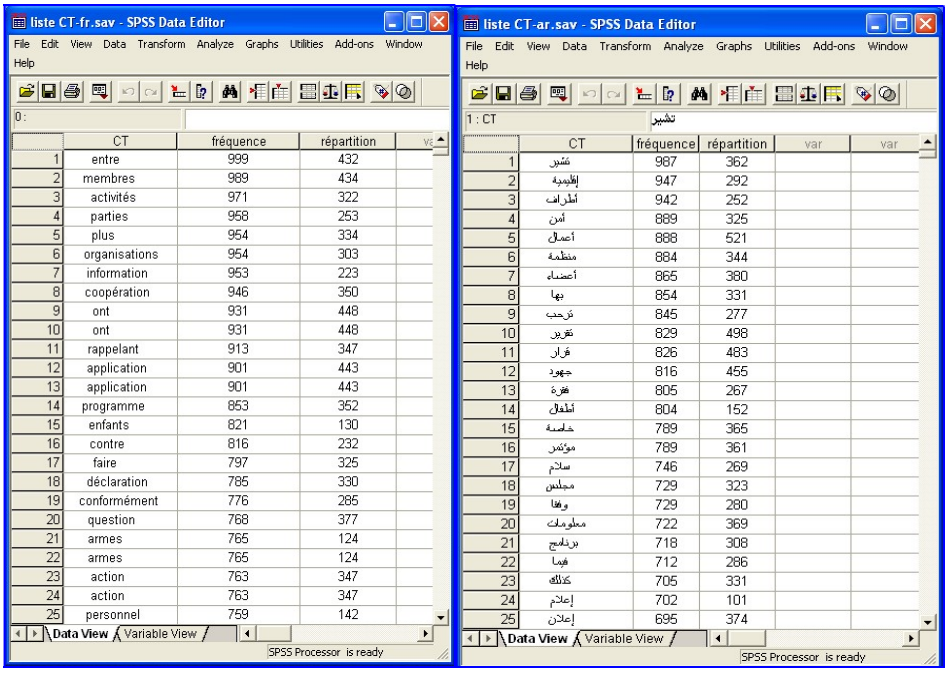
Figure 2 : Comptage statistique XQuery dans les deux parties du CA

A l'issue du comptage statistique, il a fallu procéder à l'exclusion des formes dont la fréquence dépasse les deux seuils minimal et maximal fixés pour l'extraction. Nous avons déterminé pour le seuil maximal la fréquence 1000 et pour le seuil minimal la fréquence 2. Les CT qui seront retenus donc sont ceux situés dans l'intervalle de 2 à 1000 occurrences sur la base de leur forme unique.

[8] On oppose la fréquence brute à la fréquence relative qui désigne le nombre relatif d'occurrences d'une forme graphique par rapport au nombre total des formes graphiques du corpus.

[9] Cette étape concernant la lemmatisation des CT et leur comptage statistique relève de la phase de dépouillement.

Il en résulte de l'extraction deux listes de candidats-termes (CT) simples comportant des unités nominales et adjectivales susceptibles d'être des unités terminologiques. Ces CT sont accompagnés par deux mesures statistiques et triés par ordre décroissant de la fréquence et la répartition, comme le montre la figure 3 suivante :



liste CT-fr.sav - SPSS Data Editor

	CT	fréquence	répartition
1	entre	999	432
2	membres	989	434
3	activités	971	322
4	parties	958	253
5	plus	954	334
6	organisations	954	303
7	information	953	223
8	coopération	946	350
9	ont	931	448
10	ont	931	448
11	rappelant	913	347
12	application	901	443
13	application	901	443
14	programme	853	352
15	enfants	821	130
16	contre	816	232
17	faire	797	325
18	déclaration	785	330
19	conformément	776	285
20	question	768	377
21	armes	765	124
22	armes	765	124
23	action	763	347
24	action	763	347
25	personnel	759	142

liste CT-ar.sav - SPSS Data Editor

	CT	fréquence	répartition
1	تشير	987	362
2	إطلاعية	947	292
3	أطراف	942	252
4	أمن	889	325
5	أعمال	888	521
6	منظمة	884	344
7	أعضاء	865	380
8	بها	854	331
9	ترحب	845	277
10	تقرير	829	498
11	قرار	826	483
12	جهود	816	455
13	فقرة	805	267
14	أطفال	804	152
15	خاصة	789	365
16	مؤتمر	789	361
17	سلام	746	269
18	مجلس	729	323
19	وفقا	729	280
20	معلومات	722	369
21	برنامج	718	308
22	فيما	712	286
23	كذلك	705	331
24	إعلام	702	101
25	إعلان	695	374

Figure 3 : Listes CT simples issues de l'extraction automatique

Il faut toutefois noter que ces listes renferment certes des coquilles et des imperfections dont l'interférence de certaines unités qui n'appartiennent pas à la catégorie nominale, lesquels seront rectifiés lors de la phase du dépouillement.

2) Dépouillement des listes des candidats-termes simples : La première étape du dépouillement consiste dans le nettoyage des coquilles dans les deux langues à travers soit leur élimination totale soit leur rectification. Parmi ces coquilles figurent les espaces blanches, les signes de ponctuation, des caractères spéciaux, des lettres dissociées des mots, des mots déformés, la numérotation par les lettres dans les deux langues. Cette élimination et/ou rectification s'appliquent automatiquement aux deux listes[10].

La deuxième étape du dépouillement figure dans l'élagage des listes des CT comportant deux grandes étapes : l'élimination et la normalisation. L'élimination dans cette phase vise des unités possédant un statut terminologique mais qui ne sont pas relatives à notre problématique de recherche, telles les sigles, qui ne seront pas partie intégrante dans des STN. Cette élimination est filtrée automatiquement dans une liste d'exclusion. Quant à la normalisation, elle se manifeste dans la récupération de la forme standard d'un CT considéré comme unité terminologique, citons entre autres : la récupération des traits d'union dans des CT simples formés par composition, étant omis lors de la coupure de lignes abusive ou de la segmentation, comme [souscomités✗ - sous-comités✓] - [cessezle-✗ - cessez-le-feu✓], la normalisation des CT arabes assortis de morphèmes flexionnels finals indiquant une marque diacritique spéciale [... / قرار ←*قرارٍ / مجلس ←*مجلسٍ], la récupération des CT arabes auxquels sont accolés des prépositions [←*بتقرير ... / برنامج ←*لبرنامج / تقرير], la récupération des CT arabes ayant à leur final des morphèmes flexionnels indiquant le duel ou le pluriel, ou marquant des pronoms accolés en arabe et qui sont indépendants en français [/ موظف ←*موظفو ... / حقوق ←*حقوقهم / قرارات ←*قراراتها / قرار ←*قرارين]. Parmi l'élagage figure la fusion des doublons et l'addition de leur comptage statistique.

La troisième étape du dépouillement consiste dans le tri ou la mise en ordre des CT dans les listes. Il faut préciser que le tri est une étape récurrente qui s'applique au fur et à mesure de l'avancée du travail du dépouillement. La première fois tout au début du dépouillement suivant l'ordre décroissant de fréquence et de répartition, la deuxième fois suivant l'ordre alphabétique des CT pour faciliter le repérage des unités ayant une parenté morphologique en vue de leur normalisation, ou des doublons en vue de leur suppression.

[10] Toute élimination se fait dans des listes d'exclusion qui seront objet d'étude ultérieurement dans la section consacrée à l'examen de la résolution d'extraction.

3) Validation des listes CT simples : Nous avons jugé bon d'accomplir la validation dans le processus d'extraction automatique des TSN avant l'alignement, étant donné que la validation est faite automatiquement à l'aide d'une liste de termes de référence déjà extraits et validés d'un corpus de référence volumineux. Le principe de validation se base sur l'émulation automatique entre la nouvelle liste des CT et une liste référentielle volumineuse d'unités terminologiques simples dans le même domaine, laquelle était le fruit d'études [1] et [8] et de projets postérieurs en terminologie computationnelle. Après comparaison automatique des listes, les unités communes sont retenues comme termes définitifs, et les CT ne figurant pas dans la liste sont objet d'une validation manuelle à travers la base de données de référence pour l'ONU, à savoir UNTERM[11]. Seules les unités retenues après validation manuelle sont celles figurant parmi les entrées UNTERM. Il en résulte après validation deux listes de TS : la liste française compte 1474 TS et la liste arabe 1327 TS. Il s'agit dorénavant d'unités terminologiques simples à proprement dit et il n'est plus question de candidats-termes pour ce premier processus.

4) Alignement des listes des TSN : Il s'agit à ce stade d'un alignement de terminologie entre deux listes et recherche des équivalences des TS de la langue A (français) dans la liste de la langue B (arabe). Avant de commencer l'alignement, les deux listes sont exportées vers des fichiers XML. Le principe de l'alignement consiste à réunir le codage structurel XML, déjà expliqué, ainsi que le codage statistique de la décomposition concernant le comptage des éléments et des propositions du CA et de les associer à chaque unité terminologique dans les deux listes. Des algorithmes XQuery s'appliquent sur les deux listes XML sur la base des deux codes susmentionnés. La mise en comparaison des codes structurels uniques entre les TSN de la langue A et les unités terminologiques de la langue B suffit parfaitement à réaliser l'alignement. Les algorithmes de l'alignement ainsi que son résultat peuvent être illustrés dans les deux figures 4 et 5 ci-dessous :

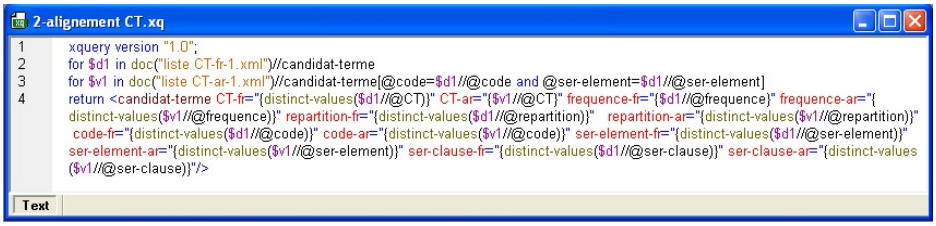


Figure 4 : Algorithmes d'alignement de terminologie XQuery

Figure 5 : Fichier XML illustrant le résultat de l'alignement

Le résultat de l'alignement est exporté de nouveau vers une base de données relationnelle, en vue d'effectuer l'appariement final qui consiste en : la dissociation des TS français qui possèdent plusieurs équivalents dans des entrées distinctes comme le terme [droit] qui possède deux équivalents [(qanu:n) قانون] et [(ħaq) حق] et la vérification des TS français dont l'équivalent arabe est un syntagme terminologique et qui seront analysés séparément (voir point 6).

5) Construction d'une base de données XML des TSN : La liste bilingue issue de l'alignement subit une dernière opération de révision et de validation pour les équivalences et une dernière exportation vers une base de données XML, en prélude à son exploitation en tant que noyaux nominaux dans l'extraction bilingue automatique des STN. La liste finale des TSN compte 820 unités terminologiques nominales simples et 107 unités terminologiques adjectivales simples, étant donné que nous avons décidé de maintenir les adjectifs dans la liste des TNSN puisqu'ils servent

[11] UNTERM est le portail terminologique des Nations Unies, dont l'interface est disponible dans les 6 langues officielles de l'ONU via le lien suivant : https://unterm.un.org/unterm2/fr/

d'expansions aux noyaux dans les unités syntagmatiques assumant des fonctions d'ordre qualificatif, déterminatif et/ou modificatif et étant donné leur apport dans la productivité et la génération syntagmatiques comme nous le verrons au point (6).

B. Extraction automatique des syntagmes terminologiques nominaux (STN)

1) Extraction automatique double de segments terminologiques du CA : Cette phase représente une particularité de notre corpus constitué à l'état brut des textes des résolutions de l'AG de l'ONU. Ces textes se composent d'une structure fixe qui s'articule sur des phrases se terminant par des virgules ou des points virgules, le point étant la marque de la dernière phrase uniquement. La normalisation des ponctuations dans la préparation du CA a unifié tous les signes en une virgule qui est le délimiteur de découpage. Ceci dit, nous avons procédé à une segmentation de notre CA selon ce délimiteur structurel. Et via le balisage structurel, ces segments sont associés à des codes structurels uniques. Le principe d'extraction automatique des segments terminologiques repose donc sur un algorithme d'appariement entre la base de données XML des TSN et le CA segmenté, à chaque fois que l'algorithme repère un segment où figure un TSN, le segment est jugé pertinent et significatif du point de vue terminologique, et par conséquent il est extrait. Cette opération d'extraction double débouche sur deux sous-corpus d'analyse (S-CA) dans les deux langues A et B. Les sous-corpus sont composés donc de segments de phrases terminologiques assortis de codes structurels et unis avec les TNSN assortis tout de même de leurs codes structurels et leurs données statistiques ainsi que leur étiquetage morphologique. Les figures suivantes ?? illustrent le résultat de cette phase en termes de S-CA FR et AR :

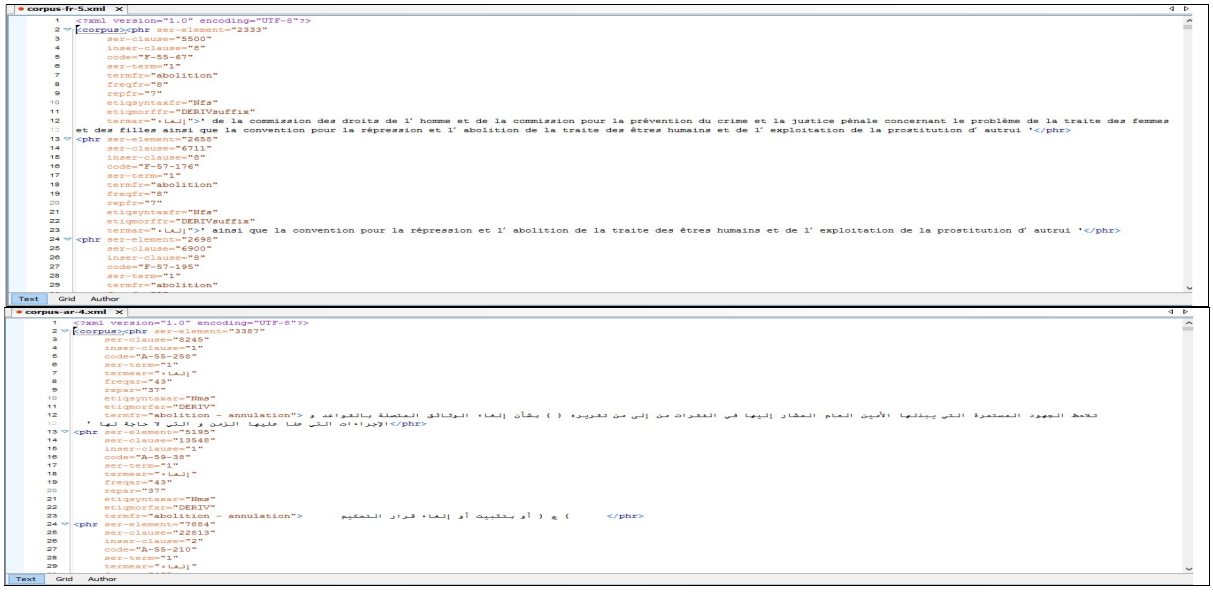


Figure 6 : Illustrations d'un échantillon des sous-corpus d'analyse FR et AR

2) Segmentation et extraction automatiques bilingue des CT syntagmatiques : Cette phase consiste en deux étapes complémentaires bilingues : la segmentation automatique et l'extraction automatique.

Il est question premièrement de la segmentation automatique des STN qui repose sur leurs noyaux, puisque chaque STN possède un noyau autour duquel se placent des expansions. Pour appliquer une segmentation automatique des STN, il faut au début déterminer la fenêtre de repérage terminologique, c'est-à-dire déterminer le nombre de mots, en l'occurrence les expansions, entourant le noyau nominal, en l'occurrence le TSN. L'ensemble du noyau et des expansions constitue la longueur du STN. Nous avons déterminé une fenêtre terminologique d'une longueur de 5 unités, y compris le TSN. Cette fenêtre couvre les deux langues française et arabe. Le schéma de la fenêtre terminologique reproduisant 6 modèles de STN peut être illustré par le tableau I suivant :

TABLEAU I

SCHEMA DE SEGMENTATION DES FENETRES TERMINOLOGIQUES

Modèle	Longueur	Fenêtre
1	2	1 - TNSN
2	2	TNSN + 1
3	3	1 - TNSN + 1
4	3	TNSN + 1 + 2
5	4	TNSN + 1 + 2 + 3
6	5	TNSN + 1 + 2 + 3 + 4

Il s'agit deuxièmement de l'extraction automatique qui s'applique sur les deux S-CA FR et AR, à travers un algorithme qui poursuit les TNSN dans les segments phraséologiques et commence leur segmentation selon le schéma de la fenêtre terminologique précité se situant entre un élément précédant ou suivant le TNSN jusqu'à 4 éléments lexicaux après le TNSN. La direction de l'expansion se lance à partir du TNSN et est réversible certes entre le S-CA FR (de la gauche vers la droite) et le S-CA AR (de la droite vers la gauche), sauf pour les modèles (1) où l'expansion se situe avant le TNSN et (3) où les expansions se situent avant et après le TNSN. Ces derniers modèles génèrent des STN, tels [bonne gouvernance (1-TNSN)] et [légitime défense individuelle (1-TNSN+1)]. L'extraction automatique bilingue basée sur le schéma de segmentation terminologique et le codage structurel XML unique reproduit 6 listes de candidats-termes syntagmatiques, comme dans la figure 7 ci-dessous :

```
<term-comp ST="bonne gouvernance "'
        count-comp="5"
        count-distr="5"
        termfr="gouvernance"
        termar="حكم"
        freqfr="38"
        repfr="21"
        etiqsyntaxfr="Nfs"
        etiqmorffr="DERIVsuffix"/>
```

```
<term-comp ST="حكم سليم"
        count-comp="1"
        count-distr="1"
        termear="حكم"
        termfr="gouvernance - jugement - sentence"
        freqar="196"
        repar="110"
        etiqsyntaxar="Nms"
        etiqmorfar="DERIV"/>
```

Figure 7 : Illustration des extraits de la phase de segmentation et extraction automatiques

3) Dépouillement des listes CT syntagmatiques : La phase du dépouillement des listes CT dans le processus de l'extraction bilingue des STN consiste dans plusieurs étapes. Il s'agit premièrement de la fusion des 6 listes dans chacune des langues dans une seule liste CT syntagmatique en français et en arabe. Après quoi, vient l'élimination des doublons résultant de la fusion ainsi que des coquilles. Une liste d'exclusion est également reproduite pour les collocations terminologiques qui ont pour noyau un TSN et qui ont été retenues par l'algorithme d'extraction des STN, à cause de leur confusion avec des unités terminologiques complexes binaires et ternaires.

Reste à signaler que certains CT syntagmatiques sont mis en veilleuse ou suspendus dans une liste d'exclusion pour s'assurer de la longueur exacte de leur fenêtre terminologique, comme à titre d'exemple le CT [commissions nationales contre la prolifération✗] dont la validation prouve une coupure incorrecte de sa longueur effective, à savoir [commissions nationales contre la prolifération des armes légères✓].

Finalement, vient le tour du tri dans le dépouillement des CT à la syntagmatique. En effet, le tri alphabétique est plus adéquat à l'ordonnancement des unités terminologiques complexes pour pouvoir détecter puis étudier des phénomènes syntagmatiques afférents à la variation ou la dérivation. Ce tri alphabétique est doublé par l'ordre décroissant de fréquence puis de longueur de la fenêtre terminologique. Il fait noter toujours que le dépouillement est appliqué sur les deux langues séparément.

4) Alignement terminologique des listes CT syntagmatiques : Notre méthode d'alignement terminologique diffère des deux méthodes précédemment signalées au point (4-C-3). Elle se base toujours sur le balisage structurel XML du CA et les codes structurels assignés à chaque segment textuel du corpus, depuis les paragraphes jusqu'aux segments les plus inférieurs. Les deux listes des CT issues du dépouillement subissent des algorithmes d'alignement qui consistent à récupérer le balisage structurel. Ces algorithmes procèdent par appariement entre les CT syntagmatiques et le CA, et chaque fois un CT est recensé au niveau des segments textuels des propositions, le codage structurel lui est assigné. Ces algorithmes XQuery sont appliqués au niveau des deux langues séparément. Une exportation des

résultats des algorithmes se fait vers deux fichiers XML correspondant aux deux langues, où chaque CT est doté de 3 types d'indices d'alignement :

- indices statistiques alignant la fréquence et répartition dans les deux listes CT syntagmatiques ;
- indices terminologiques alignant les TNSN dans les deux listes CT syntagmatiques ;
- indices structurels alignant les codes structurels dans les deux listes CT syntagmatiques.

L'alignement terminologique structurel ainsi décrit produit une liste des CT syntagmatiques de trois types : CT parfaitement alignés, CT partiellement alignés, CT non-alignés. Des opérations de vérification et de validation s'appliquent sur la liste CT issue de l'alignement pour décider du statut des CT partiellement et non-alignés.

5) Validation des listes CT syntagmatiques : Le processus de validation se fait selon deux sources, l'une manuelle et l'autre automatique. La validation manuelle se fait à travers la vérification des CT par rapport à la BDDT de l'ONU « UNTERM » à laquelle nous avons eu recours pour la validation des listes CT simples. Quant à la validation automatique, elle se fait à travers la comparaison des résultats de l'extraction syntagmatique et l'alignement terminologiques à un corpus de référence dont la taille dépasse celle de notre CA et dans le même domaine de compétence, et c'est ce même corpus duquel est extraite la liste des TS de référence qui a servi de support dans la validation des listes des TNSN. Cette validation double débouche sur une liste finale et validée comptant 3178 paires de syntagmes terminologiques nominaux (STN) FR-AR. Le tableau II suivant représente un échantillon de la liste finale des STN :

TABLEAU II

ECHANTILLON DE LA LISTE FINALE DES STN

s	code	STNFR	STNAR	TranscriptionStnAR	FreqFr	FreqAr	RepFr	RepAr
450	305	avis consultatif	فتوى	fatwa	75	45	10	7
451	482	avis consultatif de la cour internationale de justice	فتوى محكمة العدل الدولية	fatwa mahkamat alʕadl addawlja	41	44	5	6
452	305	avis consultatifs	فتوى	fatwa	9	45	3	7
453	306	avis de l'organisme paritaire	رأي الهيئة المشتركة	raʔj alhjʔa almuʃtaraka	6	1	1	1
454	2120	avis de vacance de poste	إعلان عن وظيفة شاغرة	ʔʕlan ʕn wazjfa ʃaʔira	18		4	
455	307	avis du secrétaire général	رأي الأمين العام	raʔj alaʔmjn alʕam	22	9	5	3
456	2121	avis juridiques	استشارات قانونية	ʔistʃarat qanwnjja	6		1	
457	1143	avis juridiques	المشورة القانونية	almaʃwra alqanwnjja	6	8	1	2
458	1046	banque internationale	البنك الدولي	albank addawlj	8	34	3	13
459	2122	banques multilatérales	مصارف متعددة الأطراف	maṣarif mutaʕadidat alaʔṭraf	3		1	
460	309	barème des quotes-parts	جدول الأنصبة المقررة	ʒadwal alaʔnṣiba almuqarara	27	9	4	3
461	1166	base légale	قاعدة قانونية	qaʕda qanwnjja	1	14	1	2
462	458	biens confisqués	الممتلكات المصادرة	almumtalakat almuṣadara	10	2	2	1
463	314	blanchiment d'argent	غسل أموال	ɣasl ʔamwal	26	2	6	1
464	314	blanchiment de fonds	غسل أموال	ɣasl ʔamwal	4	2	1	1
465	315	blanchiment de fonds publics	غسل أموال عمومية	ɣasl ʔamwal ʕumwmjja	2	2	1	1
466	316	blanchiment de fonds publics soustraits	غسل أموال عمومية مختلسة	ɣasl ʔamwal ʕumwmjja muχtalasa	2	2	1	1
467	317	blanchiment de l'argent	غسل الأموال	ɣasl alʔamwal	61	58	6	12
468	318	blanchiment du produit du crime	غسل العائدات الإجرامية	ɣasl alʕaidat alʔiʒramjja	4	1	1	1
469	2123	bonne foi	حسن النية	ħusn annjja	99	2	7	2
470	895	bonne gouvernance	الحكم الرشيد	alħukm arraʃjd	58	27	14	10
471	895	bonne gouvernance	حكم سليم	ħukm saljm	58	1	14	1

6 RESOLUTION DE L'EXTRACTION AUTOMATIQUE BILINGUE DES STN

Par résolution de l'extraction nous désignons le degré ou le niveau d'exactitude et de précision de notre méthode d'extraction automatique bilingue des STN qui s'est traduite en détails dans les phases susmentionnées dans la section (4). Vu la complexité de notre méthode, la résolution peut être définie selon plusieurs indices. Dans les points suivants, nous examinons ces indices sur la base des deux processus d'extraction soit des TNSN soit des STN.

A. Indices de la résolution de l'extraction

En général, il existe 4 indices en vigueur pour la résolution, soit disant l'évaluation des résultats, de la méthode d'extraction, à savoir : la précision, le rappel, le bruit et le silence. La Précision permet de mesurer le nombre de CT valides et pertinents extraits par un système donné par rapport au nombre total des CT extraits. Le Rappel calcule les CT valides et pertinents par rapport aux termes d'une liste de référence. Le Bruit calcule le nombre de CT non valides qui sont exclus de la liste finale des unités terminologiques. Le Silence concerne le nombre d'unités terminologiques non extraites dans la liste des CT. Les statistiques issues de nos listes STN nous octroient des indices de précision d'environ 77.2%, un taux de bruit d'environ 22.8% et un taux de silence estimé à 13.4%. Ces pourcentages sont loin d'être absolus ou définitifs, puisque nous possédons une méthodologie de contrôle de qualité et de compensation des lacunes appliquée à maintes reprises sur les résultats de chaque phase.

En effet, notre méthode d'extraction automatique nous a permis de créer un système de contrôle de qualité et de suivi de chaque phase, étant donné que chaque filtrage ou traitement appliqué sur les listes des CT simples ou syntagmatiques passe par des canaux dont les résultats sont enregistrés dans des listes d'exclusion. Ces listes d'exclusion subissent des phases de vérification et de récupération. A titre d'exemple, en appliquant la méthode statistique sur l'extraction des TNSN et en précisant un seuil minimal de 2 occurrences, nous risquons donc l'apparition des hapax qui sont des CT dont la fréquence compte 1. Une liste d'exclusion est produite pour les hapax et subit une vérification pour filtrer les CT à statut terminologique, et ces hapax sont récupérés dans la liste principale des CT. Ces hapax comptent 86 unités terminologiques additionnées aux listes TNSN valides. Ces listes d'exclusion nous permettent également de détecter les ST alignés qui représentent une variation formelle par rapport à la forme standard répertoriée dans la liste de référence ou dans le corpus de référence. Ces STN reflétant une telle variation atteignent un pourcentage de 7.2% pour le corpus source français et 8% pour le corpus cible arabe. Une série de phases de vérification est ainsi appliquée au fur et à mesure que nous progressons dans notre méthode.

Le principe de contrôle de qualité de la méthode d'extraction automatique des STN permet également de vérifier les règles et les fondements de travail terminologique. L'un de ces fondements figure par exemple dans la détermination de la fenêtre de segmentation terminologique reposant sur des travaux ultérieurs, tels ceux de Lapierre [16], Daille [17], Jacquemin (1996) et Drouin (2002), qui prouvent que le taux élevé de concentration des ST se situe au seuil des unités binaires et ternaires. Notre système de contrôle de qualité prouve, à travers l'analyse de la longueur de notre liste terminologique syntagmatique, que la longueur des STN bilingues est concentrée autour des unités binaires et ternaires, soit d'un taux de 70.4% pour le français et 84.3% pour l'arabe, contre 29.6% de STN dépassant trois unités lexicales en français et 15.7% en arabe. Le tableau suivant illustre la répartition des STN selon les fenêtres linguistiques :

TABLEAU III

REPARTITION DES STN BILINGUES SELON LES FENETRES TERMINOLOGIQUES

Fenêtre linguistique des STN-FR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Less than or equal 3	2243	70.4	70.4	70.4
	more than 3	944	29.6	29.6	100.0
	Total	3187	100.0	100.0	

Fenêtre linguistique des STN-AR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Less than or equal 3	2687	84.3	84.3	84.3
	more than 3	500	15.7	15.7	100.0
	Total	3187	100.0	100.0	

Notre système de contrôle de qualité permet un ajustement de la résolution de la méthode et règle ses possibles inconvénients ou imperfections. C'est là une originalité de notre processus.

B. Originalité et productivité de la nouvelle méthode d'extraction automatique

L'originalité de notre travail terminologique se trouve intimement liée au principe de productivité terminologique sur laquelle est fondée notre nouvelle méthode dés le début. Nous pouvons résumer les points forts et la particularité de notre méthode dans les lignes à suivre :

1) Dépister des équivalences entre termes simples et syntagmes terminologiques : le recours au codage structurel permet de fournir l'entourage contextuel des unités terminologiques extraites, ce qui offre la possibilité de dépister les équivalents syntagmatiques arabes des TSN français, bien que l'extraction était axée sur le niveau simple. Par contre, les équivalents simples arabes ont été dépistés pour les STN français. Cette différence de complexité syntaxique entre les équivalents des listes bilingues atteint 0.8% de la totalité des deux listes dans les deux sens. Le tableau IV suivant représente des cas de ce type :

TABLEAU IV

EXEMPLES DE CAS D'EQUIVALENCE TERMINOLOGIQUE

TS	ST
administrations	مسؤولين حكوميين [masu:ʔu:li:n ħuku:mjji:n]
antidiscrimination	مناهضة التمييز [munahaḍat attamjjz]
autodétermination	تقرير المصير [taqrjr almaṣjr]
cessez-le-feu	وقف إطلاق النار [waqf ʔiṭlaq annar]

ST	TS
avis consultatif	فتوى [fatwa]
compensation des montants	تعويض [tʕwjḍ]
établissement pénitentiaire	سجن [siʒn]
poursuite judiciaire	مقاضاة [muqaḍa:]

2) Repérer la polysémie terminologique : Notre méthode d'extraction et d'alignement automatiques nous permet de repérer les cas de polysémie où le STN possède plusieurs équivalents. Cette polysémie est réversible entre les deux parties et compte 4.5% du français vers l'arabe et 6.3% de l'arabe vers le français, comme le montre le tableau V ci-dessous :

TABLEAU V

REPERAGE DE LA POLYSEMIE DANS LA LISTE DES STN BILINGUES

Polysémie

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absence de polysémie	2842	89.2	89.2	89.2
	Polysémie FR-AR	144	4.5	4.5	93.7
	Polysémie AR-FR	201	6.3	6.3	100.0
	Total	3187	100.0	100.0	

Nous reproduisons ci-dessous quelques exemples de polysémie dans la liste bilingue des STN :

(1) Droit	Droit de l'enfant	حق الطفل	[ħaq aṭṭfl]
	Droit de l'espace	قانون الفضاء	[qanu:n alfaḍa:ʔ]
	Droits de douane	التعريفات الجمركية	[attaʕri:fa:t alʒumrukjja]
(2) Fonctions	Fonctions d'appui administratif	مؤهلات الدعم الإداري	[mu:ʔhhla:t addaʕm alʔdari:]
	Fonctions diplomatiques ou consulaires	المهام الدبلوماسية أوالقنصلية	[almaha:m addblu:masjja ʔaw alqunṣuljja]
	Fonctions judiciaires	وظائف قضائية	[waẓaʔif qaḍaʔjja]
	Fonctions militaires	مهام عسكرية	[maha:m ʕaskarjja]
	Fonctions publiques	الخدمات العمومية	[alχadama:t alʕumu:mjja]
(3) Instance	Instance judiciaire	دعوى قضائية	[daʕwa qaḍaʔjja]
	Instance multilatérale de négociation	منتدى التفاوض المتعدد الأطراف	[muntada attafawuḍ almutaʕaddid alʔṭra:f]
	Instances internationales	هيئات دولية	[hajʔa:t dawljja]
(4) Instrument	Instrument de ratification	صك تصديق	[ṣak taṣdi:q]
		وثيقة التصديق	[waθi:qat attaṣdi:q]
(5) Puissances	Puissances administrantes	السلطات القائمة بالإدارة	[assulṭa:t alqaʔma blʔidara]
		الدول القائمة بالإدارة	[adduwal alqaʔma blʔidara]
(6) Règlement	Règlement global	نظام شامل	[nẓa:m ʃamil]
		تسوية شاملة	[taswi:a ʃamila]

3) Fournir le contexte terminologique : Le balisage structurel nous offre également l'option de repérer les équivalents dans leur contexte terminologique équivalents tout de même, ce qui sert de grand support pour la traduction spécialisée ainsi que l'enseignement de la terminologie bilingue.

4) Automatiser l'extraction au niveau du corpus arabe : L'un des principaux atouts de notre méthode réside dans l'intégration de la langue arabe dans un même système automatique d'extraction terminologique avec la langue française. En effet, et eu égard à « *l'inadaptation de la langue [arabe] aux traitements informatiques ou, plutôt, de ceux-ci à la langue [arabe]* » [15 : p.25], la langue arabe est toujours problématique en matière du traitement automatique. C'est pourquoi, le balisage structurel a parfaitement assuré un moyen de rapprochement et d'appariement entre deux langues de statut différent.

5) Dresser le schéma de la productivité terminologique : C'est là toute l'originalité de notre méthode fondée sur les TNSN en tant qu'indices de repérage des STN. En effet, le nombre de TNSN bilingues extraits comme noyaux pour les STN bilingues atteint un nombre de 382 pour le S-CA-FR et 484 pour le S-CA-AR, soit respectivement 60.1% et 51.7%. Ces pourcentages équivalent au taux de productivité de génération syntagmatique des TNSN, c'est-à-dire la capacité des TNSN de reproduire des STN. Le taux de non- productivité est estimé à 39.9% et 48.3%. Or, cette non-productivité de certains TSN au niveau des noyaux extraits du S-CA prouve une productivité au niveau des expansions, autrement dit, les TSN ont opéré comme indices doubles de noyau et d'expansion, et le système procède à extraire les STN dans lesquels les TSN sont en position de noyau ou d'expansion. Et pour les STN, où le TSN figure en expansion, de nouveaux noyaux TSN sont générés. Les STN génèrent donc de nouveaux noyaux nominaux simples qui ont échappés à l'extraction des TSN, à cause de leur statut terminologique nul ou incertain (NT) à l'état simple comme [année - annuaire - certitude - comportement]. La question qui s'impose donc : comment ces noyaux ont été extraits et de quels ST ? Ces noyaux acquièrent un sens spécialisé et une valeur terminologique à cause de leurs modificateurs adjectivaux terminologiques qui figurent en tant que TS adjectivaux dans la liste de TSN, comme déjà mentionner, et qui ont position d'expansion dans les STN, tels [internationale – juridique - criminel] et ce dans des STN tels [année internationale - annuaire juridique - certitude juridique - comportement criminel]. Il est à noter

que la contribution des TS adjectivaux en position d'expansion de la liste des TS dans l'extraction des STN atteint 8.9% au niveau des STN-FR et 11.1% des STN-AR. Ce cycle ininterrompu où les TSN en position de noyau ou d'expansion génèrent des STN qui à leur tour reproduisent de nouveaux noyaux et de nouvelles expansions terminologiques est appelé « *cycle vital de génération syntagmatique* » [1 : pp.300-301]. Ceci dit, notre méthode d'extraction automatique bilingue des STN offre une vitalité et une permanence au travail terminologique et octroie une richesse et une objectivité aux résultats terminologiques, étant donné que notre fondement du travail est le TSN qui est un indice objectif et distinctif des STN, dont la complexité les confond avec d'autres unités linguistiques d'ordre général tels les syntagmes libres, les expressions à mots multiples et les expressions figées. Seul le TNSN est un trait distinctif et délimiteur automatiquement.

6) Créer un environnement de travail terminologique intégral : Il s'agit là d'un avantage et d'un inconvénient à la fois. Notre méthode s'appuie sur une charpente solide du travail terminologique où la préparation et le traitement du CA sont exécutés dans un environnement intégral et harmonieux regroupant toutes les phases terminologiques. En l'absence d'un tel environnement, il sera impossible d'envisager la même performance et les mêmes résultats.

7 CONCLUSION

La présente recherche s'insère dans le cadre des études en terminologie computationnelle qui se préoccupe du traitement automatique terminologique des corpus spécialisés. Nous avons tenté d'aborder les STN et la problématique de leur extraction automatique selon une nouvelle méthode que nous avons mise en place et pratiquée lors de plusieurs études ultérieures. Pour ce faire, nous avons passé en revue, dans un premier temps, les méthodes connues d'extraction automatique de termes, selon trois optiques : les unités visées par l'extraction, la méthodologie adoptée dans l'extraction, les langues impliquées dans l'extraction. Dans un second temps, nous avons procédé pour une exposition du principe de notre nouvelle méthode d'extraction automatique bilingue des STN basés sur leurs noyaux TNSN et sur le balisage structurel XML de leur corpus aligné. Une description détaillée des phases d'extraction automatique des TNSN et des STN est également objet d'étude déterminant le savoir-faire de chacune d'elle depuis l'extraction bilingue, passant par la segmentation, le dépouillement, l'alignement jusqu'à parvenir à la validation, phase définitive de tout travail terminologique.

Au terme de notre étude, nous avons tenté de porter un avis évaluatif sur notre méthode en examinant la résolution de l'extraction des STN, à la lumière de l'application du balisage et codage structurels XML et sur la base des TSN comme noyau terminologique non seulement des unités terminologiques, mais également du travail terminologique. Cet examen nous a conduit à des résultats qui attestent d'une efficacité et reflètent une indéniable fiabilité par rapport aux deux langues objets d'étude.

BIBLIOGRAPHIE

[1] M. El Bacha, « Les syntagmes terminologiques nominaux entre le français et l'arabe. Problématiques de l'extraction, du traitement automatique et de la traduction. Étude de terminologie computationnelle», Le Caire, 2017.

[2] M.-C. L'Homme, La terminologie : principes et techniques, Montréal : Les Presses de l'Université de Montréal, 2004, p. 278.

[3] D. Bourigault, « Lexter, un logiciel d'Extraction de TERminologie. Application à l'acquisition des connaissances à partir de textes », 1994.

[4] P. Drouin, « Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés », Université de Montréal, Montréal , 2002.

[5] E. Gaussier, "General considerations on bilingual terminology extraction", *Natural Language Processing,* vol. 2, p. 167–183, 2001.

[6] S. I. C. Cruz, « Analyse de la variation terminologique en corpus parallèle anglais-espagnol et de son incidence sur l'extraction de termes bilingue », Université de Montréal , Montréal , 2004.

[7] D. Bourigault, « Analyse syntaxique locale pour le repérage de termes complexes dans un texte », *TAL,* vol. 34, n° %12, pp. 1-15, 1993.

[8] M. M. El Bacha, « Modalités de formation des termes français et leur traduction arabe dans les textes des résolutions de l'Assemblée Générale des Nations Unies [2001-2005] - Analyse linguistique informatique », Université d'Ain-Chams, Le Caire, 2009.

[9] هـ. م. المالكي، "إشكاليات تهيئة الذخائر اللغوية وبنائها حاسوبيًا"، *أواصر*، رقم 2، 56-28 .pp، إبريل 2009.

[10] M.-C. L'Homme, Initiation à la traductique, 2e édition revue et augmentée éd., Montréal, Québec: Linguatech, 2008, p. 317.

[11] H. Nakagawa and T. Mori, "A Simple but Powerful Automatic Term Extraction Method", in *Proceedings of the*

Second International Workshop on Computational Terminology, Stroudsburg, PA, USA, 2002.

[12] M. Van Campenhoudt, « Linguistique de corpus et étude des vocabulaires spécialisés », chez *Séminaire*, Paris, 2002.

[13] M. M. El Bacha, « Plans de distinction entre terme et non-terme comme indices de repérage automatique des termes », *Logos,* n° %17, pp. 195-215, Novembre 2011.

[14] B. Habert, A. Nazarenko et A. Salem, Les linguistiques de corpus, Paris: Armand Colin, 1997, p. 240.

[15] A. Reguigui, Anatomie des syntagmes terminologiques arabes Analyse formelle et quantitative, Canada: Université Laurentienne de Sudbury, 2002, p. 421.

[16] I. Lapierre, « Recherche thématique sur le vocabulaire des valvulopathies. Étude terminologique de 50 dossiers terminographiques », Université du Québec, 1994.

[17] B. Daille, « Repérage et extraction de terminologie pour une approche mixte statistique et linguistique », *Traitement automatique des langues,* vol. 36, n°11-2, pp. 101-118, 1995.

BIOGRAPHIE

Maha El Bacha

Maitre de conférences en terminologie computationnelle et traductique, Département de français, Faculté des Langues (Al-Alsun), Université d'Ain-Chams. Titulaire d'un magistère et d'un doctorat en terminologie computationnelle et terminotique. Membre de la Société Egyptienne pour l'Ingénierie de la Langue (ESOLE), membre de la Société Egyptienne des Services et Solutions Linguistiques (EAGLS) et responsable du comité de formation dans EAGLS. Elle a publié plusieurs articles dans les domaines de la terminologie, la terminologie computationnelle, la terminotique et la traductique. Elle a participé à la traduction des chapitres 19 et 20 d'un ouvrage intitulé « Vingt ans qui bouleversèrent le monde : de Berlin à Pékin » publié en 2016 par le Centre National de Traduction. Elle a une expérience dépassant 5000 heures d'enseignement et de formation s'étalant sur 20 ans dans : la traductique (traduction assistée par ordinateur, interprétation assistée par ordinateur, sous-titrage, traduction automatique, localisation), la terminologie computationnelle et la traduction spécialisée. Elle a supervisé plus de 50 projets de transcription et sous-titrage. Elle maitrise plus de 134 applications en traductique, en terminotique et en ingénierie de la langue ainsi que nombre de langages de programmation. Concepteur d'un logiciel d'extraction automatique et de gestion de terminologie. Expert en formation dans plusieurs institutions : le Centre National de Traduction, l'Agence de Presse du Moyen-Orient (MENA), la Bibliothèque d'Alexandrie, la Société Egyptienne des Services et Solutions Linguistiques (EAGLS), Faculté des langues et de traduction - Université de Badr, l'Unité de traduction du Centre régional du Comité international de la Croix-Rouge pour le Moyen-Orient et l'Afrique du Nord.

ENGLISH ABSTRACT

New Method for Bilingual Automatic Extraction of Nominal Terminological Phrases Based on Their Cores and the XML Structural Markup of the Aligned Corpus

Maha M. EL BACHA

French department, Faculty of Languages (Al-Alsun), Ain Shams University, Egypt

Maha.elbacha@alsun.asu.edu.eg

Abstract: ***This research falls within the framework of computational terminology. It presents a new method for automatic bilingual extraction of Nominal Terminological Phrases based on their cores which are necessarily Simple Nominal Terms and exploiting the XML structural markup of their aligned bilingual corpus between French and Arabic. The research addresses in detail the founding principle and phases of our so-called statistical and structural method, starting with extraction and segmentation, going through many phases of automatic analysis and automatic alignment of terminology until reaching the final validation. At the end of the search, we carry out an examination of the resolution of the automatic extraction by virtue of the terminology results accompanied by statistics. The research also demonstrates the originality of the automatic bilingual extraction method appearing in a number of factors, the most important of which turns out to be productivity and the terminological generation of nominal cores and nominal terminological phrases.***

Keywords: ***nominal terminological phrase, automatic term extraction, simple term, terminological core, XML structural markup, aligned corpus, computational terminology.***

ARABIC ABSTRACT

منهجية جديدة للاستخراج الآلي للتراكيب المصطلحية الاسمية اعتمادًا على نواتها وعلى الترميز الهيكلي بلغة XML لذخيرتها المتوازية

مها مصطفى الباشا

قسم اللغة الفرنسية، كلية الألسن، جامعة عين شمس، مصر

Maha.elbacha@alsun.asu.edu.eg

ملخص: ***يندرج هذا البحث في إطار علم المصطلح الحاسوبي. وهو يقدم منهجية جديدة في الاستخراج الآلي ثنائي اللغة للتراكيب المصطلحية الاسمية، اعتمادًا على نواتها والتي يتحتم أن تكون مصطلحات مفردة اسمية، وانطلاقًا من استثمار الترميز الهيكلي بلغة XML والذي تم تطبيقه على ذخيرتها اللغوية المتوازية بين اللغتين الفرنسية والعربية. يتناول البحث بالتفصيل مبدأ هذه المنهجية الإحصائية الهيكلية ومراحلها، بدءًا من الاستخراج والتقطيع، مرورًا بمراحل عدة من الرصد المصطلحي الآلي وخلق التوازي المصطلحي آليًا، حتى الاعتماد المصطلحي. في ختام البحث، نقوم بدراسة مدى دقة الاستخراج الآلي في ضوء النتائج المصطلحية المدعومة بالإحصاءات. يستعرض البحث كذلك أصالة المنهجية الآلية ثنائية اللغة وخصوصيتها التي تتمثل في عدة عوامل، يتمثل أهمها في الإنتاجية والتوليد المصطلحي لمزيد من النواة المصطلحية الاسمية والمزيد من التراكيب المصطلحية الاسمية.***

الكلمات المفتاحية: ***تركيب مصطلحي اسمي، استخراج آلي للمصطلحات، مصطلح مفرد، نواة مصطلحية، ترميز هيكلي بلغة XML، ذخيرة لغوية متوازية، علم المصطلح الحاسوبي.***