La "multi-extraction" comme stratégie d'acquisition optimisée de ressources (non) terminologiques

Blandine Plaisantin Alecu[1], Izabella Thomas[2], Julie Renahy[1]
(1) Prolipsia SAS, TEMIS Innovation, 18 r Alain Savary, 25000 Besançon
(2) Centre Tesnière, Université de Franche-Comté, UFR SLHS, 30 r Mégevand, 25030 Besançon
{blandine.alecu, julie.renahy}@prolipsia.com,
izabella.thomas@univ-fcomte.fr

RÉSUMÉ

A partir de l'évaluation d'extracteurs de termes menée initialement pour détecter le meilleur outil d'acquisition du lexique d'une langue contrôlée, nous proposons dans cet article une stratégie d'optimisation du processus d'extraction terminologique. Nos travaux, menés dans le cadre du projet ANR Sensunique, prouvent que la « multi-extraction », c'est-à-dire la coopération de plusieurs extracteurs de termes, donne des résultats significativement meilleurs que l'extraction via un seul outil. Elle permet à la fois de réduire le silence et de filtrer automatiquement le bruit grâce à la variation d'un indice relatif au potentiel terminologique.

ABSTRACT

Multi-extraction as a strategy of optimized extraction of terminological and lexical resources

Based on the evaluation of terminological extractors, initially to find the best tool for building a controlled language lexicon, we propose a strategy of optimized extraction of terminological resources. Our work highlights that the cooperation of several extraction tools gives better results than the use of a single one. It both reduces silence and automatically filters noise thanks to a variable related to termhood.

MOTS-CLÉS : terminologie, extraction, langue contrôlée, potentiel terminologique, filtrage de termes.
KEYWORDS : terminology, extraction, controlled language, termhood, term filtering.

1 Introduction

1.1 Contexte, problématique et objectifs

Le présent article fait état de recherches effectuées dans le cadre du projet Sensunique (ANR-EMMA-2010-039), qui fait suite au projet LiSe (ANR-06-SECU-007) dans lequel nous avions conçu une méthodologie de contrôle de la langue française et développé un prototype expérimental d'aide à la rédaction en Langue Contrôlée (LC). Une des ambitions du projet Sensunique est d'alléger et de fiabiliser la tâche de conception d'une LC par un linguiste, en mettant à sa disposition des outils de recensement du Lexique d'une LC (désormais LLC) ; c'est sur ce point que nous concentrons cet article.

Dans notre approche, une LC doit être circonscrite à un domaine et à un environnement de rédaction précis, c'est-à-dire pour un public restreint et pour un type de textes particulier ; on peut dès lors parler d'une LC « sur mesure » ; la précision des facteurs

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 511–518,
Grenoble, 4 au 8 juin 2012.

influençant le texte à générer, et donc de la LC sous-jacente, différencie notre approche (Renahy et al., 2009) de celles des travaux menés sur le français[1]. En outre, elle est issue d'une analyse de corpus (généralement de petite taille), lequel doit recenser l'ensemble des textes en vigueur pour l'activité et le public concernés.

A notre connaissance, aucun outil dédié au recensement d'un LLC n'existe à ce jour. Cependant, puisque l'acquisition d'un LLC peut être comparée jusqu'à un certain point à l'acquisition terminologique, nous avons choisi de nous appuyer sur les Extracteurs de Termes (EdT), tout en tentant d'améliorer leurs résultats pour qu'ils répondent à nos besoins. Plus précisément, nous avons évalué le bénéfice que nous pourrions tirer de la coopération de plusieurs EdT. De multiples travaux fondés sur la coopération d'outils ont démontré son intérêt : en premier pour la reconnaissance vocale avec le système ROVER (Fiscus, 1997), repris, pour n'en citer que quelques uns[2], pour des analyseurs syntaxiques (Brunet-Manquat, 2004) ou des étiqueteurs morphosyntaxiques (Serp, 2008). Mais ce principe n'a jamais été appliqué, à notre connaissance, aux EdT.

1.2 LLC et Ressource Terminologique (RT)

Les notions de RT, résultat de l'acquisition terminologique, et LLC se recoupent et se distinguent à la fois. La principale différence concerne leurs périmètres respectifs. Parce que tous deux visent à couvrir un domaine particulier, renvoyant à des concepts spécifiques[3], RT et LLC recensent chacun des termes. Mais le périmètre d'une RT s'arrête aux unités terminologiques spécifiques du domaine alors qu'un LLC doit contenir l'ensemble du lexique nécessaire à la rédaction d'un texte dans sa globalité. Møller *et al.* (2006) parlent de « mots » (référant alors à des unités monolexémiques comme multilexémiques) afin de ne pas confondre les unités d'un LLC avec des unités terminologiques. Nous choisissons, quant à nous, de considérer comme Unités Lexicales[4] (UL) toutes les unités d'un LLC.

Pour être exhaustif, un LLC doit contenir des UL non spécifiques (par extension, non terminologiques). On ne peut pas appliquer une dichotomie spécifique-non spécifique : il s'agit, comme le dit Camlong (1996) d'un continuum allant du vocabulaire terminologique du domaine au vocabulaire général. Un texte écrit en LC peut inclure différents types d'UL, illustrés ici avec des exemples de protocoles d'immunobiologie : les termes du domaine (*anticorps monoclonaux*) ; ceux d'un autre domaine (*fenêtres informatiques*) ; les UL du lexique général prenant un sens spécifique dans le domaine traité (*population* (*bactérienne*)) ; celles du lexique général rentrant dans la composition de termes (*anticorps de chèvre*) ou gardant leur sens courant (*échantillons divers*) et les mots grammaticaux. Une RT ne suffit donc pas à l'exhaustivité d'un LLC.

Puisqu'un LLC ne contient pas que des UL terminologiques, nous avons imaginé une

[1] Dont l'unique exemple est Le Français Rationalisé, du GIFAS (1990).

[2] Voir Brunet-Manquat (2004) pour une liste plus exhaustive.

[3] En admettant que ces concepts sont dénommés par des termes.

[4] Nous reprenons ici la notion d'unité lexicale telle que définie par (L'Homme, 2005)

stratégie de mise en exergue du statut terminologique des candidats, extraits par les EdT, basée sur leur potentiel terminologique (*termhood*), c'est-à-dire le degré de spécialisation de leur sens dans le domaine à l'étude (Kageura et Umino, 1996).

1.3 Hypothèse

Nous posons l'hypothèse (H1) que l'utilisation simultanée de plusieurs EdT (que nous appellerons désormais la « multi-extraction ») est plus profitable (qu'un seul) au recensement du LLC, hypothèse que l'on peut subdiviser en :

- (H1.1) : les résultats proposés par plusieurs EdT sont les candidats-termes (CT) les plus pertinents, et peuvent être considérés comme UL terminologiques (la multi-extraction permet de déterminer le statut terminologique d'une UL en faisant ressortir son potentiel terminologique).
- (H1.2) : les CT non valides sont des UL candidates non terminologiques potentiellement pertinentes (le bruit des EdT, dans leur fonction initiale de recensement des termes, peut diminuer le silence, dans la fonction détournée de recensement des UL d'un LLC). Si cela est confirmé, la tâche de recensement d'un LLC peut être organisée, en classant les résultats par poids terminologique[5] (Pt) et/ou comme aide au filtrage du bruit pour l'acquisition de RT.

2 Matériel et méthodes

2.1 Corpus et lexique de référence

Nous avons constitué un corpus de référence de 14 modes opératoires d'immunobiologie (10 064 mots) de l'Établissement Français du Sang (EFS) Bourgogne Franche-Comté[6]. Notons que la méthodologie pensée est indépendante du domaine. Nous avons construit manuellement sur la base de ce corpus un LLC de référence, grâce à des critères linguistiques, la consultation de ressources terminologiques[7] et d'experts métier[8]. Le lexique de référence obtenu contient 1 512 UL (lemmes) pour 1 729 formes fléchies (utilisées en corpus), 7 catégories syntaxiques fonctionnelles distinctes (distinction minimale nécessaire pour un LLC : Adjectif, Adverbe, Nom, Nom propre, Verbe au participe passé, Verbe au participe présent, Verbe hors participes), 92 matrices morphosyntaxiques distinctes (exemple : *Nom Prep Det Nom Prep Det Nom Prep Nom* pour *fraction de l'immunoglobuline de l'antisérum de lapin*) et 2 statuts lexico-terminologiques distincts (terminologique et général).

2.2 Pré-sélection des EdT comme outils de recensement du LLC

Nous avons établi les critères suivants, afin d'estimer l'utilisabilité et l'adéquation

[5] Pt est un indice de fiabilité d'une ULC en tant que terme (relatif à son potentiel terminologique).

[6] Documents décrivant le déroulement détaillé et structuré des différentes étapes d'une manipulation.

[7] Le Grand Dictionnaire Terminologique, Termium Plus, le dictionnaire médical Masson 5ème édition.

[8] EFS Bourgogne Franche-Comté, partenaire Santé dans le projet Sensunique.

technique des EdT à nos besoins et de nous limiter à 3 EdT (coût raisonnable de la tâche d'évaluation) : langue (français), méthode (non purement statistique [9] : linguistique ou hybride), disponibilité (de suite), licence (libre ou commerciale : dans ce cas, coût faible ou nul), maturité de l'outil (non prototype), environnement informatique (Unix), modalité d'exécution (service web ou appel en ligne de commande), temps d'exécution (respectant le seuil d'appel en web service) et domaine d'application (non spécifique).

Ces critères nous ont menés aux EdT Acabit (Daille, 1994), TermoStat (Drouin, 2003) et YaTeA (Aubin et al., 2006). Acabit procède par identification de groupes nominaux complexes sur des matrices syntagmatiques pour extraction de bi-termes, regroupement de variantes (à partir de ces bi-termes) puis filtrage statistique. YaTeA enchaîne identification de groupes nominaux à partir de frontières morphosyntaxiques, calcul de leurs structures en tête et modifieur, puis exploitation de ces structures pour l'analyse des groupes nominaux restants. Enfin, TermoStat fonctionne par détection de CT sur patrons morphosyntaxiques puis pondération et filtrage selon la spécificité de chaque CT (méthode de mise en opposition de corpus spécialisés et non spécialisés). En outre, YaTeA et TermoStat ont l'avantage d'extraire des termes simples en plus des termes complexes ; et TermoStat est le seul à extraire également des termes non nominaux.

2.3 Evaluation

Deux des tâches du linguiste lors de la conception d'un LLC ont été évaluées :

1. Recensement des UL (de l'ensemble des UL d'un LLC) ;
2. Recensement des termes (des UL de statut terminologique d'un LLC).

Pour chacune de ces tâches, nous avons procédé à 3 expérimentations :

1. Évaluation des résultats de chaque EdT pris séparément ;
2. Évaluation des résultats cumulés de tous les EdT (union) ;
3. Évaluation des résultats consolidés, ou communs (intersection).

Pour chaque évaluation, nous avons calculé les mesures suivantes :

- Précision : P = (formes extraites correctes) / (formes extraites) ;
- Rappel : R = (formes extraites correctes) / (formes de référence) ;

2.3.1 Appariement des résultats

Notre objectif étant d'estimer la capacité des EdT à recenser les UL (terminologiques ou non) et non leur capacité de lemmatisation ou de variation terminologique, nous avons opté pour comparer les formes fléchies. Contrairement à Hamon (2000), nous ne cherchons pas les différents types d'erreurs, mais évaluons la présence d'une forme extraite dans le lexique de référence. Pour ne pas comparer les regroupements opérés par les EdT, nous avons extrait une liste des formes fléchies de tous les CT. Nous avons ainsi pu apparier les formes fléchies candidates avec celles du lexique de référence.

[9] A cause de la taille estimée des corpus utilisés pour concevoir une LC.

3 Résultats et discussion

3.1 Tâche 1 : Recensement des UL

Expérimentations	Outil(s)	P	R
Résultats d'un EdT	TermoStat	64 %	40 %
	YaTeA	43 %	**52 %**
	Acabit	44 %	17 %
Résultats cumulés (union)	TermoStat ∪ YaTeA	44 %	68 %
	TermoStat∪ ACABIT	55 %	48 %
	YaTeA ∪ ACABIT	41 %	59 %
	TermoStat ∪ YaTeA ∪ ACABIT	**42 %**	**72 %**
Résultats communs (intersection)	TermoStat ⋂ YaTeA	**74 %**	22 %
	TermoStat ⋂ ACABIT	63 %	9 %
	YaTeA ⋂ ACABIT	62 %	11 %
	(TermoStat ⋂ YaTeA) ou (TermoStat ⋂ ACABIT) ou (YaTeA ⋂ ACABIT)[10]	**69 %**	29 %

TABLE 1 – Tâche de recensement des UL

La première expérimentation montre que, dans le meilleur des cas, en utilisant un seul EdT, 52 % (valeur en gras, Table 1) des UL du LLC sont recensées, ce qui est loin de satisfaire le critère d'exhaustivité.

Pour la deuxième expérimentation, le cumul des résultats des 3 EdT permet de couvrir quasiment ¾ du lexique de référence (rappel de 72 %, en gras, Table 1). Ceci confirme l'hypothèse H1 : la multi-extraction permet de mieux couvrir le LLC que l'utilisation d'un seul EdT. En revanche, dans ce cas, il reste à filtrer manuellement près de 60 % des propositions et il devient nécessaire de filtrer automatiquement le bruit.

La troisième expérimentation démontre que la combinaison d'EdT obtenant la meilleure précision est TermoStat + YaTeA (74 %, Table 1). Cependant, il apparaît également que n'importe quelle combinaison de 2 EdT donne une précision de 69 % (donc légèrement plus faible). Nous proposons de filtrer le bruit sur cette dernière combinaison en considérant que ce cas de figure sera plus généralisable (à d'autres domaines) dans la mesure où il « suffit » qu'une UL soit proposée par 2 EdT pour être estimée pertinente. L'opération consisterait à augmenter la valeur d'un indice relatif au potentiel terminologique des UL concernées (proposées par 2 EdT), et creuser ainsi l'écart avec celles qui ne sont pas proposées que par un EdT. Cela revient à distinguer les ULC à fort potentiel terminologique de celles à faible potentiel terminologique en les

[10] Sur l'ensemble des résultats communs à (proposés par) au moins 2 EdT, quels qu'ils soient.

classant et non en supprimant ces dernières.

3.2 Tâche 2 : Recensement des termes

Expérimentations	Outil	P	R
Résultats d'un EdT	TermoStat	28 %	52 %
	YaTeA	16 %	58 %
	ACABIT	14 %	17 %
Résultats cumulés (union)	TermoStat ∪ YaTeA	16 %	76 %
	TermoStat∪ ACABIT	23 %	60 %
	YaTeA ∪ ACABIT	14 %	63 %
	TermoStat ∪ YaTeA ∪ ACABIT	15 %	**79 %**
Résultats communs (intersection)	TermoStat ⋂ YaTeA	**37 %**	33 %
	TermoStat ⋂ ACABIT	24 %	11 %
	YaTeA ⋂ ACABIT	26 %	14 %
	(TermoStat ⋂ YaTeA) ou (TermoStat ⋂ ACABIT) ou (YaTeA ⋂ ACABIT)	31 %	39 %
	TermoStat ⋂ YaTeA ⋂ ACABIT	32 %	9 %

TABLE 2 – Tâche de recensement des termes

La mesure de précision de 37 % (TermoStat ⋂ YaTeA, Table 2) pour les résultats communs permet de valider l'hypothèse H1.1. : la multi-extraction aide à déterminer le statut terminologique d'une UL en faisant ressortir son potentiel terminologique. La différence (même faible) de rappel entre les résultats cumulés des 3 EdT pour le recensement des termes (79 %, Table 2) et le recensement des UL (72 %, Table 1) démontre qu'une partie des candidats proposés ne sont pas des termes mais sont, pour le LLC, des UL correctes, de statut non terminologique (hypothèse H1.2). Bien que les résultats soient moindres que ceux escomptés, ils demeurent satisfaisants et il est possible qu'ils soient meilleurs sur des corpus plus conséquents.

En résumé, cumuler les résultats de tous les EdT permet de couvrir 79 % des termes (rappel TermoStat ∪ YaTeA ∪ ACABIT, Table 2), et le meilleur moyen d'aider à déterminer le statut d'une UL est, non pas de se baser sur les résultats communs aux 3 EdT (contrairement à ce que nous attendions), mais de se baser sur les résultats communs aux 2 EdT TermoStat et YaTeA (précision de 37 % dans la Table 2). Ceci valide tout de même l'hypothèse selon laquelle la multi-extraction aide à recenser et à organiser la validation d'un LLC.

3.3 Stratégie basée sur les observations

Le fait que la multi-extraction permette à la fois de réduire le silence et le bruit des propositions nous incite à introduire un indice, relatif au potentiel terminologique,

variable en fonction des résultats des EdT. Nous proposons d'attribuer à chaque UL candidate un poids terminologique Pt ; puis de faire varier ce Pt initialement nul en fonction des résultats de chaque EdT. Nous proposons une stratégie de variation du Pt consistant à augmenter du Pt d'une UL en fonction du nombre d'EdT qui la proposent comme candidate. Ce principe traduit bien les faits suivants :

- un EdT propose un candidat « terme » donc ayant un potentiel terminologique ;
- un candidat a d'autant plus de probabilité d'être un terme fiable qu'il y a d'EdT le proposant (comme candidat) ;
- les résultats pourront être classés et validés selon la valeur du Pt.

Notons que les 3 EdT utilisés intègrent *a priori* (TermoStat) ou *a posteriori* (Acabit et YaTeA) des indices statistiques afin de cerner les termes les plus pertinents. Bien qu'il puisse être intéressant de coupler les valeurs de ces indices (différents pour chaque candidat) au Pt, nous avons fait le choix de ne pas le faire expressément, afin de ne pas rendre l'algorithme de pondération dépendant des EdT utilisés (et puisque le calcul de Pt repose déjà indirectement sur l'efficacité des EdT utilisés).

4 Conclusion

L'exploitation des résultats des expérimentations menées nous a permis de proposer une méthode d'acquisition d'un LLC et d'optimisation de l'acquisition terminologique. Elle repose sur la coopération de plusieurs EdT et permet de faire ressortir le potentiel terminologique des candidats, de réduire le silence obtenu avec un seul EdT et de filtrer le bruit en classant les candidats sur un indice de potentiel terminologique.

Outre concevoir un outil dédié au recensement d'un LLC, l'originalité de ces travaux réside dans le fait que nous proposons de faire coopérer plusieurs EdT pour améliorer leurs résultats et mettre en place un système de filtrage, alors que les travaux antérieurs d'évaluation d'EdT visaient leur mise en opposition (ou classement) (Grabar, 2004).

Nous avons mis au point, pour améliorer l'extraction terminologique, un système à base de vote, sur la méthode dite du "vote à la majorité" (Brunet-Manquat, 2004) où plus un terme est proposé par différents EdT, plus sa fiabilité est renforcée.

Nous avons conçu une plateforme implémentant cette méthode. Elle intègre les étiqueteurs morphosyntaxiques Brill[11] et TreeTagger, le lemmatiseur Flemm (Namer, 2000) pour les analyses préalables et nécessaires à l'extraction, et les EdT Acabit, TermoStat et YaTeA. Elle permet de procéder à l'extraction et à l'organisation de lexique terminologique et non-terminologique à partir d'un corpus français au format XML TEI P5. La plateforme est paramétrée par défaut sur le principe du "vote à la majorité" mais l'utilisateur peut ajuster le poids attribué à chaque EdT, en fonction de ses besoins, afin de rendre cette plateforme aussi flexible que possible. Nous avons également intégré un module d'interrogation de ressources terminologiques ou lexicales existantes, ce qui permet de renforcer, une nouvelle fois, la fiabilité du potentiel

[11] Avec le lexique et le fichier de règles fournis par l'ATILF-CNRS, de Nancy.

terminologique des candidats.

Références

AUBIN, S. et HAMON, T. (2006). Improving Term Extraction with Terminological Resources. *In : Advances in Natural Language Processing, 5th International Conference on NLP (FinTAL'2006),* Springer, 2006, p. 380-387.

BRUNET-MANQUAT, F. (2004). Fusionner pour mieux analyser : Conception et évaluation de la plate-forme de combinaison. *In Actes de TALN-2004 (Traitement automatique des langues naturelles).* Fez, Maroc, 19-22 avril 2004. vol. 1/1, p. 111-120.

CAMLONG, A. (1996). Méthode d'analyse lexicale textuelle et discursive, Paris, Orphrys.

DAILLE, B. (1994). 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.

DROUIN, P. (2003). Term Extraction Using non-Technical Corpora as Point of Leverage. *In : Terminology,* vol.9, n°1, *John Benjamins Publishing Company: Amsterdam/Philadelphia*, p. 99-115.

FISCUS, J.G. (1997), A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER). *In IEEE Workshop on Automatic Speech Recognizer and Understanding,* p. 347-354.

GRABAR N. (2004). Terminologie médicale et morphologie : Acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique, Thèse de Doctorat en Informatique Médicale, Université Paris 6.

HAMON, T. (2000). Variation sémantique en corpus spécialisé : Acquisition de relations de synonymie à partir de ressources lexicales, Thèse de Doctorat en Informatique, Université Paris Nord.

KAGEURA, K. et UMINO, B. (1996). Methods for automatic term recognition: A review. *In : Terminology*, 3(2), p. 259–289.

L'HOMME, M.-C. (2005). Sur la notion de terme. *In Meta : journal des traducteurs*, vol. 50, n° 4, p. 1112-1132.

MØLLER, M. H., CHRISTOFFERSEN, E., HANSEN, M. (2006). Building a Controlled Language Lexicon for Danish. *In LSP and Professional Communication*, vol. 6, Nr. 1, p. 12-38.

NAMER, F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *In Traitement Automatique des Langues* ;vol. 41/2, p. 523-547.

RENAHY, J., DEVITRE, D., THOMAS, I., DZIADKIEWICZ, A., (2009). Controlled language norms for the redaction of security protocols: finding the median between system needs and user acceptability. *In Proceedings of the 11th International Symposium on Social Communication, Santiago de Cuba,* Cuba, 19-23 January 2009, p. 289-293.