Extraction terminologique : vers la minimisation de ressources

Yuliya Korenchuk[1, 2]
(1) LiLPa (Linguistique, Langues, Parole), EA 1339, Université de Strasbourg
(2) Rebuz SAS, Strasbourg
korenchuk@unistra.fr

Résumé. Cet article présente une méthode ayant pour objectif de minimiser l'apport extérieur nécessaire à la tâche d'extraction terminologique (ET) et de rendre cette tâche moins dépendante de la langue. Pour cela, la méthode prévoit des ressources morphologiques et morphosyntaxiques simplifiées construites directement à partir d'un corpus lemmatisé. Ces ressources endogènes servent à la création d'un système de filtres qui affinent les calculs statistiques et à la génération de patrons pour l'identification de candidats termes polylexicaux. La méthode a été testée sur deux corpus comparables en chimie et en télécommunication, en français et en anglais. La précision observée sur les 100 premiers candidats termes monolexicaux fluctue entre 71% et 87% pour le français et entre 44 % et 69 % en anglais ; celle des candidats termes polylexicaux s'élève à 69-78 % en français et 69-85 % en anglais en fonction du domaine.

Abstract. The article presents the method which aims to minimize the use of external resources for the terminology extraction task and to make this task less langage dependent. For that purpose, the method builds simplified morphological and morphosyntactic resources directly from a lemmatized corpus. These endogenous resources are used both in filters, which refine the statistical calculations, and in patterns for polylexical terms identification. The method was tested on two comparable corpora in chemistry and in telecommunication in French and in English. The precision observed on the first 100 monolexical terms fluctuates between 71% and 87% for French and between 44% and 69% in English ; for polylexical terms the precision was 69-78% in French and 69-85% in English depending on the domain.

Mots-clés : extraction terminologique, ressources endogènes, apprentissage automatique.

Keywords: terminology extraction, endogenous resources, machine learning.

1 Introduction

L'extraction terminologique (ET) a de nombreuses applications dans la construction de ressources linguistiques et sémantiques (par exemple, des glossaires et des bases terminologiques, des ontologies spécialisées, etc.). Ses résultats peuvent servir directement à des utilisateurs humains (des traducteurs ou des terminologues) ou bien être exploités par des moteurs de recherche ou des systèmes de traduction automatique. Les méthodes et les outils élaborés pour l'ET sont très variés. Leur diversité se manifeste à plusieurs niveaux, tels que l'approche plus ou moins dépendante de la langue, les exigences par rapport au choix du corpus (sa taille, son type, un prétraitement spécifique, etc.) et d'autres ressources nécessaires. Du point de vue de l'application industrielle, chacun de ces aspects a un coût, et les coûts de préparation et d'adaptation des ressources peuvent s'avérer assez importants.

Une des principales motivations de la méthode proposée est d'identifier automatiquement des régularités dans les textes analysés pour créer des ressources endogènes à partir du corpus. Cette approche est :
– applicable à plusieurs langues ;
– indépendante du domaine de spécialité ;
– capable d'extraire des candidats termes mono et polylexicaux ;
– capable d'apprendre les informations nécessaires pour l'analyse à partir du corpus.
Cet objectif a été atteint par le biais de patrons morphologiques et morphosyntaxiques endogènes qui résultent d'une approche multi-étapes à base de n-grammes et qui sont générés directement au cours de l'analyse à partir du corpus.

Les résultats obtenus ont confirmé la pertinence de la méthode pour sa tâche principale. En outre, ses caractéristiques permettent d'envisager l'application de la méthode à l'enrichissement multilingue des ontologies de spécialité.

La première partie de cet article propose une brève description de méthodes existantes et de ressources utilisées en ET. La

deuxième partie décrit les corpus de test et la méthode proposée. Ensuite viennent l'évaluation des résultats pour les deux langues et la conclusion.

2 État de l'art

L'ET connaît son essor dans les années 1990 et de nombreux systèmes et méthodes apparaissent presque simultanément à cette période. Depuis, le domaine est en plein développement. Les systèmes d'ET cherchent à extraire des unités lexicales simples ou complexes qui sont susceptibles d'être des termes. Plusieurs traits peuvent être pris en compte (la fréquence, la structure, le contexte ou la comparaison du corpus analysé avec un corpus de référence).

La notion du candidat terme est utilisée en extraction terminologique pour désigner les unités lexicales repérées par les systèmes automatiques (Drouin et Langlais, 2006). Cette notion sécurise le travail avec les systèmes d'ET, car le fait d'insister sur le caractère incomplet de résultats du traitement incite à valider ces résultats avant de les réutiliser.

La palette des outils disponibles pour le français comprend des systèmes récents comme TTC TermSuite (Morin et Daille, 2012) et YaTeA (Aubin et Hamon, 2006), ainsi que leurs prédécesseurs : ACABIT (Daille, 1996, 2003; Morin et Daille, 2006), FASTR (Jacquemin, 1997) et LEXTER (Bourigault *et al.*, 1996). Ces outils sont basés sur des méthodes et des ressources différentes et nous avons choisi d'en présenter les plus pertinentes par rapport à notre projet.

2.1 Méthodes d'extraction terminologique

En général, les classifications des méthodes d'ET (Bernhard, 2006) distinguent les méthodes basées sur des mesures statistiques, les méthodes basées sur des éléments linguistiques et les méthodes mixtes.

Les mesures statistiques, telles que la fréquence absolue ou pondérée, TFxIDF, l'information mutuelle ou l'indice de Jaccard mettent en évidence des unités mono ou polylexicales caractéristiques d'un document ou d'un corpus. Ces méthodes sont indépendantes de la langue et nécessitent uniquement un ou plusieurs corpus. Toutefois, leurs résultats sont meilleurs sur des corpus de taille importante. Certaines méthodes permettent d'évaluer le potentiel terminologique de candidats termes (Drouin et Langlais, 2006), ce qui peut servir de filtre pour d'autres méthodes.

La deuxième catégorie de méthodes fait appel à des éléments linguistiques parmi lesquels on peut distinguer deux groupes : les patrons morphosyntaxiques et les formants de langues classiques. L'identification des termes de certains domaines peut utiliser des traits morphologiques spécifiques à leurs nomenclatures. Tel est le cas de la nomenclature en chimie, en physique, en biologie ou encore en médecine. Cette méthode est adaptée pour des termes qui contiennent des composants savants, c'est-à-dire, des morphèmes provenant du grec ou du latin. En effet, ces morphèmes sont très productifs ; par exemple, le radical *AZOT* est à base des termes suivants : *azote, azotate, azoté, azoteux, azotique, azotite, azotémie, azothydrique, azoture, azoturie, etc.* Appliquée aux domaines listés ci-dessus, cette méthode est efficace à 87,70 % selon Estopà *et al.* (2000).

Le deuxième groupe s'approche des méthodes mixtes, car les méthodes à base de patrons morphosyntaxiques ont recours à des calculs statistiques pour affiner leurs résultats. Cette approche est utilisée dans les méthodes de (Daille, 2003; Morin et Daille, 2012), de Bourigault (Bourigault et Fabre, 2000; Bourigault, 2002) et la méthode C-value/NC-value de Frantzi *et al.* (2000). Orobinska *et al.* (2013) propose une approche plus souple en apprenant les patrons morphosyntaxiques caractéristiques à partir du corpus étiqueté. L'un des avantages des patrons morphosyntaxiques est la possibilité d'identifier la variation au sein des candidats termes polylexicaux et de définir des règles de transformation afin d'améliorer l'organisation des résultats (Bourigault et Jacquemin, 1999).

Dans notre travail, nous reprenons certains éléments des deux méthodes mixtes qui ne nécessitent pas de prétraitement de corpus. La première méthode proposée par Vergne (2003, 2004, 2005) permet d'annoter le corpus par des mots informatifs et vides pour ensuite en extraire des termes à structure contrôlée. Nous allons détailler cette méthode dans la partie 3.2. La deuxième méthode, qui s'appelle ANA (Apprentissage naturel Automatique), est développée par Enguehard (1993). A l'étape de la familiarisation, le logiciel parcourt le corpus et en extrait les listes des mots fonctionnels (*a, alors, après, etc.*), des mots fortement liés (*de la, de l', etc.*) et des mots de schéma (*en, de, du, d, de la, des*). Ces deux derniers groupes servent de liaison dans les candidats termes polylexicaux et permettent de les identifier dans le texte. Ensuite, le programme analyse le corpus de manière itérative, en identifiant les mots qui apparaissent fréquemment avec les termes de bootstrap (un ensemble de quelques termes du domaine prédéfinis manuellement dont il est question dans le corpus de

textes), et en les rajoutant dans le bootstrap.

Une autre approche intéressante consiste à utiliser le lexique transdisciplinaire comme indicateur et délimiteur des unités terminologiques (Jacquey *et al.*, 2013; Tutin, 2007). En effet, des mots comme *concept, méthode, technologie, analyser, etc.* servent souvent à introduire des termes d'un domaine sans appartenir à ce domaine. La détection de ces mots ou expressions et leur classification facilitent l'ET. Jacquey *et al.* (2013) combinent le lexique transdisciplinaire projeté sur le corpus avec l'analyse syntaxique pour identifier les cas où une unité du lexique introduit un candidat terme. Leur hypothèse a été confirmée dans environs 74 % des cas (Jacquey *et al.*, 2013).

La plupart des méthodes combinent plusieurs paramètres pour augmenter leur efficacité. Cela s'explique en partie par les particularités des termes (Estopà *et al.*, 2000) ou par les limites internes de chaque approche. Nous allons maintenant décrire les avantages et les limites des différents types de ressources employées par les méthodes d'ET.

2.2 Ressources pour l'extraction terminologique

Un corpus de domaine est la première ressource indispensable à l'ET. Il conditionne le choix de la méthode, car certaines méthodes se montrent plus efficaces sur des corpus de taille importante. Or, il n'est pas toujours facile de trouver un corpus suffisamment grand pour un domaine donné, surtout pour des langues possédant moins de ressources que le français ou l'anglais. Pour les corpus parallèles, le choix est davantage limité. Il est donc intéressant de pallier la contrainte de la taille du corpus soit par la minimisation de l'importance des opérations statistiques (en démultipliant les traits linguistiques pris en compte), soit par l'emploi de corpus de référence pour augmenter le contraste entre les termes et les mots de la langue générale.

Les ressources morphosyntaxiques, notamment les patrons d'identification de candidats polylexicaux, sont développées pour chaque méthode à part. Elles dépendent de l'étiqueteur mophosyntaxique disponible et reposent sur les résultats de ce dernier. Cependant, l'étiqueteur risque de perdre en précision lorsqu'il s'agit d'un domaine très spécifique, comme la médecine ou la chimie. Les méthodes basées sur ces patrons sont performantes, mais limitées à des langues qui ont déjà un étiqueteur disponible.

En ce qui concerne les ressources morphologiques, des listes de formants grecs et latins sont disponibles sur Internet. Cependant, elles nécessitent une vérification manuelle et une adaptation de format avant d'être introduites dans le programme. Les méthodes qui comparent des fréquences de formants dans un corpus de spécialité et dans un corpus de langue générale sont applicables pour affiner les listes existantes (Bernhard, 2006). La limite évidente des formants est leurs productivité dans le domaine analysé : certains domaines ont très peu de recours à ces morphèmes, en préférant des emprunts ou la néologie.

Enfin, des ressources lexicales, comme des anti-dictionnaires (*stop words lists*) ou le lexique transdisciplinaire (Tutin, 2007; Jacquey *et al.*, 2013), sont évidemment liées à la langue pour laquelle elles sont développées. De ce point de vue, les travaux de Drouin (2007), qui portent sur l'identification automatique du lexique transdisciplinaire, sont très intéressants. Une autre méthode permettant d'éviter l'utilisation d'un anti-dictionnaire est l'annotation de Vergne (2003, 2004, 2005). En effet, les mots étiquetés comme vides sont dans la majorité des cas éligibles pour un anti-dictionnaire.

Lever les contraintes existantes dans l'ET signifie minimiser les ressources nécessaires pour cette tâche tout en améliorant les résultats. Nous allons présenter une méthode qui tend à réunir les aspects forts des approches décrites ci-dessus tout en minimisant les ressources nécessaires pour arriver à des résultats satisfaisants.

3 Méthodologie

Les méthodes qui se montrent efficaces nécessitent des ressources externes assez importantes, comme par exemple les listes de formants savants, les patrons morphosyntaxiques, les dictionnaires de référence, etc. Les systèmes basés sur de telles méthodes sont dépendants vis-à-vis de la langue et de ces ressources. Or, ces ressources ne font que refléter des régularités linguistiques. De ce point de vue, il doit être possible de générer ces ressources à partir du corpus-même.

Ainsi, l'objectif principal de la présente méthode est de contourner la nécessité de fournir des ressources morphologiques ou morphosyntaxiques pour obtenir des candidats termes mono et polylexicaux en se basant sur les données les plus fiables obtenues à chaque étape.

3.1 Présentation des corpus

Pour notre projet, nous avons choisi deux corpus comparables bilingues (FR/EN) pour deux domaines bien distincts : la chimie et les télécommunications. Le corpus en chimie est composé de mémoires et de thèses en chimie des métaux sélectionnés manuellement à l'aide des options de recherche avancées sur Google Scholar. Tous les textes sont convertis en UTF-8 et les fragments en langues étrangères ont été éliminés.

Le corpus en télécommunication, plus précisément en technologies mobiles, vient du projet TTC [1]. Nous n'avons donc pas eu à le construire, mais nous avons pu constater qu'il contient essentiellement des documents techniques.

La différence fondamentale entre les deux domaines permet d'apporter un regard critique sur notre méthode, car les résultats peuvent varier en fonction de la nomenclature du domaine (formants savants, emprunts, etc.). Toutes les caractéristiques fournies dans la table 1 sont quantifiées après la tokenisation du corpus par le script du projet Europarl [2]. Chaque paire de corpus est assez équilibrée. Nous pouvons donc nous attendre à des résultats fiables et comparables.

Corpus/ Langue	Chimie				Télécommunication			
	Taille	*Tokens*	*Types*	*Lemmes*	*Taille*	*Tokens*	*Types*	*Lemmes*
Français	4,13 Mo	841 843	49 948	28 717	2,94 Mo	526 240	24 419	14 758
Anglais	3,49 Mo	713 369	44 684	27 274	1,95 Mo	349 656	25 425	17 186

TABLE 1 – Caractéristiques des corpus

3.2 Prétraitement et annotation automatique du corpus

Le système fait appel au script de tokenisation du projet Europarl [3] et à l'étiqueteur morphosyntaxique TreeTagger (Schmid, 1994) qui est utilisé comme lemmatiseur. La suite des traitements est effectuée sur le corpus lemmatisé.

Le corpus est traité par l'algorithme d'annotation par des mots informatifs et non-informatifs de Vergne (2003, 2004, 2005). Cet algorithme combine deux propriétés, la fréquence et la longueur de mots, pour détecter les mots vides à partir de corpus brut indépendamment de la langue de ce dernier. Selon la méthode de Vergne, le mot informatif est sélectionné selon trois critères :

1. Longueur importante
2. Fréquence réduite
3. Entourage par des mots plus courts et plus fréquents

Le reste de mots est considéré comme des mots vides. Cependant, l'application de la méthode, telle qu'elle est présentée par Vergne, rapproche les mots informatifs des candidats termes. Or, les exemples de résultats obtenus pour cette méthode sur des corpus constitués d'articles de presse (Vergne, 2005) contiennent peu de candidats termes tels que nous pouvons exploiter dans un système de gestion de la terminologie. Par exemple, dans *une nouvelle résolution de l'ONU*, les mots informatifs sont : *nouvelle, résolution, ONU*. Les mots *ONU* et *résolution* peuvent effectivement être des candidats termes et même former un candidat terme poly lexical *résolution de l'ONU*, mais le mot *nouvelle* ne s'inscrit pas dans cette terminologie. Dans d'autres exemples, les mots informatifs incluent également des mots de la langue générale, comme *cherche, utiliser, tonnes, L'or*, etc.

Nous avons remarqué que cette annotation n'est pas homogène [4] : un même mot peut être annoté comme informatif s'il est entouré par des mots plus courts ou comme non informatif s'il est à côté d'un mot plus long et encore moins fréquent. Pour cette raison, nous avons appliqué un coefficient pour corriger l'annotation. Le mot est validé comme informatif s'il a été annoté comme tel dans 90 % d'occurrences. Cette amélioration de l'algorithme de base a une conséquence positive sur les résultats de tête de la liste de fréquence absolue (table 2), mais en même temps elle neutralise le caractère local du

1. `http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html`, consulté le 3 mars 2014
2. `http://www.statmt.org/europarl/`, consulté le 3 mars 2014
3. Le choix de tokeniseur est susceptible de changer dans les futures versions du programme.
4. Il est difficile d'évaluer la performance de cette annotation, car la frontière entre les mots vides et informatifs n'est pas clairement définie. Les résultats de l'évaluation faite par Vergne (2003) ne sont pas comparables à l'usage que l'on fait de sa méthode. En effet, ils contiennent des candidats pour plusieurs langues mélangées et ne permettent pas de déduire les critères de validation.

calcul qui permettait de tenir compte des homographes (Vergne, 2004). Le corpus annoté par les mots informatifs et les mots vides servira de base pour l'extraction de candidats termes monolexicaux.

Baseline	M(I)>M(n) [5]	coef. 40 %	coef. 60 %	coef. 80 %	coef. 90 %
(	avoir	complexe	complexe	complexe	complexe
être	complexe	figure	figure	j.	énergie
en	figure	il	il	énergie	surface
ce	il	plus	deux	surface	système
dans	plus	deux	nous	système	métal
par	deux	nous	j.	différent	monsieur
au	nous	pouvoir	énergie	métal	état
que	pouvoir	j.	surface	monsieur	solution
avec	j.	on	système	état	atome
complexe	on	énergie	utiliser	solution	acide

TABLE 2 – Influence du coefficient sur le tri par fréquence absolue

3.3 Extraction de termes monolexicaux et apprentissage des ressources morphologiques endogènes

3.3.1 Termes monolexicaux

Les résultats obtenus par différentes mesures statistiques ne seront pas identiques, surtout ceux de tête de la liste triée. Pour cette raison, nous avons basé l'extraction de candidats termes monolexicaux sur le calcul de la fréquence absolue (nombre d'occurrences dans le corpus, formule (1)) et sur une formule qui retrouve des termes fréquents qui apparaissent dans un nombre réduit de documents (formule(2)). Les deux calculs tiennent compte des mots informatifs uniquement.

$$tf_i = \sum_{j=1}^{n} tf_{ij} : dj \in D \qquad (1)$$

où tf_{ij} est la fréquence d'un mot informatif dans d_j,
d_j est un élément de l'ensemble de documents D,
n est le nombre de documents

$$TF\mathrm{x}IDF(m_i) = \frac{tf(m_i)}{tf(m_{max})} \times \log \frac{D}{d} \qquad (2)$$

où m_{max} est le mot informatif le plus fréquent,
d est le nombre de documents contenant m_i

La deuxième formule appartient à la famille de TFxIDF, mais en représente une version adaptée pour être appliquée sur à une autre unité textuelle (le corpus). L'idée d'utiliser le coefficient TFxIDF pour l'extraction des termes n'est pas neuve : Witschel (2005) suggère d'identifier pour chaque document les termes fréquents qui apparaissent dans peu de documents du corpus. Nous avons modifié cette méthode en calculant la fréquence pondérée d'un mot lemmatisé par rapport à la fréquence maximale d'un mot informatif dans le corpus entier. Cette modification s'explique par la volonté d'éviter un bruit possible si un document du corpus ne porte pas vraiment sur le domaine donné, risquant ainsi de polluer les résultats de TFxIDF calculés pour chaque document.

La table 3 illustre la différence d'approche : la fréquence absolue maximale met en valeur les mots les plus fréquents dans les textes du corpus, tandis que TFxIDF identifie les mots fréquents, mais présents dans un nombre restreint de documents. Nous avons constaté que les mots qui sont en tête de la liste TFxIDF se trouvent au milieu de la liste triée par fréquence décroissante.

Paramètre	Top-10
MaxFreq	complexe, énergie, surface, système, métal, monsieur, état, solution, atome, acide
TFxIDF	perr, pile, capteur, accumulateur, peptide, inhibiteur, tension, 2phen, clo4, clhp

TABLE 3 – Top-10 résultats pour la fréquence absolue maximale et TFxIDF

Les deux listes partagent certains candidats termes, mais il est intéressant de traiter chaque liste à part : comme la précision

5. Le mot a été annoté comme informatif M(I) plus souvent que comme un non informatif M(n)

est maximale en tête de la liste triée par ordre décroissant, le premier tiers [6] de chaque liste sera retenu dans la liste des candidats termes monolexicaux qui seront utilisés pour générer des ressources morphologiques endogènes.

3.3.2 Ressources morphologiques endogènes

Notre méthode prévoit la génération de ressources morphologiques qui servent à l'apprentissage des patrons morphosyntaxiques. Ces ressources sont basées sur les n-grammes de caractères extraits à partir de candidats termes monolexicaux. En effet, les n-grammes correspondent à des quasi-morphèmes selon leurs positions, ce qui a notamment été exploité pour la tâche de tokenisation (McNamee *et al.*, 2009; McNamee, 2008).

Dans notre méthode, nous prenons en compte trois positions de n-grammes : au début du mot, au milieu (dans la fenêtre entre le deuxième et l'avant-dernier caractère du mot) et à la fin. Les listes obtenues confirment l'hypothèse que ces n-grammes se rapprochent des morphèmes de la langue. Nous avons testé le système sur les 2-grammes, 3-grammes et 4-grammes (table 4) à partir de mots dont la longueur est supérieure à 4 caractères. Les résultats de cette comparaison permettent de faire quelques observations sur le comportement de n-grammes selon leur position.

Position	Français						Anglais					
	Chimie			Télécommunication			Chimie			Télécommunication		
n	**2**	**3**	**4**	**2**	**3**	**4**	**2**	**3**	**4**	**2**	**3**	**4**
Début	co	con	comp	co	con	inte	co	con	inte	co	con	inte
	in	pro	cons	in	com	cons	in	int	comp	re	int	comp
	pr	com	élec	re	pro	comp	re	com	elec	in	com	cont
	dé	tra	inte	pr	int	cont	di	pro	nano	pr	pro	cons
	di	int	phot	dé	tra	comm	pr	dis	spec	de	mul	tran
Milieu	at	ect	ectr	at	ent	icat	at	ect	ectr	en	ter	tion
	ti	rat	ctro	ti	rat	tion	er	ter	ctro	ti	ica	icat
	ct	ctr	tion	em	ica	isat	ti	ica	izat	at	ent	atio
	em	tro	isat	ra	ter	fica	en	tro	lect	en	ect	izat
	ro	ent	omét	te	ect	enta	ct	ctr	tion	ra	tio	erat
Fin	er	ion	tion	er	ion	tion	on	ion	tion	on	ion	tion
	on	ent	ment	on	ent	ment	ly	ate	ally	ng	ing	ally
	nt	que	ique	nt	ter	ique	er	ing	tive	ly	ent	tive
	re	eur	aire	re	eur	teur	al	ent	ical	er	ate	able
	ue	ire	teur	ur	que	aire	te	ity	onal	ed	ity	lity

TABLE 4 – Top-5 de n-grammes dans les corpus

En début de mots français, les bi-grammes et les tri-grammes contiennent des préfixes (*co-, in-, dé-, re-, ré-, pro-, com-, dis-, pré-*) ; ils sont largement partagés par les deux corpus, ce qui permet d'affirmer que ce sont plutôt les éléments de la langue générale. Les quadri-grammes, au contraire, sont plus caractéristiques de chaque corpus (5 sur 10 résultats sont différents) ; ces éléments contiennent notamment les formants classiques (*hydr-, phot-, micr-, mult-, télé-,* etc.)

Au milieu, les bi-grammes ne représentent pas d'intérêt, tandis que les tri-grammes et les quadri-grammes sont bien propres à chaque corpus. Dans ces listes, il est possible de deviner les racines fréquentes (*-électr-, communic-,* etc.)

A la fin, la quasi-totalité des n-grammes est partagée par les deux corpus. Nous y retrouvons les suffixes propres aux verbes (*-er, -ir*), aux substantifs (*-ion, -ité, -eur, - ment*) et aux adjectifs (*-que*). Certes, une partie de ces suffixes peut correspondre aux adverbes (*-ment*) ou aux participes présent (*-ant*), mais l'usage que l'on en fait minimise l'influence de ces éléments.

Le choix entre les tri-grammes et les quadri-grammes n'est pas simple : d'un côté il est possible que les quadri-grammes puissent donner plus de précision, mais le nombre de candidats termes sera assez réduit ; les tri-grammes peuvent générer plus d'imprécision, mais le nombre de candidats termes sera plus élevé. Nous optons pour les tri-grammes qui se placent dans le premier tiers de la liste triée par fréquence décroissante, combinés avec un score de confiance : chaque tri-gramme du début et du milieu apporte un point ; les candidats dont le score est supérieur à trois pourront être retenus. Pour illustrer

6. Nous avons pris les premiers 30 % des deux listes.

la méthode sur le tableau 4, le mot *électronique* aura un score égal à 3, car il contient les tri-grammes *ect, ctr, tro*. Ce score s'élèvera à 4 si le tri-gramme *éle* est ajouté à la liste des tri-grammes du début.

Ce filtre sera appliqué au reste des deux listes de mots pour en extraire des candidats termes monolexicaux qui ne sont pas en tête de la liste par le biais de fréquences.

Les tri-grammes de la fin des mots qui jouent le rôle de suffixes seront utilisés pour l'apprentissage de patrons morpho-syntaxiques endogènes.

3.4 Patrons morphosyntaxiques endogènes et extraction de termes polylexicaux

Nous avons observé deux types de patrons pour l'extraction de candidats termes polylexicaux : les patrons morphosyntaxiques et les structures contrôlées de Vergne (2005).

Les deux types de patrons représentent des inconvénients. Les patrons morphosyntaxiques sont dépendants de la langue et l'efficacité de la méthode dépend de la richesse de la liste fournie, tandis que les structures contrôlées basées sur les annotations mot informatif – mot vide manquent de cohérence : n'importe quel mot informatif peut se trouver en tête de l'expression.

Notre méthode permet de profiter des points forts des deux approches sans avoir à résoudre les problèmes cités ci-dessus. Tout d'abord, la structure de patrons endogènes n'est pas figée : ils sont appris dans la fenêtre de 5 mots. Les patrons validés satisfont les conditions de commencer et de terminer par des mots informatifs, et de ne pas contenir de chiffres ou de signes de ponctuation, à l'exception de l'apostrophe qui est à ce stade annoté comme un mot vide.

La deuxième particularité de ces patrons est d'utiliser les tri-grammes à la fin de mots informatifs pour remplacer les étiquettes morphosyntaxiques. En effet, l'expérience démontre que les patrons obtenus contiennent des syntagmes nominaux, verbaux et adjectivaux. Par exemple, les patrons les plus fréquents pour le corpus français en chimie : *ion n*[7] *n ion (632), ion n n n ion (406), ion n ion (268), ion que (261), ent n n ion (219), que n n ion (201), ide que (197), tre n n ion (157), etc.*, correspondent aux patrons morphosyntaxiques suivants :
– NOM (PREP ?(DET ?)) NOM
– NOM ADJ
– VER (PREP | DET) NOM
L'apprentissage de patrons endogènes ne nécessite pas d'étiquettes morphosyntaxiques : les tri-grammes remplissent cette fonction. Certes, les patrons endogènes n'excluent pas un certain niveau de bruit. Afin d'éviter ce bruit, nous pouvons restreindre les mots vides valides à la liste de mots de schéma[8] générée à partir du corpus à l'étape de l'acquisition des patrons (table 5). Pour construire cette liste, nous avons repéré les mots non-informatifs les plus fréquents qui apparaissent entre les mots informatifs à l'intérieur des patrons.

Corpus	Mots de schéma
Chimie	le, de, du, ', un, être, et, à, en, l
Télécommunication	le, de, du, un, et, à, être, en, pour, au

TABLE 5 – Mots de schéma endogènes

En appliquant les patrons endogènes combinés avec les mots de schéma, nous arrivons à extraire les syntagmes susceptibles d'être des candidats termes polylexicaux, tenant compte de la règle du premier tiers (Top-30 %) de la liste. Cependant, il serait intéressant d'inclure une validation définitive par les ressources morphologiques endogènes.

Les syntagmes retenus ainsi varient en fonction du corpus et de la langue. Vu qu'en français la morphologie flexionnelle est plus prononcée et que la différence de longueur entre les mots vides et informatifs est plus évidente qu'en anglais, la méthode fournit de meilleurs résultats pour le français. Afin d'augmenter la précision des patrons morphologiques endogènes, nous avons ajouté une vérification complémentaire : l'un des mots doit obligatoirement être un candidat terme monolexical.

Il faut noter également que ce ne sont pas toujours les patrons les plus fréquents qui donnent le plus de résultats pertinents, car certains patrons contiennent les mots non-informatifs extérieurs à la liste de mots de schéma retenue. La variation de

7. *n* est un mot non-informatif
8. Pour simplifier, nous avons choisi ce terme pour regrouper les mots de schéma et les mots fortement liés (Enguehard, 1993).

résultats pour un même patron peut être assez intéressante. Certains patrons sont très productifs, d'autres le sont moins. Plus un patron est long, moins il fournit de résultats.

4 Évaluation des résultats

La table 6 contient l'évaluation [9] des composantes de la méthode sur les 100 premières occurrences des listes retenues. Un candidat terme monolexical a été validé soit s'il appartient de manière non-ambiguë au domaine en question, soit s'il peut éventuellement participer à la formation d'un candidat terme polylexical. Pour les candidats termes polylexicaux, nous nous sommes limités au premier critère.

Nous avons utilisé la plate-forme TERMOSTAT [10] pour obtenir des résultats de référence pour l'extraction des termes monolexicaux. L'utilisation de cette ressource pour évaluer les candidats polylexicaux semble injuste, car TERMOSTAT se limite à des syntagmes nominaux et fournit, évidemment, une précision très élevée.

Corpus	Langue	Précision Mono				Poly
		Baseline	Fréq	TFxIDF	RME [11]	PME [12]
Chimie	FR	54 %	**81 %**	**75 %**	**87 %**	69 %
	EN	**72 %**	44 %	64 %	69 %	85 %
Télécommunication	FR	77 %	**82 %**	76 %	53 %	78 %
	EN	**88 %**	58 %	77 %	53 %	69 %

TABLE 6 – Précision dans les Top-100 candidats

4.1 Candidats termes monolexicaux

Pour le corpus français en chimie le système a identifié 8 292 candidats termes monolexicaux qui résultent de la combinaison de la fréquence absolue maximale, du TFxIDF et des ressources morphologiques endogènes (table 6).

Nous pouvons constater que la fréquence absolue des mots informatifs fournit des bons résultats pour le français (81,5 % de précision en moyenne), mais se montre beaucoup moins efficace en anglais (51 % en moyenne). Dans les résultats en anglais, nous avons des bons candidats, comme *temperature, concentration, synthesis, electrochemical*, mais aussi des mots outils comme *same, first, thus, etc*. Cela s'explique par la baisse de l'efficacité de la méthode de J. Vergne sur les langues où la différence de longueur entre les mots informatifs et vides n'est pas importante.

La mesure TFxIDF est relativement plus stable (75,5 % pour le français et 70,5 % pour l'anglais en moyenne). Il faut remarquer que TFxIDF met en évidence un grand nombre de formules chimiques comme *C6F5* qui sont difficiles à reconnaître si l'on n'est pas un spécialiste du domaine.

La combinaison des deux méthodes permet de retrouver une partie de la terminologie du domaine, mais il faut tenir compte du bruit présent dans les résultats. Certes, nous pouvons appliquer les ressources morphologiques endogènes en tant que filtre complémentaire, mais de cette manière une grande partie des résultats corrects sera perdue. Par exemple, nous pouvons perdre les candidats termes, comme *eau* ou *ADN* pour la chimie et *Mac, chunk, AMRT, etc.* pour la télécommunication. Cela aura un effet négatif sur les candidats termes polylexicaux qui contiennent ces candidats termes et qui ne passeront pas la validation par la condition d'avoir un terme monolexical confirmé. Cependant, cela reste à vérifier au cours de futurs essais.

Le filtre constitué des ressources morphologiques endogènes défini dans la section 3.3.2 a donné des résultats positifs pour les deux langues, mais s'est montré nettement plus performant dans le domaine de la chimie, ce qui doit s'expliquer par la nomenclature forte dans ce domaine. L'évaluation des résultats est faite sur la liste ordonnée suivant le nombre décroissant des tri-grammes retenus, ce qui explique la longueur des mots de tête de la liste (table 7).

9. L'évaluation a été faite manuellement par l'auteur de l'article. En cas de doute, le candidat était annoté comme invalide. À terme, il est prévu de nous adresser à des experts pour préciser l'évaluation.
10. `http://termostat.ling.umontreal.ca`
11. Ressources morphologiques endogènes
12. Patrons morphologiques endogènes

Chimie		Télécommunication	
FR	EN	FR	EN
variationnellement	bioelectrochemistry	proportionnellement	internationalization
interribonucleotide	methylphenylboronic	internationalisation	internationalisation
metallointercalation	electroconductivity	telecommunication	implementations
proportionnellement	electropolymerization	repositionnement	differentiability
environnementaliste	electropolymerization84	significativement	operationalizes
spectrophotometric	spectroelectrochemistry	transactionnelle	cooperativeness
photosensibilisation	photoelectrochemical	defphysicallayerconfiguration	generalization
cristallographiea	electropolymerized	perfectionnement	radiocommunications
surdimensionnement	triphenylmethylborate	interconnections	considerations
electrocatalytic	methoxyphenylboronicacid	recommendation	externalization

TABLE 7 – Top-10 résultats pour les ressources morphologiques endogènes

Il faut remarquer que cette méthode permet de retrouver de bons candidats même dans les résultats identifiés après la ligne 300 :

– autoregistration, intersection, multiprotocol, videoconference, multiplexer, etc. (télécommunication, EN) ;
– monocarbonyles, transcriptionnel, nanolithographie, acidification, isocarbonyle, submicroscopique, aminopropylmercaptotriazole, etc. (chimie, FR).

4.2 Candidats termes polylexicaux

Les patrons endogènes ont permis d'extraire 8 631 candidats termes pour le corpus français en chimie. Nous avons imposé la condition que le candidat terme polylexical doive contenir au moins un candidat terme monolexical. La taille de patrons varie entre 2 et 5 éléments. La performance de la méthode est illustrée dans la table 6.

Tout comme dans le cas de patrons morphosyntaxiques traditionnels, il est difficile de limiter le résultat aux bons candidats termes. La méthode présente un grand avantage : les patrons sont extraits à partir du texte sans aucune analyse préalable. Ainsi, nous arrivons à extraire des candidats termes assez complexes, comme :

– réduction abiotique du sulfate
– décroître de façon monotone
– nettoyage de l ' électrode
– système de conversion de énergie
– polarisation anodique et cathodique
– complexe tétranucléaire

Nous pensons que ces candidats sont complexes pour les systèmes classiques parce que l'étiqueteur morphosyntaxique [13] se perd dans une terminologie inconnue de son vocabulaire, ce qui met en question toute l'analyse par des patrons prédéfinis. Nous illustrons cela sur les deux occurrences du candidat terme *complexes tétranucléaires* dans le corpus en chimie qui sont étiquetées avec des erreurs différentes :

– des PRP :det du
– complexes ADJ complexe
– tétranucléaires NOM tétranucléaires

OU

– complexes NAM Complexes
– tétranucléaires ADJ tétranucléaires

En même temps, la méthode de patrons endogènes a une faiblesse. Comme le patron ne part pas d'étiquette précise, il est assez difficile de classer les résultats. Par exemple, les patrons contenant le tri-gramme -ent renvoient aussi bien aux substantifs qu'aux adverbes (*traitement du eau, méthode être fortement, etc.*).

La classification est un point d'autant plus difficile qu'il existe plusieurs critères de départ :

1. patron
2. fréquence
3. terme(s) monolexical(aux) présents

13. Dans cet exemple, il s'agit de l'étiquetage morphosyntaxique du corpus en chimie (FR) par TreeTagger (Schmid, 1994)

Corpus en chimie			Corpus en télécommunication		
Patron	Collocation	Fréq	Patron	Collocation	Fréq
[nce, que]	insuffisance technique	3	[ure, n, ert, n, née]	procédure de transfert de donnée	6
	séquence nucléotidique	3	[eur, ire]	valeur binaire	2
	résistance mécanique	3		leur propriétaire	5
	séquence peptidique	6		utilisateur stationnaire	2
	puissance massique	10		erreur binaire	11
	distance interatomique	46	[ire, n, mps, n, ion]	réduire le temps de création	3

TABLE 8 – Quelques exemples de résultats par patron sur les deux corpus en français

Classer les résultats par patron semble assez risqué, car nous avons déjà évoqué la variabilité de parties du discours pour le même n-gramme. En même temps, cela a un aspect intéressant qui consiste à voir la productivité du patron (table 8). La vue par patrons pourra servir à constituer un anti-dictionnaire de patrons pour les futures analyses. La fréquence est peut-être le critère le plus commode pour présenter le résultat (table 9).

Candidat	Fréq	Patron	Candidat	Fréq	Patron
acide nitrique	152	[ide, que]	spectrométrie de masse	89	[rie, n, sse]
mettre en évidence	127	[tre, n, nce]	atome de carbone	88	[ome, n, one]
métal de transition	124	[tal, n, ion]	énergie de adsorption	82	[gie, n, ion]
être également	124	[tre, ent]	résultat et discussion	78	[tat, n, ion]
complexe de ruthénium	123	[exe, n, ium]	corrosion du cuivre	76	[ion, n, vre]
acide phosphorique	123	[ide, que]	composant électrochimique	75	[ant, que]
température ambiant	120	[ure, ant]	génie électrique	72	[nie, que]
transfert de charge	118	[ert, n, rge]	milieu acide	71	[ieu, ide]
centre métallique	98	[tre, que]	prendre en compte	70	[dre, n, pte]
être possible	90	[tre, ble]	passage de drake	70	[age, n, ake]

TABLE 9 – Top-20 candidats polylexicaux pour le corpus en chimie (FR)

La fréquence permet de voir les associations polylexicales les plus utilisées dans le corpus. Ces associations peuvent être soit des candidats termes polylexicaux (*atome de carbone, acide phosphorique, etc.*), soit tout simplement des éléments de la langue générale largement utilisés dans les textes du même style (*résultat et discussion, mettre en évidence, être également*). De ce point de vue, il sera intéressant de vérifier la possibilité d'appliquer la méthode à l'identification du lexique transdisciplinaire.

Cependant, le tri par fréquence a ses inconvénients. La tête de la liste est majoritairement occupée par les patrons à deux ou à trois éléments. Les candidats termes à quatre ou à cinq éléments apparaissent moins fréquemment, ils ont donc peu de chances d'entrer dans la première partie de la liste. Par exemple, le candidat *optimisation du paramètre de maille* est à la fin de la liste avec la fréquence égale à deux.

La troisième possibilité est de prendre en compte les candidats termes monolexicaux. Cette option a un certain avantage : elle permet de trier les candidats termes du point de vue de l'ensemble. Le lemmatiseur de TreeTagger se trompe des fois sur les lemmes des mots de schéma (*de le* ou *du*), ce qui démultiplie les associations et corrompt les fréquences.

Cette classification par les mots en commun est également intéressante car elle donne une idée du comportement du candidat terme monolexical dans le corpus spécialisé (termes associés, variantes, exemples, etc.). Par exemple, le programme a identifié plusieurs candidats termes polylexicaux contenant le candidat terme monolexical *chunk* (table 10).

Nous pouvons observer que les candidats *chunk de donnée* et *chunks de donnée* sont séparés car le mot *chunk* est un emprunt et TreeTagger n'arrive pas à le lemmatiser. Cela joue sur la fréquence : si les deux candidats étaient réunis, leur fréquence serait égale à 57.

La paire *nouveau chunk de contrôle* et *chunk de contrôle* représente un autre cas : un candidat terme contient l'autre. Dans ce cas précis, il n'y a pas d'intérêt de retenir le plus grand, mais il existe d'autres cas où il est intéressant de garder les deux candidats, par exemple *partition du générateur photovoltaïque* et *générateur photovoltaïque*.

Collocation	Fréq	Patron
chunk de donnée	25	[unk, n, née]
format du chunk cookie-ack	2	[mat, n, unk, ack]
format du chunk heartbeat-ack	2	[mat, n, unk, ack]
format du chunk init-ack	2	[mat, n, unk, ack]
format du chunk asconf-ack	2	[mat, n, unk, ack]
nouveau chunk de contrôle	6	[eau, unk, n, ôle]
chunks de donnée	32	[nks, n, née]
chunk de contrôle	10	[unk, n, ôle]

TABLE 10 – Candidats polylexicaux contenant le candidat terme monolexical chunk

Il est alors intéressant de proposer les candidats termes polylexicaux lors de la validation d'un candidat terme monolexical, car cela fournit à l'utilisateur plus d'informations sur ce dernier. Les résultats obtenus à cette étape du travail permettent de planifier quelques améliorations de l'algorithme général du système, notamment l'élagage de certains types de termes polylexicaux et la présentation des résultats.

5 Conclusion et perspectives

La méthode proposée permet d'extraire les candidats termes mono et polylexicaux à partir d'un corpus de spécialité. Le principal avantage de la méthode est son applicabilité à une large palette de langues et de corpus. Elle ne nécessite aucune ressource morphologique ou morphosyntaxique extérieure. Ainsi, cette méthode peut être appliquée sur un corpus de tout domaine de spécialité et toutes les ressources sont alors générées par le programme.

La méthode peut être utilisée pour d'autres langues que le français et l'anglais à condition qu'un tokeniseur et un lemmatiseur soient disponibles pour la langue ciblée. Même si la tâche de lemmatisation nécessite un étiquetage morphosyntaxique, notre méthode ne l'utilise pas. De cette manière, la méthode exploite l'outil d'étiquetage morphosyntaxique différemment des autres méthodes de l'ET.

La méthode combine des méthodes statistiques (fréquence absolue et TFxIDF) et des approches mixtes (annotation par mots informatifs et vides, apprentissage des mots de schéma endogènes). La particularité de la méthode est d'utiliser des n-grammes de caractères pour remplacer des listes de formants classique et des étiquettes morphosyntaxiques.

Les résultats obtenus sur le prototype permettent d'envisager des améliorations de la méthode. Notamment, il est intéressant d'automatiser le choix des coefficients et de la taille de n-grammes en fonction de la langue, car cela pourrait donner des meilleurs résultats pour l'anglais ou pour d'autres langues.

La deuxième piste consistera à tester la possibilité de corréler les n-grammes avec les syllabes pour obtenir les ressources morhologiques et morphosyntaxiques endogènes. Cela pourrait également jouer sur l'efficacité de la méthode appliquée sur d'autres langues.

Le système sera testé sur des corpus multidomaines composés par genre afin de voir si la méthode pourrait servir à l'identification du lexique et des expressions transdisciplinaires.

Une validation des résultats par des experts de domaines est prévue dans le cadre de notre projet de recherche.

Références

ANTHONY, L. (2005). AntConc : design and development of a freeware corpus analysis toolkit for the technical writing classroom. *In International Professional Communication Conference Proceedings*, pages 729–737.

AUBIN, S. et HAMON, T. (2006). Improving term extraction with terminological resources. *In Advances in Natural Language Processing*, pages 380–387. Springer.

BERNHARD, D. (2006). *Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales*. Thèse de doctorat, Université Joseph Fourier – Grenoble I.

BOURIGAULT, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de TALN*, pages 75–84, Nancy.

BOURIGAULT, D. et FABRE, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, 25:131–151.

BOURIGAULT, D., GONZALEZ-MULLIER, I. et GROS, C. (1996). LEXTER, a Natural Language Processing Tool for Terminology Extraction. *7th EURALEX International Congress*.

BOURIGAULT, D. et JACQUEMIN, C. (1999). Term extraction+ term clustering : An integrated platform for computer-aided terminology. *In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 15–22.

DAILLE, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *In The Balancing Act : Combining Symbolic and Statistical Approaches to Language, Workshop at the 32nd Annual Meeting of the ACL (ACL'94)*, Las Cruces, New Mexico, USA.

DAILLE, B. (2003). Conceptual structuring through term variations. *In* BOND, F., KORHONEN, A., MACCARTHY, D. et VILLACICENCIO, A., éditeurs : *ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pages 9–16.

DROUIN, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, XII:45–64.

DROUIN, P. et LANGLAIS, P. (2006). Évaluation du potentiel terminologique de candidats termes. *In Actes de JADT*, volume 2006, page 8.

ENGUEHARD, C. (1993). Acquisition de terminologie à partir de gros corpus. *Informatique & Langue Naturelle*, pages 373–384.

ESTOPÀ, R., VIVALDI, J. et CABRÉ, T. (2000). Extraction of monolexical terminological units : requirement analysis. *In Workshop Proceedings Second International Conference on Language Resources and Evaluation. Terminology Resources and Computation*, volume 56, pages 51–56, Athens, Greece.

FRANTZI, K., ANANIADOU, S. et MIMA, H. (2000). Automatic Recognition of Multi-Word Terms : the. *International Journal on Digital Libraries*, 3(2):115–130.

JACQUEMIN, C. (1997). *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. Thèse de doctorat, Institut de Recherche en Informatique de Nantes.

JACQUEY, E., TUTIN, A., KISTER, L., JACQUES, M.-p., HATIER, S. et OLLINGER, S. (2013). Filtrage terminologique par le lexique transdisciplinaire scientifique : une expérimentation en sciences humaines. *In Terminologie et Intelligence Artificielle (TIA)*, Paris.

MCNAMEE, P. (2008). N-gram Tokenization for Indian Language Text Retrieval. *In Working Notes of the Forum for Information Retrieval Evaluation*.

MCNAMEE, P., NICHOLAS, C. et MAYFIELD, J. (2009). Addressing morphological variation in alphabetic languages. *In 32nd Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 75–82.

MORIN, E. et DAILLE, B. (2006). Comparabilité de corpus et fouille terminologique multilingue. *TAL*, 47:113–136.

MORIN, E. et DAILLE, B. (2012). Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique. *In Conférence conjointe JEP-TALN-RECITAL 2012*, pages 141–154.

OROBINSKA, O., LYON, E. et CHAUCHAT, J.-h. (2013). Enrichissement d'une ontologie de domaine par extension des relations taxonomiques à partir d'un corpus spécialisé. *In Terminologie et Intelligence Artificielle (TIA)*, volume 704, Paris.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of International Conference on New Methods in Language Processing*, Manchester, {UK}.

TUTIN, A. (2007). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, XII:5–14.

VERGNE, J. (2003). Un outil d'extraction terminologique endogène et multilingue. *Actes de TALN*, 2:139–148.

VERGNE, J. (2004). Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. *In Actes de JADT*, Louvain.

VERGNE, J. (2005). Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. *In Actes de CIDE*, pages 155–168.

WITSCHEL, H. F. (2005). Terminology Extraction and Automatic Indexing. *In Terminology and Knowledge Engineering (TKE)*, pages 1–12.