Plans de la distinction entre terme et non–terme comme indices de repérage automatique des termes

Maha Moustafa El Bacha
Maître assistant au département de français
La Faculté d'Al-Alsun - Université de Ain-Chams
Le Caire - Égypte
Maha_m_elbacha@yahoo.fr

Résumé : Les techniques du repérage automatique des termes dans un corpus spécialisé ne se passent pas de principes théoriques distinguant entre terme et non-terme. Cette distinction théorique, quoique difficile, est parsemée de par les ouvrages de terminologie. Nous avons tenté de récapituler les principes distinctifs entre terme et non-terme selon notre conception terminologique, en vue de formuler un schéma complet exprimant note apport à cette problématique terminologique. Ce schéma sera constitué de plusieurs plans lesquels seront exploités, selon notre optique, comme indices d'identification automatique de termes, servant de support pour l'acquisition automatique de termes ainsi que pour le dépouillement automatique des listes des candidats-termes (CT), en prévision de leur validation comme termes.

Mots-clés : Terme - non-terme - langue de spécialité (LSP) - repérage automatique des termes - acquisition automatique de termes - dépouillement automatique de termes - candidats-termes.

عنـوان البحث باللغـة العربيـة :

مستويات التمييز بين المصطلح واللامصطلح كمؤشرات للاستدلال الآلي على المصطلحات

ملخص باللغة العربية: إن التقنيات الحاسوبية المتعلقة بالرصد الآلي للمصطلحات داخل الذخائر اللغوية المتخصصة لا تستغني في عملها عن المبادئ النظرية للتمييز بين المصطلح واللامصطلح. ونظرًا لأن عناصر التمييز النظري بين المصطلح واللامصطلح توجد مبعثرة فيما بين مراجع علم المصطلح الفرنسية، فقد أدى هذا الوضع إلى صعوبة التوصل إلى رؤية متكاملة لهذه الإشكالية. ولذلك، فقد استهدف البحث تجميع المبادئ الفارقة بين المصطلح واللامصطلح،

وذلك بهدف رسم صورة متكاملة يمكن أن تساهم بإضافة لتلك الإشكالية المصطلحية. وتتكوَّن هذه الصورة من عدة مستويات، سوف يتم توظيفها، وفقًا للمنهجية المُتبعة في البحث، كمؤشرات مُساعدة على الاستدلال الآلي على المصطلحات، مما يُدعم عمليات الاستخراج الآلي للمصطلح وكذلك الفرز المصطلحي لقوائم المصطلحات المُنتخبة تمهيدًا لاعتمادها كمصطلحات.

الكلمات المفتاحية: مصطلح - لامصطلح - اللغة المتخصصة - الاستدلال الآلي على المصطلحات - الاستخراج الآلي للمصطلحات - الفرز الآلي للمصطلحات - المصطلحات المُنتَخَبَة.

1. Introduction :

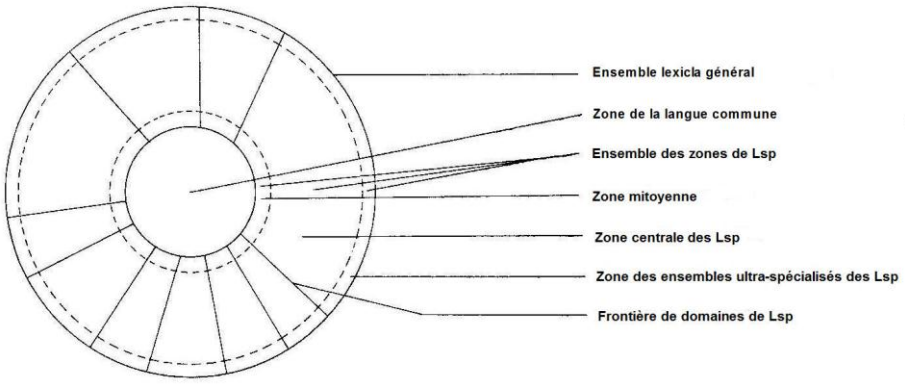
Le travail de repérage automatique de terminologie se base, dans un premier temps, sur une identification manuelle des termes. Laquelle nécessite que soient déterminés certains principes de distinction entre terme et non-terme. Cette distinction, plutôt théorique et méthodique, permet d'isoler le terme, dans un contexte spécialisé, du mot général ou de ce qu'on appelle le « non-terme ». Dans un second temps, Cette identification manuelle, vers la voie de l'automatisation, se prête les critères de la distinction théorique pour alimenter le système automatique par des étapes concrètes d'isolement du terme.

> « **L'étape principale du dépouillement est le repérage de termes qui implique, lors de la lecture systématique du corpus, une identification manuelle des termes. Ce travail, plutôt fastidieux et mécanique, de par son aspect répétitif, est un candidat idéal pour l'automatisation**. » (Drouin, 1997 : 1)

C'est cette dichotomie entre distinction manuelle entre terme et non-terme et indices d'identification automatique de termes que nous tenterons de traiter dans la présente étude. Laquelle suit une démarche méthodique en dressant six plans de distinction entre terme et non-terme partant du plus abstrait au plus concret (EL Bacha, 2009 : 53), puis aborde l'exploitation de ces plans distinctifs dans le repérage automatique des termes, partant primo de l'explication du concept du repérage automatique de terminologie pour accéder aux indices d'identification automatique de termes de par les plans distinctifs précités.

2. Terme et non-terme dans le discours spécialisé

Toute délimitation d'une unité terminologique, simple ou complexe, doit se situer dans le cadre d'un contexte spécialisé qu'est le discours spécialisé que l'on qualifie aussi de « langue de spécialité (LSP) » par opposition au « langue commune (LC) ». Il faut tout d'abord préciser que la LSP constitue un sous-système de la LC[(1)]. Guy Rondeau, quant à lui, nous a visualisé comment ce système compliqué de la langue englobe ces sous-systèmes par le schéma suivant (1984 : 25) :

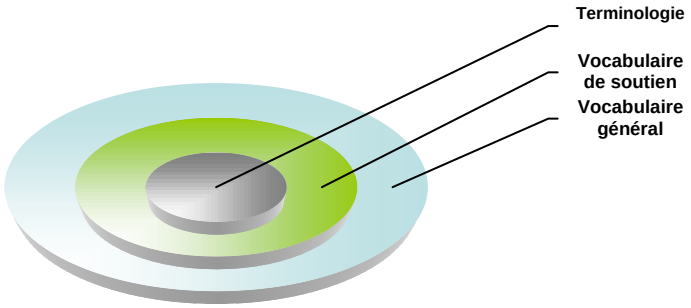


C'est ainsi que, dans le discours spécialisé, LSP et LC se côtoient par des proportions variantes selon le degré de spécialisation du texte ou du corpus : plus le degré de spécialisation augmente plus la proportion des termes augmente par rapport aux mots généraux, et plus les zones des LSP deviennent imperméables vu la technicité du texte qui bloque toute interférence entre mot général et terme. Le degré de spécialisation du corpus décide donc de la perméabilité des frontières entre LC et LSP.

D'autre part, certaines études des caractéristiques du discours spécialisé ont tendance à effacer les frontières rigides entre LSP et LC et à simplifier la tâche de distinction. Deux études sont à citer en la matière.

Le discours spécialisé, selon toutes tentatives de description, se présente par une structure hiérarchisée que nous reproduisons selon deux optiques presque identiques. La

première optique décrit la structure du discours spécialisé comme trois *« cercles concentriques »* (Gémar, 1991: 276-277) dont le noyau est la terminologie, le cercle suivant est le vocabulaire de soutien composé des cooccurrents du domaine spécialisé en cause et le dernier cercle la plus large est le vocabulaire général, autrement dit la langue générale dont le nombre des unités est illimité. Le schéma des cercles de Gémar se présente ainsi:



Suivant le même principe de classification, Maria Teresa Cabré (1991: 58) distingue entre trois couches de lexique dans le discours spécialisé, à savoir: le lexique général (équivalent à la langue générale), le lexique spécialisé ou lexique charnière (équivalent au vocabulaire de soutien), et la terminologie proprement dite.

Ceci dit, le terme dans le discours spécialisé se trouve mêlé aux autres éléments du texte y compris les éléments de la langue commune. D'où la nécessité de pouvoir relever le défi de distinction entre le terme proprement dit et le mot général en vue de servir les phases ultérieures du travail terminologique dont le repérage automatique est la plus importante, et ce comme l'indique Patrick Drouin :

> « **Il faut cependant se rendre à l'évidence que les documents spécialisés ne contiennent pas que des termes et qu'on y trouve aussi des mots. Ces derniers agissent, notamment, comme des charnières nécessaires à l'articulation du discours du spécialiste. Le défi qui se présente au terminologue est donc de réussir à distinguer les termes des mots dans l'ensemble des unités lexicales qui se présentent à lui.** » (Drouin, 2002 : 16)

Aussi, la section suivante sera-t-elle axée sur les différents plans de distinction entre terme et non-terme.

3. Principes de distinction entre terme et non-terme

La frontière entre terme et non-terme, comme nous avons vu, est tantôt perméable et élastique et tantôt rigide et imperméable selon le degré de technicité du corpus. Ce va et vient continuel peut être, théoriquement, réglé : primo par une définition pertinente du terme et secundo par des critères palpables délimitant la frontière minimale entre terme et non-terme.

3.1. Définition du terme

Le terme constitue l'unité opératoire du travail terminologique et le noyau dur d'un corpus spécialisé autour duquel gravitent les notions, la structure et les éléments outils de la terminologie. Cette unité centrale de la terminologie peut être définie selon plusieurs optiques.

La première optique, représentée dans presque la quasi-totalité des ouvrages de terminologie, conçoit le terme en tant que la forme linguistique d'un concept. Autrement dit, le terme est décrit comme étant la dénomination linguistique d'une, et une seule, notion spécialisée. C'est cette dichotomie entre dénomination et notion qui préside toute définition du terme.

La seconde optique de définition du terme est moins notionnelle et plus applicationnelle; étant donné qu'elle est issue de la linguistique de corpus, domaine interdisciplinaire de la terminologie:

> **« Une certaine expérience de la linguistique de corpus nous amène à suggérer que le terme mérite d'être défini comme le mot ou la suite de mots qui pose un problème de recherche d'équivalence, de compréhension ou d'usage phraséologique. Cette définition n'est, certes, pas la plus simple à modéliser en ingénierie linguistique, mais elle est celle qui se rapproche**

> **sans doute le plus de la réalité quotidienne du traducteur. Ce point de vue rejoint l'opinion nouvelle qui voudrait que la terminologie ne soit pas seulement une discipline consacrée à l'élaboration de glossaires. Elle doit aussi fournir des informations sur les collocations dans les textes spécialisés et sur les métamorphoses de termes qui ne sont pas toujours des syntagmes figés. »** (Campenhoudt, 2002 : 25)

Cette définition, libérant le terme de la dépendance de la notion, renforce la dimension linguistique pratique de l'étude du terme par rapport au non-terme. Il importe, subséquemment, de prendre en considération ce rapport entre terme et non-terme dans notre étude terminologique en vue de combler les définitions du terme et de mieux cerner les points de divergence entre ces deux unités contradictoires.

3.2. Plans de la distinction entre terme et non-terme

Se basant sur les diverses définitions du terme ainsi que sur les tentatives de distinction entre terme et mot général préalablement effectuées dans les études terminologiques, nous tenterons dans le présent article d'établir une distinction nette entre terme et non-terme, laquelle sera ultérieurement applicable sur le plan automatique.

En effet, plusieurs études ont tenté de mettre au point les caractéristiques du terme qui le distingue du mot général. Dans la présente section, notre tâche sera donc de récapituler, dans un premier temps, tous les éléments distinctifs précités pour en dresser une liste exhaustive ; et dans un second temps, d'analyser ces éléments pour ajouter notre apport à cette distinction, soit par analyse ou interprétation desdits éléments, soit par adjonction de nouveaux éléments plus pratiques(2).

Notre conception de la distinction entre terme et non-terme suit *« une démarche de dévoilement partant du plus abstrait pour accéder au plus concret »* (El Bacha, 2009 : 53) divisant les plans de distinction en six plans, lesquels sont : le plan pragmatique, le plan sémantique, le plan formel, le plan fonctionnel, le plan statistique et enfin le plan traductionnel(3).

3.2.1. Plan pragmatique

Nous désignons par plan pragmatique le plan de la communication, en d'autres termes, le plan des situations de communication où fonctionne un terme. Tout terme se caractérise par son appartenance à un domaine spécialisé. C'est cette référence spécialisée du terme à un domaine précis de l'activité humaine qui le distingue du non-terme :

> **« D'un point de vue formel ou sémantique, les termes ne manifestent pas une grande différence par rapport aux mots. Les différences deviennent claires si on les examine du point de vue de la pragmatique et de la communication. En effet, la particularité la plus notable de la terminologie, par rapport au lexique commun, est qu'elle sert à désigner les concepts propres des disciplines et des activités de spécialité. »** (Cabré, 1998 : 149)

Contrairement au mot général, le terme n'existe que dans un contexte spécialisé ou un domaine spécialisé. Le mot général quant à lui apparaît soit dans le contexte de la langue commune soit dans le contexte de la langue de spécialité comme c'est déjà décrit dans la section 2. Des mots comme [*jour - siècle - caractère - ouvrir - dormir - ...etc.*] sont en usage courant dans tous types de textes. Par contre, les termes comme [*Loi - législature - saisie-arrêt - criminaliser - assignation - ...etc.*] sont inhérents au domaine juridique, et c'est dans les textes juridiques que nous les recensons (abstraction faite du degré de spécialisation du texte depuis le niveau de la haute technicité jusqu'au niveau de la vulgarisation).

3.2.2. Plan sémantique

Tout comme le mot général, le terme se compose d'un signifiant (dénomination) et d'un signifié (notion). La différence entre terme et non-terme réside donc en la matière dans « *le rapport que le terme établit avec le référent* » (El Bacha, 2009 : 53-54) ; ainsi que dans le point de départ sémantique ou la direction de la démarche terminologique. Pour le non-terme, c'est le signifiant qui décide du ou des signifiés ; tandis que pour le terme, c'est de la notion (signifié) que dépend la dénomination (signifiant). Aussi la démarche du lexicologue diffère-t-elle de celle du terminologue.

Une deuxième distinction est à relever sur le plan sémantique. Il s'agit de la monosémie du terme contre la polysémie du non-terme (mot général) : un terme ne possède qu'un, et un seul, sens dans un domaine de spécialisation donné. Cette univocité représente un caractère saillant de la terminologie et la démarque du lexique général :

> **« On trouvera une troisième caractéristique du terme dans le fait que, pour une notion donnée, il y a, théoriquement, une dénomination et une seule. Cette caractéristique du terme se fonde sur un autre important postulat de la terminologie, celui du rapport d'univocité entre dénomination (signifiant) et notion (signifié), rapport du type réflexif. »** (Rondeau, 1984 : 19)

Il est à noter que cette univocité terminologique est à discuter quand le terme est transfrontalier entre deux ou plusieurs domaines. C'est ainsi que le terme (dénomination) possédera plusieurs sens (notions). À titre d'exemple, le terme [*Droit*] signifie, dans le domaine du droit civil : « [l']ensemble des règles en vigueur dans une société donnée, qui sont sanctionnées, au besoin, par une autorité juridictionnelle » (le grand dictionnaire terminologique GDT) ; tandis qu'il acquiert un sens différent dans le domaine de l'art de gestion théâtrale : « Sorte de taxe perçue par les sociétés compétentes à l'occasion de l'exécution publique d'une œuvre dramatique protégée » (GDT).

3.2.3. Plan formel

Nous entendons par plan formel les modalités de formation des unités de la langue. Il est communément admis que les modalités de formation du mot général, le non-terme en l'occurrence, s'appliquent au terme. Toutefois, la formation des termes est régie par certaines particularités de prédilection, à savoir :

- La modification sémantique (soit par élargissement, restriction ou modification du sens) pour transformer un mot général en un terme. Par exemple, le mot [bureau] devient un terme de l'informatique désignant, selon la définition du grand dictionnaire terminologique, le « *contenu de l'écran qui apparaît quand on ouvre un micro-ordinateur muni d'une interface utilisateur graphique, et sur lequel s'affichent les fenêtres et les icônes* ».

- Les modes de la création néologique[4] du mot s'appliquent toujours aux termes, à savoir : la dérivation, la composition et la troncation. Or, certains modes sont privilégiés (parlant de la langue française) comme surtout l'affixation, recourant à certains préfixes ou suffixes vu leur *« densité sémantique spécialisée »* (El Bacha, 2009 : 54) pour des domaines spécialisés comme la chimie et la physique. La siglaison est également de mise, vu la tendance terminologique à circuler plus d'informations via des formes brèves et condensées, à tel point que nous parlons de nos jours de dictionnaires de sigles.
- L'emprunt est également un mode récurrent dans la formation des termes, malgré les critiques, et s'emploie comme solution pour pallier au retard des langues par rapport à la langue anglaise. C'est pourquoi, on le rencontre rarement en anglais, un peu plus en français et beaucoup plus en arabe[5].
- La parenté morphologique impliquant les différents dérivés d'une même unité terminologique, comme l'indique Marie-Claude L'Homme : « *La parenté morphologique - nécessairement accompagnée d'une parenté sémantique - est un autre indice permettant de confirmer un sens spécialisé. Si des termes ont été retenus en vertu des critères [précédents], leurs dérivés sont forcément spécialisés.* » (L'Homme, 2004 : 65)
- La formation syntagmatique constitue également un mode de formation terminologique privilégié, étant donné que la quasi-totalité des études terminologiques est axée sur le niveau syntagmatique des termes et néglige le niveau des termes simples.

3.2.4. Plan fonctionnel

Sur le plan fonctionnel, terme et non-terme se démarquent par la restriction qu'entraîne la terminologie vis-à-vis de certaines parties du discours (PDD). Il est communément admis en terminologie que le terme est représenté toujours par une entité nominale vu la capacité des noms à mieux véhiculer les notions par rapport aux verbes et aux adjectifs qui ne sont

pas représentatifs du point de vue terminologique sauf s'ils sont dérivés d'une structure nominale d'un terme. Les autres PDD comme l'adverbe, les pronoms, les noms propres, etc. sont exclus du cadre fonctionnel des termes, contrairement aux mots généraux qui sont des non-termes :

> **« 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 les noms de processus, d'actions. Les systèmes terminologiques excluent tout signe linguistique dont la fonction de dénotation classificatrice ou de symbole conceptuel est nulle ou dérivée: ainsi des marques de l'énonciation (pronoms personnels, qui, comme les noms propres linguistiques, dénotent un objet unique, mais ne le classent pas selon des critères notionnels ; adjectifs possessifs ; adverbes de temps et de lieu), des unités à fonction relationnelle (mots dits ‹ grammaticaux ›), des unités qu'on peut considérer comme des transformés sémantiques (adverbes qu'on peut ramener à un adjectif, verbes nominalisables…). Restent donc, comme on le constate aisément en consultant un vocabulaire terminologique, des noms communs, des syntagmes nominaux, quelques verbes dont le contenu notionnel ne peut se ramener à celui d'un nom, et des adjectifs se trouvant dans une situation analogue (qui seront des noms dénotant une qualité). »** (Rey, 1979 : 24-25)

Nous sommes parvenu à prouver ces acquis terminologiques qui distinguent entre terme et non-terme dans notre thèse de magistère (El Bacha, 2009). Sur 990 termes équivalents français arabes, nous avons recensé un taux élevé de substantifs dans les deux langues françaises et arabes soit 70% et 67% respectivement. Au second degré, nous trouvons les formes adjectivales qui comptent 23.6% pour le français et 18.3% pour l'arabe. Le taux des verbes était de 5.6% et 2.5% pour le français et l'arabe respectivement.

L'adverbe figure seulement dans 1.2% des termes français et 0.6% des termes arabes et il était recensé vu sa forme dérivée d'une forme nominale terminologique. Toutes autres PDD sont exclues de la liste des termes alignés.

3.2.5. Plan statistique

Nous avons déjà vu dans la section 2 les frontières qui séparent LSP et LC et, subséquemment, distinguent entre terme et non-terme. L'une de ces frontières est le degré de spécialité du corpus qui plus il augmente plus le nombre de termes augmente. Ceci dit, la fréquence d'apparition octroie à un terme plus de valeur et le distingue du mot général. Nous allons voir dans la section 5.1 comment ce plan est hautement exploité comme indice d'identification automatique des termes.

3.2.6. Plan traductionnel

Finalement, le terme se distingue du non-terme par la difficulté qu'il entraîne pour le traducteur au niveau de la *« transition bilingue du texte vers une […] autre que la langue maternelle du traducteur »* (El Bacha, 2009 : 56). Un terme doit se traduire vers un, et un seul, équivalent dans la langue cible au sein d'un même domaine de spécialité. Par contre, la traduction d'un mot général, un non-terme, reflète une liberté extrême au traducteur à cause de sa polysémie. Ce plan constitue, ipso facto, une distinction marquante et pratique entre terme et non-terme.

En effet, les six plans précités contribuent à distinguer entre terme et non-terme au sein d'un corpus spécialisé. Au niveau du travail manuel, ces critères peuvent servir de guide au terminologue pour extraire les unités opératoires de son étude qui sont les termes. Cependant, au niveau automatique, ces plans doivent acquièrent une interprétation plus applicationnelle qui doit être appréhendée par la machine au service du repérage automatique des termes.

4. Le repérage automatique des termes

Le processus de repérage des termes chevauche entre travail manuel et travail

automatique en terminologie. Il constitue une opération terminologique fondamentale qui consiste à identifier les termes appartenant au champ de l'étude, c'est-à-dire au domaine de spécialisation prédéterminé par le terminologue. Une telle identification nécessite, ipso facto, une distinction entre terme et non-terme. Cette tâche terminologique primordiale peut s'effectuer par le terminologue, comme l'indique Robert Dubuc :

> **« Le repérage est une opération qui permet au terminologue, en face d'un texte, d'identifier les expressions appartenant au thème de sa recherche. Il suppose de la part de celui qui le fait : 1° une excellente connaissance de la langue courante, lui permettant de rejeter toutes les expressions qui n'ont pas réellement un caractère technique : 2° la possession des rudiments de la technique étudiée pour identifier les notions qui appartiennent au thème de sa recherche et non pas à des spécialités connexes. »** (Dubuc, 1980 : 25)

Cependant, le travail du repérage des termes, effectué manuellement, perd son efficacité du point de vue qualitatif, du fait que le terminologue humain est incapable à poursuivre les conditions requises dans la définition ci-dessus tout au long de son travail ; ainsi que du point de vue quantitatif, puisque le travail manuel reste lacunaire face aux corpus spécialisés volumineux. Ce sont exactement ces causes qui impliquent une automatisation du repérage des termes:

> **« L'étape principale du dépouillement est le repérage de termes qui implique, lors de la lecture systématique du corpus, une identification manuelle des termes. Ce travail, plut6t fastidieux et mécanique, de par son aspect répétitif, est un candidat idéal pour l'automatisation. En effet, l'ordinateur est infiniment plus systématique que l'humain pour accomplir des tâches répétitives. Lance sur un corpus de grande taille, un ordinateur peut l'analyser selon un ensemble de règles qui seront respectées du début à la fin de la procédure ; un humain peut difficilement accomplir la même tâche avec autant de brio durant plusieurs jours, vingt-quatre heures sur vingt-quatre. »** (Drouin, 1997 : 45)

De manuel le repérage des termes devient automatique. Tâche facilitée ou compliquée ? Certainement, le travail automatique facilite beaucoup la tâche du repérage comme signalé ci-dessus par Patrick Drouin. Mais, les plans de distinction entre terme et non-terme qui sont accessibles au terminologue dans le repérage manuel des termes le seront-ils pour l'ordinateur dans le repérage automatique ? Cette interrogation sera examinée dans la section suivante.

5. Plans de distinction et indices d'identification automatique des termes

En effet, le processus d'identification automatique des termes revêt deux phases complémentaires dans le travail terminologique, telles sont : l'acquisition automatique de termes et le dépouillement de listes des candidats termes (CT) issues de l'acquisition automatique.

5.1. Acquisition automatique des termes

Parler de l'acquisition automatique des termes, c'est parler du *« processus d'automatisation du repérage des termes par un terminologue »* (El Bacha, 2009 : 179). Guidé par son intuition linguistique et son expérience terminologique, le terminologue accomplit le repérage manuel des termes. Toutefois, l'intuition du terminologue s'estompe au plan automatique et se trouve remplacée par la rigueur de la machine. Le rôle du terminologue sera donc de traduire les plans de distinction entre terme et non-terme vers des règles claires ; et de les formuler dans des algorithmes bien définis, soit qu'ils sont conçus par le terminologue informaticien via le recours à un langage de programmation, soit qu'ils sont implantés dans un extracteur automatique conçu, tout de même, par un langage de programmation mais sans implication du terminologue qui se transforme en utilisateur du logiciel. Ces algorithmes sont donc la traduction pratique automatique des plans de distinction théoriques.

Selon la classification de Marie-Claude L'Homme (2004), les méthodes d'acquisition automatique de termes sont réparties sur trois catégories : la méthode linguistique, la méthode statistique et la méthode mixte (ou hybride selon la désignation de Patrick Drouin).

Chacune de ces méthodes exploite les principes prédéterminés de distinction entre terme et non-terme, comme nous le verrons dans les points suivants.

5.1.1. Méthode linguistique

Le principe de cette méthode repose essentiellement sur les trois plans sémantique, formel et fonctionnel de distinction entre terme et non-terme. Selon la nature des connaissances requises, cette méthode opère suivant deux types d'analyse : la première consiste en une analyse de surface de la phrase dans le cas des connaissances lexicales et morphologiques ; la seconde consiste en une analyse profonde de la phrase pour extraire des connaissances syntaxiques :

> **« Les systèmes présentés ici sont qualifiés de linguistiques puisqu'ils font appel à des techniques d'analyse reposant sur les connaissances actuelles de la langue et de sa structure. On distingue principalement les systèmes utilisant des informations syntaxiques et ceux qui utilisent des informations lexicales ou morphologiques. Les premiers reposent sur une analyse complète de la phrase en ses constituants afin d'en dégager les syntagmes intéressants selon les objectifs de la recherche. Dans le second cas, des grammaires locales procèdent à une analyse de surface de la phrase à la recherche de syntagmes potentiels. Ces derniers sont décrits, à leur tour, à l'aide de grammaires qui permettent de circonscrire l'ensemble des réalisations potentielles. »** (Drouin, 2002 : 66-67)

Il faut noter que certains logiciels comme Adepte-Nomino(6) réunit les trois types de connaissance lexicales, morphologiques et syntaxiques. D'autres logiciels adoptant la méthode linguistique, comme Lexter(7), fonctionne suivant le système de « frontières de termes » qui repose sur l'analyse syntaxique de la phrase par son découpage en mots assortis d'étiquettes morphologiques où sont désignées leurs catégories grammaticales et leurs lemmes. L'étape suivante de l'extraction automatique dans Lexter repose sur l'identification des frontières des termes complexes, c'est-à-dire les unités qui ne peuvent faire partie

intégrante d'un terme. Les verbes conjugués, les adverbes, les pronoms, les conjonctions de coordination, les conjonctions de subordination et enfin les signes de ponctuation (à l'exception du tiret) sont considérés parmi les frontières de termes. C'est ainsi que les plans formel et fonctionnel de la distinction entre terme et non-terme sont exploités et automatisés.

5.1.2. Méthode statistique

Cette méthode repose sur les calculs statistiques sans aucun recours aux connaissances linguistiques, comme le dicte le principe du plan statistique qui exploite la fréquence d'un terme comme indice de repérage automatique. Sur le plan théorique, la fréquence est conçue comme un moyen de comptage. Sur le plan automatique, que veut-on dire par fréquence ?

Par fréquence, nous désignons le nombre d'occurrences d'un terme dans un corpus. Deux types de fréquence sont à étudier : la fréquence absolue ou brute, c'est-à-dire le nombre total de fois où un terme apparaît dans un corpus et la fréquence relative qui est le pourcentage d'occurrences d'un terme par rapport au nombre total des mots dans un corpus. L'indice de la fréquence est hautement exploitable dans les études terminologiques du fait de son efficacité. S'ajoute à l'indice de la fréquence la répartition, c'est-à-dire le nombre de textes dans lesquels est recensée la fréquence d'un terme dans corpus. À noter que le calcul de la fréquence et de la répartition se fait automatiquement via un algorithme compteur.

5.1.3. Méthode mixte

La méthode mixte, comme son nom l'indique, est un amalgame entre les deux méthodes précédentes. Sa raison d'être est de combler les lacunes de chacune des deux méthodes linguistique et statistique et d'en maximiser les avantages. Les travaux appliquant cette méthode sont presque identiques, leur différence réside dans l'ordre qu'ils adoptent dans le cheminement du travail terminologique : certains commencent par la méthode statistique, d'autres par la méthode linguistique.

Jusqu'à ce stade, se termine la phase de l'acquisition automatique de termes. Une deuxième phase est à étudier et avec laquelle l'exploitation des plans de distinction entre terme et non-terme se trouve maximisée.

5.2. Dépouillement des listes des candidats-termes (CT)

Le dépouillement terminologique est un processus *« tributaire »* de l'acquisition automatique de termes. Il désigne la *« distinction entre les unités qui ont un statut terminologique saillant et les autres unités appartenant au vocabulaire de la langue générale »* (El Bacha, 2009 : 205). Nous revenons ainsi aux plans de distinction entre terme et non-terme, mais cette fois-ci dans une liste de candidats-termes (CT)[(8)]. À la lumière des six plans de distinction entre terme et non-terme, cinq indices sont mis en œuvre pour l'identification du statut terminologique des CT dans la liste issue de l'acquisition automatique, à savoir : la prédominance des termes de nature nominale, la problématique de traduction imposée par un CT, l'appartenance du CT au domaine de spécialité prédéterminé, la parenté morphologique entre CT et finalement, la complexité et l'étrangeté morphologique du CT.

5.2.1. Prédominance des termes de nature nominale

Nous avons expliqué, dans le plan fonctionnel, un des acquis terminologiques qui prévalent certaines PDD aux dépens des autres. Ce principe est préservé dans l'identification automatique des CT. Les CT assortis d'étiquettes morphosyntaxiques dénotant leur nature nominale sont recensés par les extracteurs. Or, il faut se méfier de cet indice, appliqué à lui seul, car tout substantif n'est pas nécessairement un terme et une intervention humaine est à recommander en l'occurrence pour combler les lacunes de cet indice.

5.2.2. Problématique de traduction des termes

Une unité polémique pour la traduction spécialisée revêt, indubitablement, un statut terminologique dans une liste des CT. Appliqué aux listes des CT, cet indice est prouvé dans notre thèse de magistère où un nombre important des CT pose problème dans la traduction, ce qui décide de son statut terminologique. À titre d'exemple, les CT français [*coercition* - *détectabilité* - *plénipotentiaires*] représentent une problématique de traduction pour le terminologue dans les listes CT, ce qui est considéré comme indice fort saillant de leur statut terminologique, et ils sont donc retenus comme termes. Après alignement des listes CT, leurs

équivalents arabes ont renforcé leur statut terminologique : [(coercition = إكراه) - (détectabilité = كشف عن الألغام) - (plénipotentiaires = (مفوضين].

Il faut, cependant, prendre en considération que certains CT, qui sont accessibles dans la traduction comme [*droit* - *paix* - *paragraphe* - *loi*], doivent être retenus puisqu'ils sont des termes juridiques, nécessitant ainsi des restrictions dans l'application de cet indice.

5.2.3. Domaine de spécialité prédéterminé

Les termes reconnus dans une liste CT doivent figurer dans le cadre du domaine de spécialité prédéterminé par l'étude. Aussi, pour une étude de la terminologie juridique, ne faut il pas retenir des termes de la science, de la médecine ou de la sexualité, tels : [*biodiversité* - *clonage* - *paludisme* - *pandémies* - *antisexiste* - *pédopornographie* - *proxénétisme*]. D'où l'importance du plan pragmatique et dans la distinction entre terme et non-terme et dans la délimitation entre termes appartenant à des domaines différents.

5.2.4. Parenté morphologique entre les CT

La parenté morphologique est détectée automatiquement se basant sur la parenté formelle entre les CT. Par exemple, les CT [*légalité* - *législation* - *témoin*] sont retenus comme termes ; il sera donc facile de détecter par la machine leurs formes connexes dérivées :

Ex.(1) Légalité → *légal - légalement*

Ex.(2) Législation → *législative - législature*

Ex.(3) Témoin → *témoignage - témoigner*

La parenté morphologique, comme l'affirme Marie-Claude L'Homme, est un indice fiable et récurrent.

5.2.5. Complexité et étrangeté morphologiques

Nous sommes parvenu à concevoir cet indice de la complexité et l'étrangeté morphologique se basant sur le dépouillement de la liste des CT dans notre thèse de magistère.

Par complexité et étrangeté morphologiques, nous entendons : *« [soit] la combinaison inédite de plusieurs morphèmes libres dans une structure morphologique composée, … [soit] la soudure inhabituelle de deux termes dans un seul terme composé »* (El Bacha, 2009 : 210-211). Le premier cas se manifeste dans les exemples suivants : [*contre-mesures / sous-comité / sous-programme / …*] ; le second cas est relevé dans des termes tels : [*interétatiques / intergouvernementale / interinstitutions / …*]. Dans les exemples précédents, le morphème [*inter-*] ne désigne uniquement en l'occurrence une relation interne [*entre*] ; mais signifie également [entre nations], c'est-à-dire international, ce qui apparaît clairement dans les équivalents arabes de ces CT français : [interétatiques =فيما بين الدول / intergouvernementale= حكومي دولي / interinstitutions = فيما بين الوكالات / …]. La quasi-totalité des CT qui revêtent une formation inédite ou inhabituelle sont retenus comme termes.

Il s'avère nettement de par les cinq indices du dépouillement des listes CT, décrits ci-dessus, l'apport des plans de distinction entre terme et non-terme pour le repérage automatique des termes dans un corpus spécialisé. Or, il faut garder présent à l'esprit que la tâche de la distinction entre terme et non-terme demeure une problématique qui nécessite l'intervention constante du terminologue humain pour dûment assurer le filtrage et la validation des résultats, comme nous avons déjà vu avec l'application de certains de ces indices :

> **« Il est […] impossible, pour le moment, de penser à distinguer automatiquement les termes des non-termes sans une intervention humaine non négligeable »** (Drouin, 2002 : 37)

6. Conclusion

Au terme de notre étude, nous sommes parvenu à réaliser deux objectifs complémentaires. Le premier objectif consiste à fonder les principes théoriques de la distinction entre terme et non-terme de par six plans : le plan pragmatique, le plan sémantique, le plan formel, le plan fonctionnel, le plan statistique et le plan traductionnel. Ces six plans distinctifs entre l'unité terminologique et l'unité lexicale, bien que théoriques,

permettent de cerner les points de divergence entre le terme et le mot général, considéré dans l'optique terminologique comme un non-terme.

Le second objectif est atteint lors de l'explication des phases de l'identification automatique des termes dans le présent article, lesquelles sont : l'acquisition automatique de termes et le dépouillement automatique des listes CT. Ces deux phases sont basées sur l'automatisation desdits plans de distinction entre terme et non-terme. Ceci dit, l'apport des plans de distinction entre terme et non-terme n'est pas à nier étant transformés en des indices d'identification automatique de termes.

Bibliographie

1- CABRÉ (M. T.), 1998 : *« La Terminologie. Théorie, méthode et applications »*, Canada, Québec, les presses de l'université d'Ottawa, Armand Colin, 322 pages.

2- DROUIN (P.), 2002 : *« Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés »*, thèse en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.) en linguistique, Université de Montréal – Faculté des études supérieures, 307 pages, disponible sur : <http://www.olst.umontreal.ca/pdf/DrouinPhD2002.pdf>, page consultée le 20-3-2008 à 11:54 pm.

3- Drouin (P.), 1997 : *« Une méthodologie d'identification automatique des syntagmes terminologiques : l'apport de la description du non-terme »*, Meta, vol. 42, n° 1, pp. 45-54, disponible sur :< http://www.erudit.org/revue/meta/1997/v42/n1/002593ar.pdf>, page consultée le 12-12-2011 à 05:27 pm.

4- DUBUC (R.), 1980 : *« Manuel Pratique de Terminologie »*, Paris, publié en coédition par linguatech et le conseil international de la langue française, 98 pages.

5- DUBUC (R.), 2002 : *« Manuel Pratique de Terminologie »*, 4ème édition, Paris, publié en coédition par linguatech et le conseil international de la langue française, 98 pages.

6- El Bacha (M.), 2009 : « *Modalités de formation des termes français et leurs équivalents arabes dans les textes des résolutions de l'Assemblée Générale des Nations Unies – Analyse linguistique informatique* », thèse de magistère, Faculté d'Al-Alsun, Université de Ain-Chams, Le Caire, Egypte.

7- L'HOMME (M. C.), 2004 : *« La terminologie: principes et techniques »*, Canada, Québec, les Presses de l'université de Montréal, 278 pages.

8- REY (A.), 1979 : *« La terminologie. Noms et notions »*, Paris, Presses universitaires de France, collection Que sais-je ?, 127 pages.

9- RONDEAU (G.), 1984: *« Introduction à la terminologie »*, 2ème édition, Québec, gaëten morin éditeur, 238 pages.

10- Le Grand dictionnaire terminologique (GDT): http://www.oqlf.gouv.qc.ca/ ressources/gdt.html.

* * * *

Notes et remarques:

1- Selon Guy Rondeau (1984 : 23-24), la LSP et la LC recouvrent un *« sous-ensemble »* de la langue.

2- Ce travail est le fruit d'une étude de terminologie computationnelle issue de notre thèse de magistère intitulée « *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* ».

3- Dans notre thèse de magistère, le plan traductionnel était intitulé le plan de transition bilingue. Mais, nous avons jugé plus adéquat de modifier cette appellation, étant donné qu'elle sied mieux à notre visée pratique axée sur l'apport de la distinction entre terme et non-terme pour le repérage automatique de termes.

4- Pour la terminologie, nous appelons les modes de création de nouvelles unités néonymie ; quant à la néologie, elle désigne uniquement les modes de formation des mots de la langue générale.

5- Selon les résultats de l'équivalence terminologique issus de notre thèse de magistère, le taux d'emprunt était presque identique dans les deux langues française et arabe : d'un nombre de 990 termes simples, 1.3% termes français et 2.6% termes arabes sont empruntés.

6- Ce logiciel est développé par Pierre Plante, Lucie Dumas et André Plante (ATO, Département de linguistique, Université du Québec à Montréal). Il fonctionne dans l'environnement Macintosh et dans l'environnement Windows (Le Réseau international de néologie et de terminologie (Rint) a participé à l'adaptation au logiciel dans ce dernier environnement). Pour

une description détaillée du logiciel, consultez la page sur le lien suivant : www.ling.uqam.ca/nomino/synopsis.htm.

7- Lexter est un logiciel conçu par Didier Bourigault en 1994 dans le cadre de sa thèse intitulée « *Un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir des texte* », soutenue à L'École des Hautes Études en Sciences Sociales à Paris.

8- Campenhoudt définit le candidat-terme (CT) ainsi : *« On nomme candidat-terme toute suite de caractères identifiée comme susceptible de constituer un terme spécialisé. »* (Campenhoudt, 2002 : 18).

*** * * ***