Filtrage semi-automatique des variantes de termes dans un processus d'indexation contrôlée

Fabienne Ville-Ometz, Jean Royauté, Alain Zasadzinski

▶ **To cite this version:**

Fabienne Ville-Ometz, Jean Royauté, Alain Zasadzinski. Filtrage semi-automatique des variantes de termes dans un processus d'indexation contrôlée. Jun 2004. sic_00001255

HAL Id: sic_00001255

https://archivesic.ccsd.cnrs.fr/sic_00001255v1

Submitted on 8 Dec 2004

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Filtrage semi-automatique des variantes de termes dans un processus d'indexation contrôlée

Fabienne Ville-Ometz (1), Jean Royauté (2), Alain Zasadzinski (3)

(1) et (3) URI - INIST CNRS, 2 allée du Parc de Brabois, 54514 Vandoeuvre-lès-Nancy Cedex

(2) LIF – CNRS, 163 Av. de Luminy, Case 901, 13288 Marseille Cedex 9

ometz@inist.fr, jean.royaute@lif.univ-mrs.fr, zasadzinski@inist.fr

RÉSUMÉ. L'extraction de l'information pertinente contenue dans les textes par des procédures automatisées de type TALN constitue une opération essentielle dans le processus de fouille de données textuelles. Nous réalisons cette opération à partir de la plate-forme d'ingénierie linguistique ILC qui reconnaît et extrait du corpus les termes ainsi que leurs variantes linguistiques. Nous présentons une méthodologie de constitution de règles visant à améliorer la reconnaissance de la variation terminologique en anglais par l'exploitation de critères syntaxiques et morpho-syntaxiques. Ces améliorations ont pour objectif d'obtenir un meilleur filtrage de la variation et d'aider l'expert dans la tâche de validation de l'indexation.

ABSTRACT. The extraction of relevant information in texts constitutes a fundamental process of text mining. We realize this process with a linguistic platform (ILC) for recognition and extraction of terms and their linguistic variants from corpora. We present a methodology to enhance the recognition of syntactic term variation in English using syntactic and morpho-syntactic features. Those modifications contribute to improve filtering of the variants and to assist the expert in validating indexation.

MOTS CLÉS. Extraction d'information, traitement automatique des langues, reconnaissance terminologique, variation, syntaxe, morpho-syntaxe, indexation, fouille de texte.

KEYWORDS. Information extraction, natural language processing, term recognition, variation, syntax, morpho-syntax, indexation, text mining.

Actes de CIFT 2004, pages 89 à 101

1 Introduction

Il est maintenant communément admis que les termes représentent l'expression linguistique de concepts et forment des indicateurs privilégiés de la connaissance portée par les documents. Cet ensemble de termes constituent les données d'entrée des autres niveaux d'analyse basés sur des méthodes symboliques ou statistiques (Cherfi et Toussaint, 2002 ; Polanco et al., 2000). Nous proposons d'améliorer la procédure d'indexation automatique réalisée à partir de la plate-forme d'ingénierie linguistique ILC qui reconnaît et extrait les termes d'un corpus à partir de ressources terminologiques de référence. Cette étude s'inscrit dans un processus de fouille de données textuelles appliqué à la génomique et la protéomique du cancer de la thyroïde[1]. Il s'agit d'aider l'expert à extraire et analyser l'information pertinente contenue dans les textes spécialisés grâce à des procédures automatiques.

Extraire et représenter la connaissance contenue dans les textes à partir d'une approche par reconnaissance terminologique permet d'obtenir une indexation plus homogène qu'une indexation non contrôlée. Cette qualité se répercute sur les traitements postérieurs tels que la classification et la cartographie. En contrepartie, une telle approche se heurte au problème de la variabilité des termes dans les sous-langues. En partant de notre expérience dans le domaine de l'indexation automatique contrôlée, nous allons montrer comment il est possible d'améliorer la reconnaissance de la variation terminologique. Ces variantes sont extraites par l'outil FASTR (Jacquemin, 1994) qui a été intégré dans la plate-forme d'ingénierie linguistique ILC (section 3). L'analyse des données issues de la reconnaissance des variantes terminologiques lors de cette procédure initiale a révélé un manque de précision des règles qui gèrent les transformations linguistiques (section 5). Ces règles spécifiques, appelées « métarègles », engendrent des erreurs récurrentes liées à des problèmes de dépendance non préservée et nécessitent une validation humaine de toutes les transformations produites. Nous montrons (section 6) comment exploiter des critères syntaxiques (propriété syntaxique de la variante) et morpho-syntaxiques (flexion et catégories grammaticales) afin d'améliorer la reconnaissance de la variation syntaxique. Ces critères interviennent à la fois à l'intérieur de la variation, comme contraintes filtrantes, et à l'extérieur, comme contraintes bornantes, afin de préserver les dépendances et d'assurer la correcte délimitation de la transformée en corpus. L'originalité de notre approche réside dans l'utilisation de ces critères bornants et se distingue en ceci de celle proposée par (Fabre et Jacquemin, 2000) qui exploite des critères dont la portée est limitée à la séquence correspondant à la variation. Nous cherchons également à assister l'expert dans sa tâche de validation et à réduire au maximum son intervention dans le processus en

[1] Projet scientifique soutenu par l'INSERM dans le cadre de l'Appel d'offre 2000 « Bioinformatique Inter-EPST » (CNRS, INRA, INRIA, INSERM).

augmentant la part d'automatisation à ce niveau. Les règles que nous développons se veulent indépendantes du domaine et applicables à tout terme linguistiquement motivé. Une réindexation de nos données initiales à l'aide des métarègles modifiées (section 7) apporte des résultats très positifs concernant la pertinence des critères utilisés et la faisabilité d'une telle approche avec une précision globale de 0.88.

2 Le point de vue de la variation terminologique : état de l'art

Les termes existent en dehors des fiches terminologiques ou des lexiques d'indexation. Ils subissent toutes les contraintes de la langue dans les textes et font preuve d'une très grande flexibilité qui les fait osciller entre variation et figement, entre lexicalisation et compositionnalité. G. Gross (1988) a montré dans la langue courante que les mots composés tels que *carte bancaire* étaient peu sensibles à la variation. Wagner (1991) étend cette observation aux domaines de spécialité. Il fait remarquer, à partir de lexiques français de termes économiques, qu'environ 75% d'entre eux sont des groupes nominaux partiellement ou totalement figés. Pourtant, Barkema (1994) s'est intéressé à donner une mesure de leur flexibilité en comparant les fréquences attendues d'occurrence telle que *cold war* par rapport aux fréquences observées de séquences libres « Nom Adjectif ». D'autres travaux, (Polanco et al., 1995 ; Royauté et al., 1996 ; Royauté, 1999) ont cherché à quantifier ce phénomène comme indicateur de l'activité scientifique. La variation révélerait une expression non encore stabilisée dans son usage pouvant être interprétée comme l'indice d'une activité scientifique. Dans une perspective similaire, (Ibekwe-SanJuan et SanJuan, 2002) exploitent ce phénomène pour parvenir à une classification des termes d'indexation et des documents associés. Depuis les travaux fondateurs de Spark-Jones et Tait (1984) en recherche d'information, la variation terminologique est passée du statut d'épiphénomène au statut de problématique linguistique à part entière (Habert et Jacquemin, 1993). La création d'outils linguistiques appropriés (Jacquemin, 1994 ; Daille, 1996) a permis de mesurer son importance et d'en faire l'étude en corpus (Jacquemin et Royauté, 1994). Ces outils ont été intégrés dans des applications terminologiques ou de recherche d'information (Aronson, 2001), et ont fait émerger des phénomènes complexes comme la variation morpho-dérivationnelle (Jacquemin et Tzoukermann, 1999) ou sémantique (Fabre et Jacquemin, 2000). La variation terminologique s'est révélée féconde dans l'acquisition de nouveaux termes (Jacquemin, 1999), la structuration de terminologies (Schmidt-Wigger, 1999), leur gestion (Carl et al., 2002) ou encore la construction d'ontologies (Daille, 2003). Dans les systèmes questions/réponses (Rinaldi et al., 2002), le repérage des variantes représente une aide précieuse pour apparier correctement une question à une phrase ou un segment de texte. Enfin en fouille de textes, (Cherfi et Toussaint, 2002) utilisent le logiciel FASTR afin de représenter le contenu de leur corpus d'étude à partir d'un réseau de termes extraits des textes pour identifier, dans un second temps, des règles d'association.

3 ILC : une plate-forme d'indexation contrôlée pour la fouille de texte

L'une des principales étapes en fouille de texte correspond à la collecte de documents et la représentation du sens qu'ils véhiculent à partir d'un ensemble de termes extraits de ces textes. Extraire les termes à partir d'une approche par reconnaissance, telle que celle développée dans ILC – Infométrie, Langues et Connaissances (Royauté, 1999), permet d'obtenir une représentation cohérente et homogène du corpus. ILC constitue un environnement ouvert pour une indexation contrôlée de textes français ou anglais. Il intègre des outils de traitement de la langue et des ressources linguistiques pour la reconnaissance des termes et de leurs variantes en corpus. Les traitements terminologiques impliquent que les termes soient étiquetés grammaticalement et lemmatisés. Cette phase, réalisée avec le TreeTagger[2], permet de transformer les termes en règles, selon le formalisme PATR-II de l'analyseur FASTR (Jacquemin, 1994; Jacquemin, 1999), pour lesquelles les informations morpho-dérivationnelles sont extraites de la base CELEX[3]. Le corpus doit subir une transformation similaire où chaque mot est étiqueté puis transformé en PATR-II. L'indexation porte ainsi sur ces deux types de données transformées : lexiques et corpus textuel. Un ensemble de règles particulières, nommées métarègles, permet à l'analyseur de reconnaître les variantes des termes en corpus. Les traitements permettent d'indexer les documents à partir de termes appartenant à ces ressources (Jacquemin et al., 2002; Daille et al., 2000).

Trois types de variations terminologiques sont exploitées : la variation flexionnelle, la variation syntaxique d'insertion (neural tissue ← neural crest derived tissues), de permutation (metabolism studies ← studies of iodine metabolism), et de coordination (residual tumor ← residual, recurrent or metastatic tumors) et, enfin, la variation morpho-dérivationnelle (hormone production ← produce some others hormones).

Les transformations linguistiques interviennent sur des termes composés de deux ou trois unités (« tumor cells », « thyroid function test », « cell of bone »). Ainsi, la métarègle de coordination *X2 N3 → X2 PUNC4 < {A|N|Np|V} PUNC? > C5 < {A|N|Np|V} > N3* permet de reconnaître en corpus la variante *residual, recurrent or metastatic tumors* à partir du terme « residual tumor ». Cette métarègle établit une équivalence entre un terme composé de deux unités lexicales X2 et N3, appartenant, respectivement, à n'importe quelle partie du discours (symbolisée par

---

[2] Le TreeTagger à été développé par Helmut Schmid à l'Université de Stuttgart. (http://www.ims.uni-stuttgart.de/~schmid/)

[3] CELEX est une base de données lexicales conçue par le « Centre for Lexical Information, Max Plank Institute for Psycholinguistics, Nijmegen.» (http://www.kun.nl/celex/). Pour le français, faute de disposer d'une ressource terminologique équivalente, des outils de morphologie robuste fondés sur des suffixes appropriés ont été développés pour la reconnaissance de dérivés possibles.

X) et à la classe des substantifs (symbolisée par N) et une transformée de ce terme formée du mot X2, d'une ponctuation (PUNC), de la présence d'un adjectif (A), nom (N), nom propre (NP) ou (représenté par une barre oblique) d'un verbe (V), suivi optionnellement (?) d'une seconde ponctuation, puis une coordination (C) suivie de l'insertion d'un adjectif, nom ou verbe avant le nom N3.

Les métarègles sont très permissives afin de privilégier le rappel sur la précision. Ce manque de précision est amplifié par le fait qu'elles opèrent dans une fenêtre de quelques mots où toutes les relations de dépendance ne peuvent être présentes.

4 Corpus d'analyse et données initiales

Ce travail s'appuie sur les résultats d'une indexation menée dans le cadre d'un processus de fouille de données textuelles appliqué à la génomique du cancer de la thyroïde (Royauté et al., 2004). Le corpus textuel représente 6253 données bibliographiques issues de la base Medline. La procédure d'extraction terminologique menée à l'aide d'ILC a porté sur les champs textuels des titres et des résumés en langue anglaise. Elle a été réalisée à partir d'une ressource terminologique composée de 360281 termes provenant de l'UMLS[4] (256290 préférentiels et 103991 synonymes, composés de deux mots et plus) et susceptibles de subir les variations linguistiques du langage naturel. 26108 séquences textuelles ou *ST* (réalisations des termes en corpus) ont été reconnues par ILC dont plus de la moitié (52%) proviennent d'une variation dont 10007 d'une variation syntaxique. Une phase de validation suit l'extraction où l'expert a accepté 70 % des termes issus d'une variation. Nous n'avons retenu pour cette étude que les termes récupérés à partir d'une variation syntaxique[5].

5 Variation : discussion et typologie des erreurs

Une variation est incorrecte lorsqu'elle modifie les relations initialement entretenues par les unités constituant le terme. Les termes renvoient à des groupes nominaux plus ou moins complexes exprimant des dépendances entre une tête et son expansion. Deux phénomènes en sont à l'origine. Le premier, externe à la séquence,

[4] L'UMLS est un projet de l'U.S. Department of Health and Human Services, National Institutes of Health (NIH) – National Library of Medicine (NLM).

[5] Les mauvaises variations morpho-dérivationnelles sont, pour une grande part, imputables à la base lexicale CELEX à partir de laquelle FASTR extrait les informations dont il a besoin. CELEX repose sur la notion de famille morphologique au sens large. Le passage d'un dérivé à un autre altère souvent le sens de départ (Human body ⇜ human cd44 monoclonal antibody).

correspond à une mauvaise délimitation de cette dernière en corpus. Le second est interne à la variation et concerne l'insertion d'unités grammaticale(s) ou lexicale(s) dans le syntagme qui éclate la cohésion syntaxique initiale (Ville-Ometz & al., 2004).

5.1 *Erreur de délimitation de la séquence en corpus*

Toutes les métarègles utilisées ici reposent sur un même principe : les éléments qui délimitent les frontières du terme bornent l'expression renvoyant à la variation. Lorsque le système reconnaît en corpus le motif décrit par la métarègle, il extrait la ST correspondante et renvoie au terme de référence. Rien ne permet d'assurer une correspondance syntaxique stricte entre la ST et le SN qu'elle est censée parfaitement recouvrir en corpus. Ces erreurs de délimitation touchent les contextes gauches et droits (en italiques dans les exemples ci-dessous) :

- le nom tête du terme initial et qui définit la tête apparente de la ST après variation (*bone* en (1)) constitue réellement dans le texte l'expansion d'une autre tête (*marrow*) :

  (1) skeletal bone ⇜ skeletal survey and bone *marrow examination*

- l'expansion du GN initial et de la ST après variation (*primary*) constitue en corpus l'expansion d'une autre tête (*myxedema*) :

  (2) primary hypothyroidism ⇜ hypothyroidism (primary *myxedema)*

- l'expansion dans le GN initial et dans la ST après variation (*thymus*) se révèle être, en corpus, la tête d'une autre dépendance (*rat*) constituant ainsi une nouvelle unité syntagmatique indépendante de la tête initiale (*gland*) :

  (3) thymus gland ⇜ *rat* thymus and adrenal gland

5.2 *Dépendances modifiées par insertion d'unités grammaticales ou lexicales*

L'insertion d'unités lexicales ou grammaticales est également susceptible de provoquer des ruptures syntaxiques tout à fait manifestes pour un être humain. Nous cherchons à formaliser ces phénomènes linguistiques afin que le système soit capable de filtrer automatiquement les mauvaises variations (4 et 5) tout en préservant les bonnes variations (6 et 7) :

(4) breast tissue ⇜ breast lesions or normal tissues
(5) temperature receptor ⇜ temperature dependent and receptor
(6) dividing cell ← dividing follicular and stroma cells
(7) physical development ← physical growth and development.

6 Traitements locaux appropriés pour une meilleure reconnaissance de la variation en corpus

6.1 *Modification des métarègles*

Nous proposons d'intervenir, d'une part, au niveau même de la ST pour empêcher que l'insertion de certaines unités lexicales ou grammaticales disloque les dépendances et, d'autre part, au niveau du contexte syntaxique externe à la séquence pour s'assurer de sa correcte délimitation en corpus. Quatre types de critères linguistiques sont exploités : la structure syntaxique de la variation produite, la catégorie grammaticale à laquelle appartient le ou les unités introduites lors de la transformation, la marque du pluriel sur le ou les noms têtes (lors d'une coordination) et enfin, la catégorie grammaticale des unités appartenant aux contextes gauche et droit de la séquence en corpus. Ce dernier critère va permettre de spécifier et d'introduire dans la métarègle des éléments linguistiques bornant le segment correspondant à la variation. Selon le principe de l'analyse syntaxique en chunks, nous utilisons certaines parties du discours – ici les mots outils et les verbes – ainsi que des séparateurs (tels que la ponctuation) pour définir des frontières de syntagme nominal et assurer un meilleur découpage syntaxique du segment lors de l'extraction terminologique. En ceci, notre démarche diffère de celle de (Fabre et Jacquemin, 2000) : nous ne nous contentons pas de définir des contraintes internes pour filtrer les mauvaises variations. De plus, dans le cas des variations syntaxiques, l'utilisation de critères bornants permet de faire l'économie de critères sémantiques qui se révèlent plus laborieux à manipuler car ils requièrent, notamment, une annotation sémantique du lexique.

L'introduction de ces critères linguistiques est illustrée à partir de deux métarègles différentes. Les critères ne s'appliquent pas uniformément à toutes les métarègles quelque soit leur type entraînant ainsi différents degrés de confiance que l'on peut accorder aux variations produites.

6.1.1 *Métarègle de coordination*

La métarègle de coordination sur les têtes, initialement formalisée à partir de l'expression

X2 N3 → X2 < {A|N|Np|V} 1-3 PUNC/',' ? > C4 < {A|N|Np|V} ? > N3

engendre différentes structures de variantes :

(8) Breast tissue ⇜ breast lesions or normal tissues
(9) Tumor cells ⇜ tumor patterns or cell
(10) Dividing cell ← dividing follicular and stroma cells
(11) Endocrine cel ← endocrine tissues and cells.
(12) ? Thyroid vein ← thyroid artery and vein.

Certaines de ces variations sont mauvaises (11 à 13), d'autres correctes (14 et 15) ou encore ambiguës (16), particulièrement pour une machine. Dans ce dernier cas les informations syntaxiques fournies par la séquence ne permettent pas de déterminer la validité de la transformation.

La métarègle a été modifiée et éclatée en trois métarègles distinctes. En plus de la syntaxe, les modifications concernent les contraintes liées à la métarègle par l'insertion d'une partie ou de l'ensemble des critères linguistiques définis précédemment.

1ère modification : utilisation du critère morphologique d'appartenance à une catégorie lexicale spécifique.

X2 N3 → X2 < {A|N|Np|V} 0-2 > A4 C5 < {A|N|Np|V} > N3 X6
avec : < X6 cat > ! A, N et Np ; <X6 lem> ! 'of' [6]. Cf. exemple 10.

Seule l'insertion d'un adjectif est autorisée à gauche de la coordination C5. Cette structure implique obligatoirement que A4 soit un modifieur de N3 et autorise ainsi l'introduction de toute unité lexicale à droite de C5 sans altérer les dépendances. La correcte délimitation de la ST en corpus est garantie par la présence d'une unité (X5) qui soit susceptible de marquer la frontière du groupe nominal auquel renvoie la séquence. La délimitation du contexte gauche n'est pas indispensable dans une telle structure où X2 est obligatoirement rattaché à la tête définie par N3.

2ème modification : utilisation des critères morphologiques de flexion et d'appartenance à une catégorie lexicale spécifique.

X2 N3 → X2 < {A|N|Np|V} 0-2 > N4 C5 N3
avec < N3 agr num > = plu ; < N4 agr num > = plu. Cf. exemple 11.

Lorsque l'unité lexicale introduite à gauche de C5 renvoie à un substantif il faut interdire l'insertion de toute unité à droite de la coordination ((11) Breast tissue ~~✓~~ breast lesions or normal tissues). L'ambiguïté liée à la délimitation de la ST (contexte droit) peut être levée en introduisant un critère flexionnel. La marque du pluriel sur N4 et N3 indique l'appartenance des deux substantifs au même SN et une coordination sur les noms têtes.

3ème modification : utilisation des critères morphologiques de flexion et d'appartenance à une catégorie lexicale spécifique.

---

[6] « agr num » pour *agreement number,*; « cat » pour *category*; « lem » pour *lemma*. Le symbole « = » exprime l'égalité entre les éléments, l'inégalité étant formalisée par « ! ». L'expression « n-m » indique que le ou les items la précédant doivent apparaître au minimum n fois et au maximum m fois dans la forme variante (respectivement 0 et 2 dans cette métarègle).

X2 N3 → X2 < {A|N|Np|V} 0-2 > N4 C5 N3 X6
avec : < N3 agr num > ! plu ; < N4 agr num > ! plu et < X6 cat > ! A, N, Np ; <X6 lem> ! 'of'.

(13) Thyroid function ← thyroid artery and vein are.
(14) Nuclear Transport ↚ nuclear compartmentalization and transport of

Lorsque N3 et N4 portent la marque du singulier, il faut s'assurer que N3 définisse bien la limite droite du syntagme en introduisant le critère d'appartenance à une catégorie grammaticale appliqué à X5 dans l'expression.

6.1.2 *Métarègle d'insertion*

Il est apparu plus problématique d'appliquer nos critères internes sur les métarègles d'insertion. C'est notamment le cas de la métarègle initiale X2 N3 → X2 < {A|N|Np|V} 0-3 > N3 pour laquelle nous avons décidé de n'intervenir que sur la borne droite de la variante. Dans une telle structure syntaxique, avec juxtaposition de modifieurs à gauche du nom tête N3, le modifieur initial X2 peut être précédé d'un autre modifieur sans que cela altère les relations de dépendance entre N3 et X2.

1ère modification : utilisation du critère morphologique d'appartenance à une catégorie lexicale spécifique.

X2 N3 → X2 < {A|N|Np|V} 0-3 > N3 X4
avec X4 n'appartenant pas à la catégorie des noms, des adjectifs et des noms propres (< X4 cat > ! A, N, Np).

(15) Cultured Cell ← cultured neoplastic human thyroid cells and
(16) Cultured Cell ← cultured peripheral blood mononuclear cells (
(17) DNA analysis ← dna image cytometric analysis of

Cette métarègle est simplement bornante mais non filtrante. On peut donc considérer a priori qu'elle présente un niveau de confiance inférieur à celui des métarègles de coordination précédentes.

De plus, nous n'avons accepté l'insertion d'un verbe qu'immédiatement après X2 comme suit :

2ème modification : utilisation du critère morphologique d'appartenance à une catégorie lexicale spécifique et élargissement du segment au contexte gauche.

X2 N3 → X5 X2 V4 < {A|N|Np} 0-2 > N3 X6
avec X6 n'appartenant pas à la catégorie des noms, des adjectifs et des noms propres < X4 cat > ! A, N et Np) et aucune contrainte sur X5.

(18) Cell differentiation ↚ cell line exhibiting differentiation
(19) Thyroid tumor ← the thyroid are unrelated malignant tumors.

Ces modifications permettent d'éliminer les variations résultant d'un mauvais découpage syntaxique à droite du segment. Toutefois, X5 n'a été introduit que pour élargir la fenêtre et faciliter ainsi la prise de décision. Cette métarègle se montre moins filtrante que la précédente.

7 Réindexation à partir des nouvelles métarègles : observation et discussion

Nous avons opéré une nouvelle indexation sur notre corpus initial à partir des métarègles modifiées. Afin d'évaluer ces résultats, un échantillon de 100 ST a été constitué de manière aléatoire pour chaque type de métarègles. Les ST issues des métarègles les moins productives ($10 \leq n \leq 100$ séquences) ont été examinées dans leur ensemble. On obtient ainsi un échantillon total de 1591 ST. En revanche nous n'avons pas tenu compte de certaines métarègles, principalement des coordinations, dont la productivité était trop faible (moins de 10 ST) pour que l'évaluation de leur précision soit réellement significative. Nous indiquerons simplement que sur les six métarègles concernées, seule l'une d'entre elles (une métarègle de coordination) a donné lieu à une mauvaise variation. Concernant les données obtenues en sortie de l'indexation, nous noterons que les insertions se montrent les plus productives avec 4908 ST retrouvées, suivies des permutations – 3435 – puis des coordinations – 1016, soit un total de 9359 ST.

Les modifications apportées aux métarègles ont permis d'obtenir une précision totale de 0.88 soit un gain de près de 20 % par rapport à la métagrammaire initiale. Une vue d'ensemble des trois principales catégories n'indique pas de différence notable en termes de précision puisque l'on obtient respectivement 0.9, 0.87 et 0.83 pour les coordinations, les permutations et les insertions. Les coordinations atteignent le meilleur score, comme nous l'avions prévu en section 6, étant donné qu'elles se sont révélées les plus propices à l'introduction des critères bornants et filtrans ; les insertions se présentant comme les moins favorables à de telles modifications.

Les disparités apparaissent clairement à l'intérieur même de chacune des trois catégories de coordination, de permutation et d'insertion. Ainsi, les quatre métarègles de coordination les plus productives présentent une précision qui s'élève, respectivement, à 0.93, 0,94, 0.99 et 1. Dans cette catégorie, les scores les plus bas s'échelonnent entre 0.64 et 0.77 et correspondent à des transformations peu productives. On retrouve une distribution similaire concernant la permutation où la moitié des transformations a une précision supérieure ou égale à 0.97. Les transformations les plus permissives se situent entre 0.66 et 0.87. La catégorie des insertions présente un profil légèrement inversé puisque la métarègle la plus productive (près de 89 % des séquences textuelles ramenées) atteint une précision moins élevée – 0.84 – au regard des deux autres catégories (précisions supérieures à

0.9)[7]. Par rapport aux trois nouvelles métarègles de coordination présentées en section 6.1.1., nous obtenons les scores suivants : 0.77, 0.92 et 0.94. Les résultats concernant ce premier score sont assez décevants ; des erreurs de sélection des termes avant indexation avec ILC, qui ne peut être appliqué que sur des termes motivés linguistiquement, ont perturbé les résultats. Enfin, de manière générale, les transformations syntaxiques les moins précises correspondent à des métarègles peu filtrantes qui ont été préservées afin de ne pas pénaliser le rappel sur la précision.

8 Conclusion et perspectives

La pertinence de nos critères syntaxiques et morpho-syntaxiques pour améliorer la reconnaissance de la variation syntaxique et aider l'expert dans sa tâche de validation est confirmée par les résultats obtenus en termes de précision. Toutefois l'exploitation de ces critères ne peut s'effectuer de manière homogène et uniforme sur l'ensemble de la métagrammaire mais elle dépend étroitement des propriétés linguistiques des structures mises en jeu. Un grand nombre de métarègles, dont les plus productives, sont maintenant en mesure de préserver les relations de dépendance initiales. Nous avons vu que certaines métarègles présentent des chiffres de précision très élevés – supérieures ou égales à 0.9. Concernant ces métarègles montrant un haut degré de confiance, on pourrait se permettre, en fonction des objectifs applicatifs liés au processus d'indexation, de faire l'économie sur une validation humaine qui se révèle très coûteuse en temps. Partant de ce principe, l'expert pourrait ne plus intervenir sur la totalité des séquences extraites mais se concentrerait uniquement sur celles dont le degré de confiance est plus faible en raison d'une impossibilité d'appliquer à la fois les critères filtrants et bornants.

La qualité des métarègles par rapport à la délimitation des ST montre également que l'exploitation d'éléments frontières au niveau même des métarègles est suffisante pour garantir une correcte délimitation de nos segments. Ceci permet de faire l'économie dans l'ensemble du processus d'une véritable analyse syntaxique partielle (grammaire des dépendances ou chunking) qui présenterait l'inconvénient d'alourdir les traitements sans toutefois être en mesure de garantir une analyse fiable.

Suite à ces résultats, nous envisageons d'étudier la manière de transposer nos critères bornants et filtrants à la reconnaissance de la variation morpho-dérivationnelle. Nous étudierons également la manière de combiner notre approche à celle de (Fabre et Jacquemin, 2000). Enfin, nous testerons la métagrammaire sur

[7] Nous attendions une précision plus élevée, toutefois nous nous sommes aperçus d'une erreur d'encodage de la métarègle dans la métarègrammaire qui a empêché son bon fonctionnement alors qu'elle était considérée comme très filtrante.

un jeu de données provenant d'un autre domaine scientifique et technique afin de vérifier la généricité de nos règles.

Bibliographie

Aronson A. R., « Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program », *Proceedings of the American Medical Informatics Association Symposium (AMIA 2001)*, 2001, pp.17-21.

Barkema H., « Determining the syntactic flexibility of idioms », in U. Fries, G. Totties et Schneider P., (eds), *Creating and using English language corpora*, Rodopi, Amsterdam, 1994, p. 39-52.

Carl, M., Haller J., Horschmann C., Maas D., Schütz J., « The TETRIS Terminology Tool », *Traitement Automatique des Langues*, Vol.43 : 1, 2002, p.73-102.

Cherfi H. et Toussaint Y., « Adéquation d'indices statistiques à l'interprétation de règles d'association », In P. Sébillot et A. Morin (eds), *Actes des 6èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 2002)*, Saint-Malo, France, 2002, p. 233-244.

Daille B., « ACABIT : une maquette d'aide à la construction automatique de banques terminologiques monolingues ou bilingues », in A. Clas, P. Thoiron and H. Béjoint (eds), *Lexicomatique et Dictionnairiques*, FMA, Beyrouth, 1996, p.123-136.

Daille B., « Conceptual structuring through term variations », in F. Bond, A. Korhonen, D. MacCarthy and A. Villacicencio (eds.), *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 2003, p.9-16.

Daille B, Royauté J. et Polanco X, « Evaluation d'une plate-forme d'indexation de termes complexes", *Traitement Automatique des Langues* , Vol.41 : 2, 2000. p. 396-422.

Fabre C. and Jacquemin, C., « Boosting Variant Recognition with Light Semantics », *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000*), Luxemburg, 2000, p.264-270.

Gross G., « Degré de figement des noms composés », *Langage*, Vol. 90, 1988, p.57-72.

Habert B. et Jacquemin C., « Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques », *Traitement automatique des langues*, Vol.34 : 2, 1993, p.5-42.

Ibekwe-SanJuan F., SanJuan E., « From term variants to research topics », *Journal of Knowledge Organization (ISKO), Special Issue on Human Language Technology*, Vol.29 : 3/4, 2002, p.181-197.

Jacquemin C., « FASTR : A unification-based front-end to automatic indexing », *Proceedings of the Intelligent Multimedia Information Retrival Systems and Management (RIAO'94*), New-York. Paris : CID, 1994, p.34-37.

Jacquemin, C., « Recognition and acquisition: two inter-related activities in corpus-based term extraction », *Terminology*, Vol. 4 : 2, 1999, p. 245-274.

Jacquemin, C., and Tzoukermann, E., « NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax », In T. Strzalkowski, (ed.), *Natural Language Information Retrieval*, Kluwer, Boston, MA, 1999, p. 25-74.

Jacquemin C. et Royauté J., « Retrieving terms and their variants in a lexicalized unification-based framework », *Proceedings of the 17th Annual International ACP SIGIR Conference on Research in Informatioin Retrieval (SIGIR'94),* Dublin, Juillet, Springer Verlag, 1994, p.132-141.

Polanco X., François C., « Data Clustering and Cluster Mapping or Visualization in Text Processing and Mining », *Proceedings of the 6th International ISKO Conference*, Toronto, Canada, *Advances in Knowledge Organization*, Vol. 7, 2000, p. 359-365.

Polanco X., Grivel L., Royauté J., « How To Do Things with Terms in Infometrics : Terminological Variation and Stabilization as Science Watch Indicators », in Koening and A. Bookstein (eds), *Proceedings of the 5th International Conference of the International Society for Scientometrics and Informetrics*, M.E.D, Medford, USA : Learned Information Inc., 1995, p. 435-444.

Rinaldi F., Dowdall J., Hess M., Kaljurand K., Koit M. and Kahusk N., « Terminology as Knowledge in Answer Extraction », *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE-2002),* Nancy, 2002, p. 107-112

Royauté J., Muller C., Polanco X., « Une approche linguistique infométrique de la variation terminologique pour l'analyse de l'information », *Actes du Colloque Informatique et Langage Naturel (ILN'96)*, Nantes, 1996, p. 563-581.

Royauté J., *Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information*, Thèse de doctorat en informatique, Université Henri Poincaré-Nancy I, 1999, 228 pages.

Royauté J., François C., Zasadzinski A., Besagni D., Dessen P., Maunoury M.T. et Le Minor S., « Relation entre gènes impliqués dans les cancers de la thyroïde », », *Revue des Nouvelles Technologies de l'Information (RNTI-E-2)*, Extraction et Gestion des Connaissances (EGC'2004), vol. 2, 2004, p 465-476.

Schmidt-Wigger A., « Building Consistent Terminologies », *COLING-ACL'98, COMPUTERM Workshop*, Montréal, 1999.

Spark-Jones K. et Tait J. I., « Automatic search term variant generation », *Journal of Documentation*, vol.40 , 1984, p.50-66.

Ville-Ometz F., Zasadzinski A, Besagni D., « Apport de l'analyse linguistique pour l'extraction terminologique en corpus : application au domaine de la génomique », *Actes des 3èmes Journées linguistiques de Corpus JLC'03*, Lorient, Presses Universitaire de Rennes, 2004.

Wagner Horst, « Dictionnaires, Bases de Données Lexicales & Lexicographie des Langue de Spécialité : Le traitement des Unités Complexes *», Actes du Colloque Informatique et Langue Naturelle (ILN'91),* Nantes, Décembre 1991, p. 389.1-389.10.