Meta
Journal des traducteurs
Translators' Journal

META

Une analyse terminométrique pour le repérage automatique des descripteurs complexes dans les textes de spécialité

Jacques Ladouceur et Patrick Drouin

Volume 42, numéro 1, mars 1997

Lexicologie et terminologie

URI : https://id.erudit.org/iderudit/003372ar
DOI : https://doi.org/10.7202/003372ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Ladouceur, J. & Drouin, P. (1997). Une analyse terminométrique pour le repérage automatique des descripteurs complexes dans les textes de spécialité. *Meta*, *42*(1), 207–218. https://doi.org/10.7202/003372ar

Résumé de l'article

On décrit une méthodologie pour le dépistage des descripteurs complexes (sous-ensemble des termes complexes) en langue de spécialité. On propose, dans un premier temps, un survol des approches ayant été utilisées pour le repérage des descripteurs en repérage d'informations et on présente ensuite une approche terminométrique du problème. L'analyse terminométrique se divise en quatre étapes principales d'analyse : l'analyse statistique, le filtrage des non-termes, le filtrage selon le degré de figement et l'analyse des candidats en contexte. On donne ensuite les résultats préliminaires du prototype NOTIONS.

UNE ANALYSE TERMINOMÉTRIQUE POUR LE REPÉRAGE AUTOMATIQUE DES DESCRIPTEURS COMPLEXES DANS LES TEXTES DE SPÉCIALITÉ

JACQUES LADOUCEUR et PATRICK DROUIN[1]
Université Laval, Sainte-Foy et Université de Montréal, Montréal, Canada

Résumé

On décrit une méthodologie pour le dépistage des descripteurs complexes (sous-ensemble des termes complexes) en langue de spécialité. On propose, dans un premier temps, un survol des approches ayant été utilisées pour le repérage des descripteurs en repérage d'informations et on présente ensuite une approche terminométrique du problème. L'analyse terminométrique se divise en quatre étapes principales d'analyse : l'analyse statistique, le filtrage des non-termes, le filtrage selon le degré de figement et l'analyse des candidats en contexte. On donne ensuite les résultats préliminaires du prototype NOTIONS.

Abstract

This article describes a methodology used for detecting compound descripters (subgroups of complex terms) in specialized language. It begins with an overview of current approaches to information retrieval and then discusses a terminometric approach consisting of four phases: statistical analysis, filtering of non-terms, filtering of set expressions according to their degree of fixedness, analysis of candidate-terms in context. It concludes with preliminary results obtained using the NOTIONS *prototype.*

1. INTRODUCTION

1.1. TERMINOLOGIE ET GESTION DOCUMENTAIRE

La parenté entre le travail de repérage automatique d'information dans le domaine de la gestion documentaire et celui de repérage de la terminologie est grande (voir Sager 1990). En gestion documentaire, notamment pour la constitution de bases de données bibliographiques, on élabore, pour représenter le contenu des documents, des listes de mots clés, ou de descripteurs. En terminologie, on cherche notamment à identifier l'ensemble des termes qui caractérisent un domaine donné, domaine souvent circonscrit par un corpus de textes. L'élaboration de listes de termes et de listes de descripteurs sont des opérations qui se ressemblent beaucoup, et pour cause : le descripteur est un terme.

Étant donné que les experts de la gestion documentaire ont effectué beaucoup de travail dans une optique de repérage informatique des descripteurs, les travaux du domaine possèdent une certaine avance sur ceux effectués en terminologie. Il est donc avantageux de tenter de réutiliser les techniques de dépistage automatique des descripteurs pour le dépistage automatique des termes.

1.2. TERMES ET DESCRIPTEURS

En gestion documentaire, le descripteur est défini comme un mot graphiquement simple ou complexe qui caractérise le contenu d'un document ou d'une partie de document qui correspond à une notion du domaine auquel appartient le document (ou la partie de

texte) qu'il caractérise. Le descripteur a une fréquence moyenne ; les mots les plus fréquents n'étant généralement pas porteurs d'information[2] et les mots les moins fréquents n'étant pas représentatifs de l'ensemble textuel décrit.

En terminologie, le descripteur correspond à un terme récurrent, un terme bien établi dans son domaine. La figure 1 présente le lexique qui se divise en unités simples et en unités complexes selon un axe vertical. Selon un axe horizontal, les unités se divisent en termes et non-termes. Cette vision simplifiée du lexique permet de bien saisir le lien de parenté qui existe entre le sous-ensemble des termes complexes et le sous-ensemble des descripteurs complexes. Les descripteurs font partie du sous-ensemble des termes et les descripteurs complexes (DC) font donc partie de l'ensemble des termes complexes. Les autres blocs de l'ensemble des unités complexes sont les segments[3] répétés qui ne feraient pas partie d'une description terminologique d'un texte. Le sous-ensemble des termes que nous cherchons à identifier dans le cadre de cette étude correspond à la portion de la figure 1 qui est hachurée verticalement.

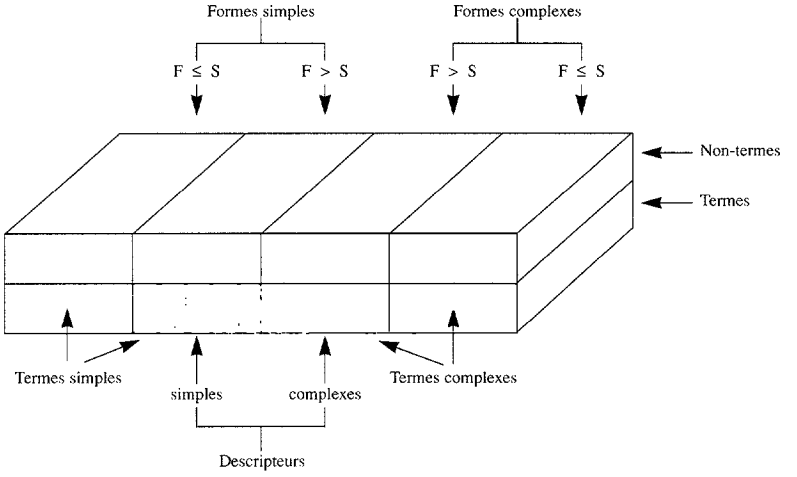

Figure 1 :
Structure lexicale

1.3. IDENTIFICATION DES DESCRIPTEURS

La problématique de l'identification automatique des descripteurs n'est pas la même selon que ceux-ci sont graphiquement simples ou complexes. De nombreux systèmes de gestion documentaire ou de repérage d'information permettent de reconnaître, avec un certain succès, les descripteurs graphiquement simples mais très peu, et ceux qui le font obtiennent des résultats médiocres, permettent de reconnaître les descripteurs graphiquement complexes.

Le descripteur est généralement identifié à partir de critères statistiques comme la fréquence et la répartition (voir Salton 1988). Le descripteur est défini comme un mot à contenu notionnel dont la fréquence ne se situe ni parmi les plus hautes, ni parmi les plus basses ; on dira ainsi que le mot k est un descripteur du texte i si :

(1) La fréquence de *k* dans *i* est supérieure à 1
$\text{FRÉQUENCE}_{ik} > 1$

(2) La fréquence de *k* dans *i* est inférieure ou égale au seuil fixé
$\text{FRÉQUENCE}_{ik} \leq \textit{Seuil}$

L'identification de cet ensemble de descripteurs permettra donc de recenser tous les termes dont l'emploi est bien stabilisé dans un corpus de textes représentatif d'un domaine.

On utilise également la fréquence pour caractériser l'importance des descripteurs. Plus la fréquence est élevée, plus le descripteur est important. D'ailleurs, en terminologie, il est généralement admis que la fréquence est l'indice statistique le plus productif pour le repérage des termes (Daille 1993 ; 1994). La fréquence d'occurrence est un bon indice du degré de terminologisation d'une forme dans un texte de spécialité.

La reconnaissance du descripteur s'appuie donc sur une reconnaissance préalable du terme. Un descripteur graphiquement simple est un terme graphiquement simple qui possède les caractéristiques statistiques que nous venons de décrire. La reconnaissance du mot simple ne pose pas de problème car ce dernier est formellement délimité par des blancs typographiques (ou autres séparateurs comme la ponctuation, les parenthèses, etc.). La distinction entre terme et mot s'effectue d'après des critères sémantiques et extralinguistiques qui ne peuvent être utilisés dans le cadre d'une analyse automatique, et nous devons donc chercher une autre solution. Il est possible de distinguer assez efficacement le terme simple du mot simple à l'aide d'un simple critère de fréquence si l'on procède à l'exclusion des mots grammaticaux par le biais de dictionnaires d'exclusion.

La problématique de reconnaissance du descripteur complexe réside, entre autres, sur le fait que la reconnaissance du mot complexe est extrêmement difficile autant d'un point de vue sémantique que syntaxique. L'utilisation d'un simple critère de fréquence, comme celui utilisé pour le repérage des descripteurs simples, conduit à l'identification de segments qui sont des enchaînements syntaxiques non nominaux fréquents (*il y a*), des segments nominaux qui ne sont pas des termes (*panier de pommes*) et des segments nominaux qui sont des termes (*pomme de terre*). Comme le montrent les deux derniers exemples, l'utilisation de critères syntaxiques ne permettrait pas non plus de faire la distinction entre des segments nominaux qui ne sont pas des termes et des segments nominaux qui sont des termes[4]. Une méthodologie d'identification des termes doit donc aller chercher ailleurs des moyens lui permettant de faire une distinction claire entre les deux types de segments.

2. UNE ANALYSE TERMINOMÉTRIQUE POUR L'IDENTIFICATION DU DESCRIPTEUR COMPLEXE

Si les techniques statistiques pures permettent de relever très facilement les segments récurrents, elles ne sont d'aucune utilité pour faire la distinction entre un segment qui ne correspond pas à une unité lexicale complexe (*uranium est*), une unité lexicale complexe qui n'est pas un terme (*a eu*) et un terme complexe (*particule* α).

Comme nous l'avons mentionné plus haut, l'identification automatique des termes à partir de patrons de formation syntagmatique est limitée par l'ampleur de l'intersection entre les termes et les enchaînements syntaxiques possédant une structure semblable (*panier de pommes, kilo de pommes*). De plus, la rigidité des patrons de formation (N_1 *de* N_2) ne permet pas de représenter les phénomènes de dynamique discursive d'insertion et de réduction auxquels sont parfois soumis les termes complexes en situation textuelle. Ce flou concernant la structure de termes rend difficile, voire impossible, la représentation formelle de la structure du terme selon des patrons de formation syntagmatique dans un système informatique. S'il est difficile de décrire formellement le terme à l'aide de règles,

il est beaucoup moins difficile de décrire ce qui assurément ne peut pas être un terme. L'approche que nous proposons fait appel à la statistique mais surtout à des descriptions formelles du *non-terme*[5].

La méthode que nous avons élaborée pour reconnaître le descripteur complexe comprend quatre étapes. Nous utilisons d'abord une analyse statistique pour élaborer une première liste brute de descripteurs complexes potentiels. Ensuite, nous appliquons deux filtres linguistiques afin d'éliminer le maximum de bruit possible :

a) filtrage des non-termes ;
b) filtrage des candidats selon le degré de figement.

Finalement, nous qualifions le potentiel des segments repérés à être des descripteurs à l'aide d'une analyse de leur contexte immédiat.

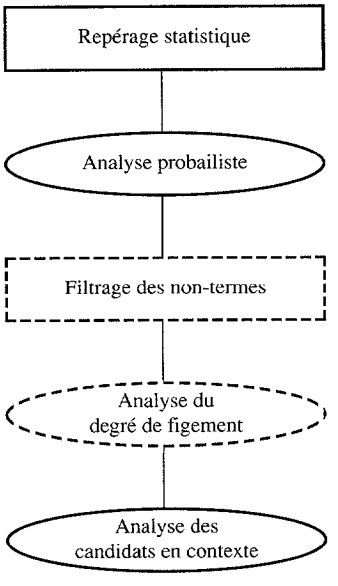

Figure 2 :
Étapes du traitement terminométrique

2.1. ANALYSE STATISTIQUE

Nous élaborons d'abord une liste de descripteurs complexes potentiels en utilisant la technique statistique décrite précédemment. Les critères utilisés pour l'identification des segments retenus sont les suivants :

(3) La fréquence observée du segment est supérieure à sa fréquence théorique
FRÉQ. OBSERVÉE(*Segment*) > FRÉQ. THÉORIQUE(*Segment*)
(4) La fréquence observée du segment est supérieure ou égale au seuil de pertinence
FRÉQ. OBSERVÉE ≥ *Seuil*
(5) Le segment ne contient pas plus de *x* éléments[6]
LONGUEUR(*Segment*) ≤ *Nombre d'éléments*

La liste présentée plus bas est un exemple des résultats d'un tel traitement statistique. On remarque que le résultat brut de l'analyse n'est pas directement satisfaisant et que les formes obtenues pourraient être filtrées à l'aide de règles linguistiques relativement simples.

d'un système d'information à
d'un système d'information à référence
d'un système d'information à référence spatiale
du système
du système d'information
du système de
du système de référence
le projet de système
le projet de système d'information
le projet de système d'information sur
système d'information
système d'information sur
système d'information sur les
système d'information sur les parcelles
système d'information traditionnel
système d'information un
système d'information à
système d'information à référence
système d'information à référence spatiale
système d'information à référence spatiale à

2.2. ANALYSE LINGUISTIQUE

Les résultats de l'analyse statistique sont très intéressants parce qu'ils permettent de repérer l'ensemble des formes complexes qui se produisent d'une façon non aléatoire[7]. Un grand nombre de candidats n'ont toutefois ni statut, ni intérêt linguistique. Afin d'éliminer ce bruit tout en conservant l'ensemble maximal des formes intéressantes pour le traitement linguistique ou notionnel, nous devons procéder à l'élaboration de stratégies linguistiques d'épuration.

2.2.1. Filtrage des non-termes

Comme nous l'avons mentionné précédemment, la description du terme est formellement difficile à réaliser car sa structure ne se laisse pas facilement maîtriser. À partir de cette constatation, nous avons opté pour une approche différente qui, selon nous, permet d'obtenir de meilleurs résultats et de conserver un plus grand nombre de bons candidats. Pour y arriver, nous avons élaboré une série de patrons syntagmatiques des non-termes[8] :

(6) **Pronom**
pronom ω
ω pronom
ω pronom φ

(7) **Préposition**
préposition ω
ω préposition

(8) **Article**
article ω
ω article

(9) **Conjonction**
conjonction ω
ω conjonction

(10) **Adverbe**
adverbe ω
ω adverbe

(11) **Verbe conjugué**
ω verbe conjugué
verbe conjugué ω
ω verbe conjugué φ
ω ~prép. verbe infinitif

Nous identifions ensuite les segments qui correspondent à l'un de ces patrons et les éliminons de notre liste :

~~d'un système d'information à~~
~~d'un système d'information à référence~~
~~d'un système d'information à référence spatiale~~
~~du système~~
~~du système d'information~~
~~du système de~~
~~du système de référence~~
~~le projet de système~~
~~système d'information sur~~
~~système d'information sur les~~
système d'information sur les parcelles
système d'information traditionnel
~~système d'information un~~
~~système d'information à~~
système d'information à référence
système d'information à référence spatiale

~~le projet de système d'information~~ ~~système d'information à référence spatiale à~~
~~le projet de système d'information sur~~ ~~système d'information à référence spatiale peut~~
système d'information

Le filtrage opéré par NOTIONS n'est que partiel et l'étape est interrompue dès que l'une des règles n'est pas respectée. Par exemple, dans le cas du dernier segment de la liste, *système d'information à référence spatiale peut*, le système utilisera la règle (11) pour l'éliminer de la liste ; dans le cas du segment *le projet de système d'information sur*, le système peut tout aussi bien appliquer la règle (8) ou la règle (9) pour filtrer le segment.

Nous travaillons ainsi à dépister et à éliminer le non-terme plutôt que le terme. Cette approche nous semble beaucoup plus simple pour le moment et elle a pour avantage d'éviter les problèmes de l'analyse syntaxique que nous avons décrits en introduction.

2.2.2. Filtrage des candidats selon le degré de figement

Diverses techniques statistiques ont été utilisées pour illustrer le lien qui existe entre deux unités lexicales ou plus dans un texte (voir Daille 1993 : 115-150). Nos tests démontrent que la simple comparaison des segments entre eux permet d'obtenir de bons résultats. La comparaison de la fréquence d'un segment avec celle d'une partie de ce segment (sous-segment) peut fournir des renseignements intéressants sur leur autonomie linguistique respective. Les règles qui suivent permettent d'éliminer des candidats retenus par les étapes de filtrage statistique et de filtrage des non-termes. Le principal problème de la liste obtenue, à l'étape, en est un de redondance et de recoupement entre les segments retenus. Les règles[9] utilisées pour résoudre ce problème sont les suivantes :

a) Si la fréquence d'un segment (*événements géologiques*) est la même que la fréquence d'un segment plus long qui le contient (*datation d'événements géologiques*), on peut affirmer que le segment plus court ne se produit jamais indépendamment du segment plus long et l'éliminer.

(12) Si $\text{FRÉQ}_{ab} = \text{FRÉQ}_{a}$ alors conserver seulement ab
ex. : ~~système d'information à référence~~ (FRÉQ. = 132)
système d'information à référence spatiale (FRÉQ. = 132)

b) Si la fréquence d'un segment (*événements géologiques*) est supérieure à la fréquence d'un segment plus long qui le contient (*datation d'événements géologiques*), on peut affirmer que le segment plus court se produit à certains moments indépendamment du segment plus long.

(13) Si $\text{FRÉQ}_{ab} < \text{FRÉQ}_{a}$ alors conserver ab et a
ex. : système d'information à référence spatiale (FRÉQ. = 132)
système d'information (FRÉQ. = 175)

L'inclusion constitue un indice valable sur le fonctionnement linguistique indépendant d'un segment, sur son degré de figement linguistique et sur son intérêt pour le repérage d'information notionnelle.

2.2.3. Analyse des contextes

Le filtrage par analyse contextuelle, contrairement à l'étape précédente, ne procède pas à l'élimination de candidats. Il s'agit d'une analyse sommaire du contexte du terme, à l'aide de grammaires locales (Silberztein 1993), qui vise à assigner aux candidats une pondération représentant la probabilité, qu'il s'agisse ou non d'un terme ou d'un descripteur. La notion de «probabilité» dont il est ici question ne correspond pas à la

notion mathématique de probabilité. Notre analyse prend en considération divers critères qui permettent de caractériser le comportement d'un terme, et la pondération obtenue sera d'autant plus élevée que le segment se comporte comme un descripteur complexe. La pondération est calculée à partir de diverses analyses locales qui prennent en considération, en plus de la probabilité d'enchaînement des segments côte à côte (distribution aléatoire) et du degré de figement des segments, le comportement textuel et graphique du segment. Les grammaires locales procéderont à une analyse du contexte précédant et suivant le mot pour vérifier si le segment répond aux règles d'analyse contextuelle. Par la suite, ces résultats sont fusionnés dans une cote globale avec la cote relative à l'aspect graphique du terme. La figure 3 illustre la mise en place de la cote globale qui représente le potentiel d'un segment à être un terme ou son caractère «terminogénique».

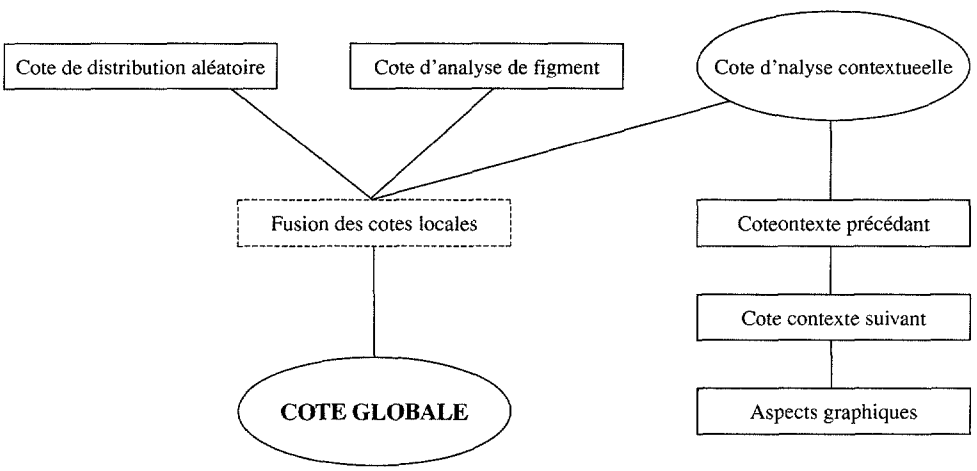

Figure 3 :
Calcul du caractère terminogénique d'un segment

L'élaboration des filtres et le poids de chacune des pondérations locales dans le calcul de la pondération globale (étape de fusion des cotes locales) ont été mis de l'avant à partir de l'observation de plusieurs milliers de contextes et de l'identification de phénomènes redondants. Cette technique, bien que hautement empirique, permet d'obtenir des résultats concluants. Notre technique d'élaboration de règles rejoint celle utilisée pour le logiciel LEXTER qui se fonde sur des règles de nature empirique qui n'ont pas nécessairement de fondement en linguistique théorique mais qui permettent l'obtention de résultats de bonne qualité (voir Bourigault 1992). Les règles, telles qu'elles sont présentées, ont un effet sur les cotes locales uniquement dans les cas où elles sont satisfaites. Si la condition n'est pas satisfaite, la cote locale n'est pas modifiée.

(14) aspects graphiques
Certains aspects graphiques du segment répété peuvent fournir une indication sur la validité d'un segment repéré. Par exemple, la présence d'une majuscule au début du segment et d'un séparateur fort (voir règle (16)) en fin de segment permettent de supposer que ce dernier correspond à une notion.
ex.: [...] *il est chef du Service «**S**ystème d'information» et responsable [...]*

(15) segment précédé d'un déterminant[10]
DÉT + ω
ex.: *[...] accéder aux données d'**un** système d'information [...]*

(16) segment précédé ou suivi d'un séparateur fort
SÉP + ω
ω + SÉP
ex.: [...] *envisager la mise en place d'un système d'information **?***

(17) segment suivi d'un verbe conjugué
ω + VERB
ex.: *Ce système d'information **s'étend** à la gestion [...]*

(18) segment suivi d'un pronom relatif
ω + PREL
ex.: [...] *grâce au système d'information **que** nous [...]*

3. RÉSULTATS PRÉLIMINAIRES

Pour évaluer le prototype, nous avons traité deux textes extraits de la collection «Que sais-je ?» des Presses Universitaires de France. Après avoir été numérisés, les textes ont été soumis à un indexeur humain qui a identifié les termes complexes et qui a ensuite extrait, à partir de critères statistiques, les descripteurs complexes. Les consignes qui ont été données à l'indexeur humain sont les mêmes que celles qu'utilise le prototype NOTIONS pour le repérage; c'est-à-dire les règles (3), (4) et (5) qui définissent le descripteur. L'indexeur devait donc commencer son travail en identifiant l'ensemble de termes de l'ouvrage analysé tout en identifiant les termes complexes au passage à l'aide du trait de soulignement (ex.: *système_d_information*). Les termes complexes ont par la suite été divisés en deux sous-ensembles selon qu'ils étaient des descripteurs ou non d'après un critère de fréquence.

Les mêmes textes ont été soumis au prototype NOTIONS et les résultats ont été comparés. L'évaluation de la performance du prototype peut s'effectuer à l'aide de divers indices mais nous avons décidé de nous en tenir aux indices les plus utilisés dans le domaine du repérage d'information: le taux de rappel (*R*) et la précision (*P*) (Salton 1988).

Le taux de rappel représente le potentiel d'un système à fournir l'ensemble de l'information pertinente. Les valeurs possibles de *R* oscillent entre 0 et 1, qui représentent respectivement un résultat nul et un résultat parfait. Ainsi, un système qui peut recenser 15 descripteurs pertinents sur une possibilité de 15 obtiendra un taux de rappel de 1,0. Par contre, il est aussi possible que le même système ne retienne pas que ces 15 descripteurs et que ces derniers soient noyés dans un taux de bruit très élevé[11]. C'est à ce niveau que le taux de précision entre en jeu.

Le taux de précision représente le potentiel d'un système à fournir uniquement l'information pertinente. Tout comme pour *R*, *P* peut prendre des valeurs entre 0 et 1, qui indiquent respectivement un résultat passable à un résultat parfait. Nous ne pouvons pas parler de résultat nul dans le cas de *P* car il est possible qu'un système puisse identifier 15 descripteurs sur une possibilité totale de 15 mais que le système identifie aussi 30 autres

segments comme étant des descripteurs. Nous pouvons difficilement qualifier ce résultat de «nul» étant donné qu'il est tout de même utilisable. En bref, nous pouvons dire que le taux de précision reflète la pureté des résultats obtenus.

Résultats	Test 1	
Communs à l'indexeur	49 (52)	
Carence	3	
Surplus	17	
Positif	**Oublis de l'indexeur**	13
Négatif	**Erreurs**	4

Tableau 1 :
Résultats du Test 1

Test 1	Pertinents	Non pertinents
Repérés	62	4
Non repérés	3	

Tableau 2 :
Tableau de contingence pour l'évaluation des résultats du Test 1

Pour le premier texte, le prototype a reconnu 94,23 % des descripteurs que l'indexeur humain avait identifiés. La systématicité du travail de la machine offre un avantage sur le travail de l'indexeur qui peut être influencé par divers facteurs propres au travail ou extérieurs à ce dernier. Les résultats de la machine, sans être plus précis[12], sont plus proches d'un taux de rappel maximal, car le prototype a identifié 13 descripteurs laissés de côté par l'humain. Les quatre descripteurs suggérés par la machine et qui étaient erronés sont sans grande importance car leur nombre est peu élevé. Pour récapituler, sur un total de 65 descripteurs complexes possibles, l'indexeur en a identifiés 49 (75,4 %), alors que le prototype a pu en relever 62 (95,4 %). Le taux de rappel pour le premier test est très satisfaisant car NOTIONS nous a permis d'identifier 95 % des descripteurs complexes tout en conservant un taux de précision élevé (0,94). Un tel taux de précision indique que les manipulations par le terminologue des données fournies par le prototype sont très faciles et qu'elles ne demandent que très peu de temps. L'information est disponible rapidement et n'est pas noyée dans une liste interminable de descripteurs *potentiels*.

À partir de la table de contingence des résultats pour le Test 1, nous pouvons calculer le taux de rappel (R_1) et (P_1) de la façon suivante :

(19) $$R_1 = \frac{\textit{N. d'items pertinents repérés}}{\textit{N. total d'items pertinents}} \qquad \frac{62}{(62+3)} = 0{,}95$$

(20) $$P_1 = \frac{\textit{N. d'items pertinents repérés}}{\textit{N. total d'items repérés}} \qquad \frac{62}{(62+4)} = 0{,}94$$

Pour le deuxième texte, le prototype a identifié les 58 descripteurs recensés par l'indexeur humain et en a proposé 7 qui n'avaient pas été retenus et qui auraient dû l'être par l'indexeur humain. Le taux de rappel et la précision sont calculés de la façon suivante :

(21) $$R_2 = \frac{65}{(65+0)} = 1{,}00$$

(22) $$P_2 = \frac{65}{(65+5)} = 0{,}93$$

Résultats	Test 2	
Communs à l'indexeur	58 (58)	
Carence	0	
Surplus	12	
Positif	**Oublis de l'indexeur**	7
Négatif	**Erreurs**	5

Tableau 3 :
Résultats du Test 2

Test 2	Pertinents	Non pertinents
Repérés	65	5
Non repérés	0	

Tableau 4 :
Tableau de contingence pour l'évaluation des résultats du Test 2

Dans ce cas précis, le prototype a su identifier l'ensemble des DC pertinents contenus dans le texte et obtient ainsi un taux de rappel parfait de 1,00 (l'ensemble des DC du texte ont été identifiés). Cette valeur de R est d'autant plus impressionnante qu'elle dépasse encore une fois la performance de l'humain. Pour ce qui est de la perti-

nence, elle demeure relativement stable (0,93) pour les deux tests effectués et prouve que les résultats du prototype peuvent être directement intégrés dans la démarche du terminologue qui ne sera pas appelé à perdre la plus grande partie de son temps à éliminer les erreurs de la liste de repérage.

4. CONCLUSION

Contrairement aux performances des systèmes de repérage utilisant une approche purement statistique ou linguistique, la majorité des DC recensés par notre prototype sont intéressants et pertinents pour le travail linguistique. L'épuration des résultats de l'analyse statistique à l'aide de techniques linguistiques permet d'obtenir une nette diminution du taux de bruit[13].

Les résultats que nous avons obtenus avec NOTIONS sont probants à 95 % et ne sont donc pas parfaits. Cependant, si l'on considère les résultats de l'indexeur humain, qui disposait malgré tout d'un environnement de travail informatisé pour le calcul des fréquences des termes complexes, les résultats sont d'autant plus intéressants.

En gardant en tête que les descripteurs complexes sont un sous-ensemble important des termes complexes[14], nos résultats sont directement utilisables dans une optique d'informatisation de la démarche terminologique. L'objectif de l'informatisation d'une démarche n'étant pas d'automatiser l'ensemble du processus de travail mais plutôt de cibler l'automatisation des tâches mécaniques et répétitives, nous croyons que le prototype pourrait procéder au premier dépouillement d'un texte de spécialité pour le terminologue. Les résultats obtenus peuvent alors être utilisés dans le cadre d'une démarche interactive qui peut tout aussi bien s'insérer dans un cadre de travail traditionnel que dans un cadre de travail informatisé (poste de travail du terminologue). Pour de nombreuses années encore, le recours au langagier dans le cadre d'une démarche automatique est inévitable, et nous espérons que le langagier saura aussi se rendre compte de l'aspect inévitable du recours à la machine pour une économie de temps et d'efforts.

Notes

1. Nous tenons à remercier le CRSH pour sa contribution à nos recherche de doctorat.
2. Il s'agit en général des mots grammaticaux.
3. Un segment est une suite de formes qui ne contient pas de séparateur fort (voir Lebart et Salem 1988).
4. L'utilisation d'une matrice de formation du type N de N conduit à l'identification de *panier de pommes* et de *panier d'osier* qui n'ont pas la même valeur terminologique.
5. Le *non-terme* est un enchaînement (dans le cas de notre recherche qui porte sur les formes complexes) lexical qui ne correspond pas à une notion.
6. La longueur maximale des segments est variable et paramétrable. Nos tests ont porté sur des segments dont la longueur se situait entre 2 et 6 éléments.
7. Nous dirons qu'une forme complexe a une répartition non aléatoire lorsque la fréquence observée dans le texte dépasse de façon significative la fréquence théorique que devrait avoir cette forme dans le texte.
8. Dans les patrons, les caractères ω et φ sont utilisés pour représenter des chaînes des caractères quelconques et le caractère ~ la négation. Certains patrons qui sont décrits ici peuvent à l'occasion correspondre à la structure d'un terme, mais ces cas sont marginaux et sans grand intérêt pour une démarche visant la généralisation.
9. Dans ces règles, *a* et *b* sont des sous-segments qui constituent le segment *ab*.
10. Dans les exemples qui suivent, le symbole ω est utilisé pour signaler la présence d'un segment répété.
11. Il s'agit ici du problème principal des systèmes procédant uniquement à l'aide de patrons syntagmatiques qui recensent l'ensemble des formes ayant une structure du type N_1 *de* N_2, par exemple.
12. Le taux de précision de l'humain est généralement de 1,0 et est très difficile à atteindre par un système automatique.
13. Le taux de bruit peut être calculé de la façon suivante : $B = (1\text{-}P)$. Les taux de bruit pour nos deux tests sont donc respectivement 0,06 et 0,07 sur une échelle variant entre 0 et 1.
14. Certains systèmes visant l'identification automatique des termes ne ciblent que les descripteurs sans le dire explicitement.

RÉFÉRENCES

AUGER, P. (1979) : «La syntagmatique terminologique, typologie des syntagmes et limite des modèles en structure complexe», *Table ronde sur les problèmes de découpage du terme*, V[e] Congrès de l'Association internationale de linguistique appliquée (AILA), Montréal, Office de la langue française, Éditeur officiel du Québec, pp. 9-26.

BOURIGAULT, D. (1992) : «Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases», *Proceedings of the Fourteenth International Conference on Computational Linguistics — COLING 92*, Nantes, pp. 977-981.

CADIOT, P., HABERT, B. et C. JACQUEMIN (1992) : *Compte rendu de la Journée Noms Composés*, 26 Juin 1992, École Normale Supérieure de Fontenay, Saint-Cloud, 25 pages.

CHOUEKA, Y. (1988) : «Looking for Needles in a Haystack or Locating Interesting Collocationnal Expressions in a Large Textual Database», *Actes de colloque du RIAO 88*, Cambridge, Cambridge University Press, pp. 609-623.

CHOUEKA, Y., KLEIN, T. et E. NEUWITZ (1983) : «Automatic Retrieval of Frequent Idiomatic and Collocationnal Expressions in a Large Corpus», *ALLC Journal*, Grande-Bretagne, vol. 4, n° 1, pp. 34-39.

DAILLE, B. (1993) : «Extraction automatique de terminologie monolingue», *Actes du colloque Informatique et langue naturelle*, Nantes, 21 pages.

DAILLE, B. (1994) : «Extraction de noms composés terminologiques du domaine des Télécommunications», *5[es] Journées ERLA-GLAT (Études et Recherches Lexicales Appliquées)*, Brest, 13 pages.

DAVID, S. (1990) : «Le progiciel TERMINO : de la nécessité d'une analyse morphosyntaxique pour le dépouillement des textes», *Actes du colloque Les industries de la langue : perspective des années 1990*, 21 au 24 novembre 1990, Montréal, Office de la langue française et Société des traducteurs du Québec, pp. 71-89.

DROUIN, P. et J. LADOUCEUR (1994) : «L'identification automatique des descripteurs dans des textes de spécialité», *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*, 2 et 3 décembre, Genève, ISSCO, pp. 18-28.

JACQUEMIN, C. (1994) : «FASTR : A Unification-Based Front-End to Automatic Indexing», *RIAO 94 Conference Proceedings: Intelligent Multimedia Information retrieval Systems and Management*, 11 au 13 octobre 1994, New York, Rockfeller University, vol. 1, pp. 34-47.

KOCOUREK, R. (1991) : *La langue française de la technique et de la science : vers une linguistique de la langue savante*, 2[e] édition, Oscar Brandsetter Verlag GMBH & Co. KG, Wiesbaden, 259 pages.

LADOUCEUR, J. et A. BRISSET (à paraître) : «L'ingénierie documentaire : perspectives pour la traduction et la terminologie», *Actes du Colloque Perspectives d'avenir en traduction*, Collège de Saint-Boniface, Saint-Boniface.

LADOUCEUR, J. et A. BRISSET (à paraître) : «Multilingual Text Information Management», *Actes du colloque RIAO 94 : Intelligent Multimedia Information Retrieval Systemes and Management*, Rockefeller University, New York.

LAURISTON, A. (1994) : «Automatic Recognition of Complex Terms: Problems and the TERMINO Solution», *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, vol. 1, n° 1, John Benjamins Publishing Company, pp. 147-170.

LEBART, L. et A. SALEM (1988) : *Analyse statistique des données textuelle : questions ouvertes et lexicométrie*, Paris, Dunod, 210 pages.

MULLER, C. (1973) : *Initiation aux méthodes de la statistique linguistique*, Paris, Hachette, 187 pages.

MULLER, C. (1977) : *Principes et méthodes de la statistique lexicale*, Paris, Hachette, 206 pages.

SAGER, J. C. (1990) : *A Practical Course in Terminology Processing*, Amsterdam, John Benjamins Publishing Co., 254 pages.

SALEM, A. (1987) : *Pratique des segments répétés : essai de statistique textuelle*, Institut national de la langue française — INaLF, URL Lexicométrie et textes politiques, Publications de l'INaLF, Collection «Saint-Cloud», Paris, Klincksieck, 333 pages.

SALTON, G. (1988) : *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer*, New York, Addisson-Wesley Publishing Company, 530 pages.

SMADJA, F. (1993) : «Retrieving Collocations from Text: Xtract» *Computational Linguistics*, 19 (1), Association for Computational Linguistics, pp. 143-177.