Document généré le 19 juil. 2025 01:05

Meta
Journal des traducteurs
Translators' Journal

META

Une méthodologie d'identification automatique des syntagmes terminologiques : l'apport de la description du non-terme

Patrick Drouin

Volume 42, numéro 1, mars 1997

Lexicologie et terminologie

URI : https://id.erudit.org/iderudit/002593ar
DOI : https://doi.org/10.7202/002593ar

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

Drouin, P. (1997). Une méthodologie d'identification automatique des syntagmes terminologiques : l'apport de la description du non-terme. *Meta*, *42*(1), 45–54. https://doi.org/10.7202/002593ar

Résumé de l'article

L'utilisation de techniques purement linguistiques ne se prête pas facilement à l'identification automatique de la terminologie. Le problème se situe dans l'interaction entre les niveaux syntaxique et sémantique de la langue. Afin de pallier les problèmes que pose l'analyse linguistique, on propose une méthodologie automatique en deux grandes étapes. On se démarque des techniques traditionnelles, qui utilisent principalement une description du terme complexe, en utilisant une description syntagmatique du non-terme.

Tous droits réservés © Les Presses de l'Université de Montréal, 1997

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter en ligne.

https://apropos.erudit.org/fr/usagers/politique-dutilisation/

érudit

Cet article est diffusé et préservé par Érudit.

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche.

https://www.erudit.org/fr/

UNE MÉTHODOLOGIE D'IDENTIFICATION AUTOMATIQUE DES SYNTAGMES TERMINOLOGIQUES : L'APPORT DE LA DESCRIPTION DU NON-TERME[1]

PATRICK DROUIN*
Université de Montréal, Canada

Résumé

L'utilisation de techniques purement linguistiques ne se prête pas facilement à l'identification automatique de la terminologie. Le problème se situe dans l'interaction entre les niveaux syntaxique et sémantique de la langue. Afin de pallier les problèmes que pose l'analyse linguistique, on propose une méthodologie automatique en deux grandes étapes. On se démarque des techniques traditionnelles, qui utilisent principalement une description du terme complexe, en utilisant une description syntagmatique du non-terme.

Abstract

The use of purely linguistic techniques is not readily adaptable to computerized terminological idenfication. Difficulties occur when syntactic and semantic levels of language interact. To overcome problems posed by linguistic analysis, a two-stage methodology is proposed. This article shows how adopting a syntagmatic description of the non-term departs from the traditional approach which primarily uses a description of the complex term.

1. INTRODUCTION

Le travail du terminologue prend deux formes dont une est généralement plus connue que l'autre. La publication de dictionnaires scientifiques et techniques, lexiques, vocabulaires, glossaires, etc., donne une bonne visibilité à une portion de la chaîne du travail terminologique. Une portion beaucoup moins noble de cette dernière, le dépouillement terminologique, n'est cependant pas aussi connue du grand public. C'est cependant cette étape qui demande le plus de temps et d'efforts et qui fournit la matière première essentielle à tout travail purement terminologique.

La méthodologie de travail, en terminologie, se fonde essentiellement sur les corpus et sur leur analyse ; c'est cette étape que nous nommons *dépouillement*. Afin de recenser l'ensemble des termes d'un domaine, le terminologue doit lire l'ensemble du corpus. Ce dernier, pour être représentatif d'un domaine, doit avoir une taille considérable. Les corpus analysés sont maintenant disponibles sous forme électronique, et nous tenterons d'en tirer profit.

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

2. SOLUTIONS ENVISAGÉES

L'utilisation de l'ordinateur pour l'identification automatique de termes pose tout de même des problèmes. Nous procédons dans cette section à la présentation de trois grandes approches méthodologiques envisagées jusqu'ici pour le repérage des termes :

- l'approche linguistique ;
- l'approche statistique ;
- l'approche hybride.

2.1. L'APPROCHE LINGUISTIQUE

Jusqu'à tout récemment, deux grandes avenues ont été empruntées pour repérer automatiquement les termes. La première technique, l'approche linguistique, repose sur l'analyse syntaxique : celle-ci procède à l'identification des constituants d'une phrase et au marquage grammatical des formes.

Malgré le fait qu'une telle technique permette l'obtention de bons résultats, l'intérêt de ces derniers est pondéré par la dilution de l'information causée par un fort taux de *bruit*[2]. Sur le plan de la reconnaissance des unités complexes, les principaux problèmes proviennent du fait que l'analyse purement syntaxique ne fait que très rarement appel à la sémantique et, bien souvent, de manière très partielle. Une analyse de la structure de surface ne permet pas de distinguer le *terme* du *syntagme de discours* lorsqu'ils possèdent une structure syntaxique identique comme *panier d'osier* (terme) et *panier de pommes* (syntagme de discours).

À court terme, le couplage d'un module sémantique ne peut être une solution économiquement envisageable dans le cadre de la mise en place d'un système de reconnaissance des unités terminologiques complexes qui serait appelé à traiter des textes provenant de domaines très différents. La confection de dictionnaires électroniques ou la réutilisation de l'information sémantique contenue dans les grandes banques de terminologie sont encore trop coûteuses en temps et en efforts pour être facilement intégrées dans le cadre d'une démarche flexible.

L'approche linguistique offre cependant l'avantage de se rapprocher des intuitions du linguiste ; certains y verront cependant un inconvénient. Malgré l'utilisation de règles et de métarègles dans un langage informatique cryptique pour l'écriture de grammaires, ces dernières découlent des recherches récentes sur la formalisation des structures linguistiques et sont relativement simples à maîtriser pour qui veut bien s'en donner la peine.

Un des systèmes les plus connus de dépouillement terminologique reposant sur des techniques linguistiques est le système de l'Office de la langue française du Québec conçu par une équipe du Centre d'ATO de l'Université du Québec à Montréal, TERMINO (David 1990). Ce système est décrit par ses auteurs comme un système de reconnaissance de candidats-termes. La première étape d'analyse à laquelle se prête le logiciel est une analyse syntaxique des phrases afin d'isoler les synapsies potentielles qui sont ensuite soumises à une batterie de filtres en vue d'éliminer le maximum de bruit (Lauriston 1994).

Dans le cadre des approches linguistiques appliquées au repérage automatique de la terminologie, une nouvelle piste de recherche a été explorée par Bourigault (1992, 1993, 1994a et 1994b). Bien que cette dernière repose sur une catégorisation grammaticale des éléments de la phrase, la technique ne se fonde pas sur la recherche de matrices de formation des termes mais sur le dépistage de frontières de termes. L'algorithme recherche ainsi les mots dont la catégorie grammaticale ne peut pas faire partie d'un terme (*frontières*), et fabrique de cette façon une liste de candidats. Cette liste sera ensuite revue par un algorithme d'analyse syntaxique locale qui extrait les candidats potentiels à l'intérieur du segment de texte identifié à l'étape précédente. Malgré le fait qu'elle ne

fasse pas appel à des techniques purement linguistiques, elle donne de bons résultats qui sont exploités par un logiciel commercialisé (LEXTER).

2.2. L'APPROCHE STATISTIQUE

Les méthodes statistiques ont le principal avantage d'être parfaitement systématiques et très rapides. La rapidité vient souvent du fait que ces filtres statistiques ne manient pas des mots, mais des chiffres qui se rattachent aux mots. Les procédures qui manipulent des données numériques ne demandent pas autant d'appels à la récursivité que les procédures d'analyse linguistique. Un analyseur syntaxique consacre la majorité de son temps à l'essai de pistes d'analyse qui échoueront. De plus, contrairement aux algorithmes linguistiques qui travaillent sur l'ensemble des données, les analyses statistiques identifient des groupes de données qui feront l'objet de traitements.

En outre, l'approche statistique nécessite moins de ressources logicielles et d'injection de connaissances. L'analyse linguistique repose, et reposera encore pour longtemps, sur un ensemble de dictionnaires qui sont utilisés pour la lemmatisation des formes en cours d'analyse. Du côté des connaissances, les règles grammaticales des systèmes d'analyse syntaxique actuels reposent en bonne partie sur la notion de rôle (agent, patient, etc.), et sur la désambiguïsation et la catégorisation des constituants de la phrase. L'information nécessaire à de tels traitements doit inévitablement être encodée par l'humain dans une étape de pré-traitement du corpus. L'analyse d'un domaine nouveau implique donc la création de nouveaux dictionnaires et l'ajustement des règles existantes. Ces ajustements peuvent parfois être majeurs selon les particularités syntaxiques, sémantiques ou lexicales du domaine traité.

Cette dépendance du système par rapport au domaine analysé est aussi supprimée lors d'un recours à des règles d'analyse statistiques. Ces dernières sont donc très avantageuses dans un contexte de recherche terminologique où le domaine traité varie très régulièrement et où il est impensable de songer à mettre sur pied des dictionnaires électroniques. De plus, il serait plutôt paradoxal de procéder à cette étape de pré-traitement, car il s'agit de l'objectif du terminologue à ce moment précis de la démarche.

2.3. L'APPROCHE HYBRIDE

Depuis peu, de nouvelles méthodes de traitement viennent faire le pont entre les deux courants déjà identifiés. Les approches hybrides tentent d'offrir les avantages des précédentes tout en essayant d'éviter leurs inconvénients. Elles sont systématiques, proches de l'intuition linguistique, et elles proposent un compromis intéressant entre les ressources logicielles nécessaires à l'analyse et l'efficacité des algorithmes.

Deux grands angles d'approche sont possibles dans le cadre d'une approche hybride combinant linguistique et statistique. On peut d'abord choisir comme point de départ la technique linguistique que complète un algorithme statistique ; c'est l'approche adoptée par Daille (1993 et 1994) et Lauer (1994). On peut à l'inverse, et c'est ce que nous avons choisi de faire, appliquer, dans un premier temps, des méthodes statistiques que nous couplons par la suite à des méthodes linguistiques.

2.3.1. De la linguistique à la statistique

Les fondements linguistiques utilisés pour le repérage de la terminologie sont les mêmes pour la majorité des auteurs qui se sont intéressés au domaine. Les corpus d'analyse sont des textes dont les mots ont préalablement été marqués (parties du discours). Dans le cadre de son approche hybride, Daille propose une première étape de dépistage qui se fonde sur une analyse du texte selon des matrices de formation des termes. Les séquences correspondant à des patrons syntagmatiques qui sont utilisés pour la construction de termes complexes sont relevées et ajoutées à une liste de candidats. Plutôt que de tenter

de filtrer cette liste de façon linguistique, Daille utilise des algorithmes qui prennent en considération les extrémités lexicales d'un candidat afin de le soumettre à une batterie de tests statistiques. Les résultats de ces derniers déterminent si les candidats seront conservés ou éliminés de la liste.

Pour sa part, Lauer procède à une analyse sur l'anglais, ayant pour but de repérer les composés de type N_1 N_2 ... N_n qui sont extrêmement fréquents dans cette langue. Un algorithme reposant sur des dictionnaires de catégorisation grammaticale contenant les formes nominales non ambiguës balaie un texte tout en identifiant les séquences de mots présents au dictionnaire et séparés uniquement par des blancs. Un calcul statistique évalue, par la suite, le degré d'association conceptuelle[3] (à partir d'un thesaurus) entre les éléments du segment de texte retenu. L'analyse syntaxique qui suit prend en considération le poids de l'association conceptuelle entre les divers éléments afin de construire un arbre syntaxique. Ainsi, si l'association des deux premiers éléments reçoit une valeur plus grande que l'association des deux suivants, l'arbre de ce segment composé de trois éléments sera [[N_1 N_2] N_3]. Cette technique pourrait être adaptée au français. Les matrices des termes complexes sont cependant plus complexes ; les algorithmes devraient donc être modifiés de façon considérable.

2.3.2. De la linguistique à la statistique

La méthodologie que nous adoptons procède à l'inverse de celles présentées dans la section précédente. Nous appliquons tout d'abord des méthodes statistiques à des textes afin d'en extraire des candidats, et nous les filtrons ensuite à l'aide de méthodes linguistiques. Il ne s'agit pas d'une prise de position théorique mais plutôt d'une décision découlant des ressources disponibles au moment d'entreprendre la recherche.

Les outils nécessaires à une analyse linguistique automatisée ne sont pas faciles d'accès pour la majorité des chercheurs. Les analyseurs permettant d'analyser et de marquer grammaticalement un texte d'un domaine arbitraire sont rares et nécessitent une puissance de traitement informatique non négligeable. Nous croyons que nous pouvons, tout en procédant à des analyses linguistiques plus restreintes, obtenir des résultats comparables à ceux obtenus par les autres approches. En évitant le recours aux dictionnaires électroniques, nous conservons une certaine distance par rapport au domaine traité et obtenons une autonomie d'analyse accrue. De plus, l'analyse syntaxique locale est beaucoup plus rapide qu'une analyse de la phrase en ses constituants.

3. MÉTHODOLOGIE HYBRIDE

Notre méthodologie est composée des étapes suivantes :

- analyse statistique :
  - — repérage des candidats-termes,
  - — évaluation probabiliste du statut des candidats ;
- analyse linguistique :
  - — filtrage selon la morphologie des non-termes complexes,
  - — analyse de l'autonomie des candidats,
  - — analyse des candidats en contexte.

3.1. ANALYSE STATISTIQUE

3.1.1. Repérage des candidats-termes

Le repérage des candidats-termes est, pour le moment, articulé autour d'un algorithme de repérage fort simple qui permet l'obtention de résultats indiscutables. La technique

utilisée est celle présentée dans Choueka *et al.* (1983) et Choueka (1988). Cet algorithme identifie, sans discrimination, les enchaînements de mots qui se répètent plus souvent qu'un seuil de fréquence donné ; la technique est semblable à celle mise au point par Lebart et Salem (1988) et Salem (1987). À l'instar de ces auteurs, nous utiliserons le terme *segment* pour décrire un enchaînement de mots.

Les candidats-termes, qui forment un sous-ensemble des segments, sont dilués dans une multitude de segments sans intérêt pour le travail du langagier. Il faut donc envisager de mettre au point des techniques de filtrage des segments afin d'isoler les candidats-termes, tout en éliminant le plus de bruit possible. C'est en effet en fournissant au terminologue une liste peu encombrée et rapidement utilisable que l'outil trouve toute son utilité. Dans les sections qui suivent, une étape de filtrage est dite *positive* lorsqu'elle procède à la bonification d'un segment au sein de la liste des candidats ou *négative* lorsqu'elle procède au retrait d'un segment de la liste des candidats.

3.1.2. Évaluation probabiliste du statut des candidats

Cette étape positive de filtrage évalue le potentiel des mots qui forment un segment à se retrouver côte à côte dans le texte étudié. Si la fréquence d'occurrence d'un segment dévie de façon significative de la fréquence théorique calculée à partir de la fréquence des mots qui le forment, l'importance accordée au segment comme candidat-terme est augmentée. Cette étape cherche donc à identifier des écarts, sur le plan de la fréquence, entre la théorie et la réalité, et à les quantifier. L'intuition du terminologue est à la base de cette règle statistique qui ne fait que confirmer de façon empirique ce que le langagier observe à la lecture d'un texte spécialisé.

3.2. ANALYSE LINGUISTIQUE

Pour procéder à une analyse plus fine des résultats offerts par les premiers algorithmes statistiques, nous employons d'autres techniques de filtrage, plus proche de la réalité du traitement linguistique.

3.2.1. Filtrage selon la morphologie des non-termes complexes

Un repérage automatique des termes complexes reposant sur l'utilisation de matrices terminogéniques fait appel à une suite de règles qui ne peut posséder un caractère fermé et dont l'application repose en majeure partie sur des catégories grammaticales ouvertes. Un échantillon tiré d'un texte portant sur le domaine de la *géomatique* permet de donner un aperçu des matrices nécessaires à la description d'un domaine scientifique ou technique.

1. **nom + adjectif**

   carte topographique

2. **nom + nom**

   arpenteur géomètre
   système de localisation gps

3. **nom + prép. + nom**

   système d'information sur le territoire
   centre de recherche en géomatique
   structure en quadrants arborescents

4. **nom + prép. + dét. + nom**

   système d'information sur le territoire

5. nom + de + nom

système de gestion de base de données
ordre des arpenteurs géomètres du québec
système de gestion de base de données localisées
système de gestion de base de données à référence spatiale
corporation d'information géographique du nouveau québec
base de données à référence spatiale
système d'information géographique
modèle numérique de terrain

6. nom + prép. + verbe infinitif

table à numériser

Les matrices indiquées plus haut ne correspondent qu'à la matrice globale du terme. Afin de permettre la reconnaissance de ces termes, des matrices locales doivent être appliquées récursivement afin de composer les divers constituants du terme. L'ensemble suivant contient les matrices complexes nécessaires à la reconnaissance des termes précédents :

1. nom + adjectif
2. nom + adjectif + de + nom
3. nom + de + nom
4. nom + de + nom + adjectif
5. nom + de + nom + adjectif + de + adjectif + nom
6. nom + de + nom + de + nom + de + nom
7. nom + de + nom + de + nom + de + nom + adjectif
8. nom + de + nom + de + nom + de + nom + prép. + nom + adjectif
9. nom + de + nom + nom
10. nom + de + nom + nom + prép. + nom
11. nom + de + nom + prép. + de + nom
12. nom + de + nom + prép. + nom
13. nom + de + nom + prép. + nom + adjectif
14. nom + nom
15. nom + prép. + de + nom
16. nom + prép. + nom
17. nom + prép. + nom + adjectif
18. nom + prép. + infinitif

Comme le démontrent les exemples qui précèdent, la reconnaissance d'une matrice globale de type **nom de nom** implique bien souvent un appel à la récursivité[4]. Le problème posé par le recours à la récursivité est l'imposition d'une limite au nombre de récursions possibles. La limite des enchâssements possibles pour la construction d'un terme complexe n'est pas imposée par des facteurs linguistiques, mais par des facteurs liés à la compréhension humaine (Auger 1979). L'imposition arbitraire, afin de borner la récursivité, est limitative, peu efficace, et laisse échapper de nombreux candidats intéressants.

De plus, la rigidité des matrices ne permet pas de rendre compte des phénomènes de dynamique discursive d'insertion et de réduction auxquels sont parfois soumis les termes complexes en situation textuelle. Le comportement textuel du terme complexe est souvent bien différent de celui dont on rend compte dans les lexiques. Cette lacune rend difficile, voire impossible, la représentation formelle de la structure du terme selon des patrons de formation syntagmatique dans un système informatique. Il faut donc tenter d'aborder le problème sous un nouvel angle. 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[5]. L'approche que nous proposons fait appel à la statistique mais surtout à des descriptions formelles du *non-terme*. Nous définissons le *non-terme* comme un enchaînement (dans le cas de notre recherche qui porte sur les formes complexes) lexical qui ne correspond pas à une notion.

Comme nous l'avons mentionné précédemment, la description du terme est formellement difficile à réaliser, sa structure ne se laissant pas formaliser facilement. À partir de cette constatation, nous avons opté pour une approche différente qui, selon nous, offre de meilleurs résultats. Elle permet de relever un plus grand nombre de candidats valables. Pour y arriver, nous avons élaboré une série de règles que nous appliquons à la liste des segments afin de filtrer les candidats indésirables :

1. un pronom ne peut faire partie d'un terme complexe ;
2. une préposition ne peut débuter ou terminer un terme complexe ;
3. un article ne peut débuter ou terminer un terme complexe ;
4. une conjonction ne peut débuter ou terminer un terme complexe ;
5. un adverbe ne peut débuter ou terminer un terme complexe ;
6. un verbe conjugué ne peut faire partie d'un terme.

Ces règles décrivent des restrictions générales dont la validité est relative car il est relativement simple de trouver un exemple de termes qui s'articule autour d'un adverbe. Cependant, nous croyons qu'il vaut mieux présenter des règles valables dans la majorité des cas (c'est le cas de celles que nous présentons) et risquer d'éliminer à l'occasion des termes complexes plutôt que de conserver une liste de candidats qui n'offre que très peu d'intérêt. L'algorithme de filtrage négatif identifie les segments de la liste de candidats qui correspondent à l'une de ces règles et les élimine de la liste.

Les règles citées ci-dessus, à l'exception de la dernière, reposent sur des catégories grammaticales fermées qui rendent possible le recours à des dictionnaires. La liste des prépositions est bien connue de tous, tout comme celle des conjonctions. La catégorisation grammaticale des éléments n'est cependant pas aussi simple lorsqu'on a recours à des matrices utilisant des catégories ouvertes comme les substantifs ou les adjectifs. Dans le cas des verbes, nous utilisons un algorithme qui permet de désambiguïser de façon locale et heuristique la majorité des cas qui lui sont présentés.

3.2.2. Analyse de l'autonomie des candidats

Diverses techniques statistiques ont été utilisées pour illustrer le lien qui existe entre deux unités lexicales ou plus dans un texte (Daille 1993 : 115-150). En ce qui concerne les termes complexes, nos tests démontrent que la simple comparaison des segments entre eux permet d'obtenir de l'information pertinente. 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 textuelle. Les règles qui suivent permettent d'éliminer des candidats retenus au cours des étapes de filtrage statistique et de filtrage des non-termes. Les principaux problèmes que pose la liste obtenue à la suite de l'étape précédente sont ceux de la redondance et du recoupement entre les segments retenus. Les règles utilisées pour résoudre ce problème sont les suivantes :

1. si la fréquence d'un segment est la même que la fréquence d'un segment plus long qui le contient, alors le segment plus court n'a pas de comportement textuel autonome ;
2. si la fréquence d'un segment est supérieure à la fréquence d'un segment plus long qui le contient, alors le segment plus court a un comportement textuel autonome.

La règle 1 filtre négativement les segments de la liste des candidats-termes alors que la règle 2 filtre positivement la liste par bonification du statut du segment le plus long. En effet, les segments décrits par la règle 2 se comportent comme des termes complexes qui donnent naissance à des termes plus longs par détermination et expansion à droite[6]. L'analyse de l'inclusion constitue donc 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. Il serait évidemment possible de pousser beaucoup plus loin cette analyse à l'aide de techniques statistiques élaborées, mais l'observation de la fréquence semble suffisante. La fréquence, en terminologie, est un indice particulièrement révélateur (voir Daille 1993 et 1994) dont il faudra chercher à tirer profit dans un plus grand nombre de cas.

3.2.3. Analyse des candidats en contexte

Les règles d'analyse des candidats en contexte que nous utilisons peuvent, au premier abord, sembler *ad hoc* et difficiles à justifier. Nous croyons cependant, tout comme Bourigault (1992, 1993, 1994a et 1994b), que des analyses empiriques pratiquées sur des corpus de grande taille et l'évaluation de leur pertinence suffisent pour justifier des règles qui n'ont pas nécessairement de base théorique linguistique. Ces règles ont été élaborées à partir de l'observation de plusieurs milliers de contextes dans lesquels figurent des termes complexes afin d'identifier le comportement textuel de ces derniers. Les redondances ont été notées et mises à profit dans la rédaction des règles qui suivent :

1. bonification des candidats aux aspects graphiques particuliers (majuscules, guillemets, etc.) ;
2. bonification des candidats précédés d'un déterminant ;
3. bonification des candidats précédés ou suivis d'un séparateur fort ;
4. bonification des candidats suivis d'un verbe conjugué ;
5. bonification des candidats suivis d'un pronom relatif.

Le filtrage par analyse contextuelle, contrairement à l'étape précédente, ne procède pas à l'élimination de candidats ; il s'agit d'une étape de filtrage négative. Des grammaires locales (Silberztein 1993) examinent le contexte sommairement. Une pondération est ensuite attribuée au candidat-terme : elle évalue le statut potentiel d'un candidat (terme ou non-terme) selon le nombre d'indices présents en contexte. Les grammaires locales examinent le contexte précédant et suivant la chaîne de mots et s'assurent que 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. Notre analyse prend en considération divers critères qui permettent de caractériser le comportement textuel d'un terme, et la pondération obtenue sera d'autant plus élevée que le segment se comporte en contexte comme un terme complexe.

Une étape de fusion des différentes observations réalisées localement lors des diverses étapes de filtrage positif suit l'analyse contextuelle. La figure 1 illustre la mise en place de la cote globale qui représente le potentiel d'un segment à être un terme ou son *potentiel terminogénique*.

Les candidats sont ensuite triés selon leur potentiel terminogénique et le terminologue peut ainsi accorder plus d'attention aux candidats les plus intéressants qui apparaissent en tête de liste. Le prototype utilisé possède un module paramétrable qui facilite la gestion des candidats-termes et l'exclusion automatique des candidats possédant un potentiel terminogénique inférieur à un certain seuil.

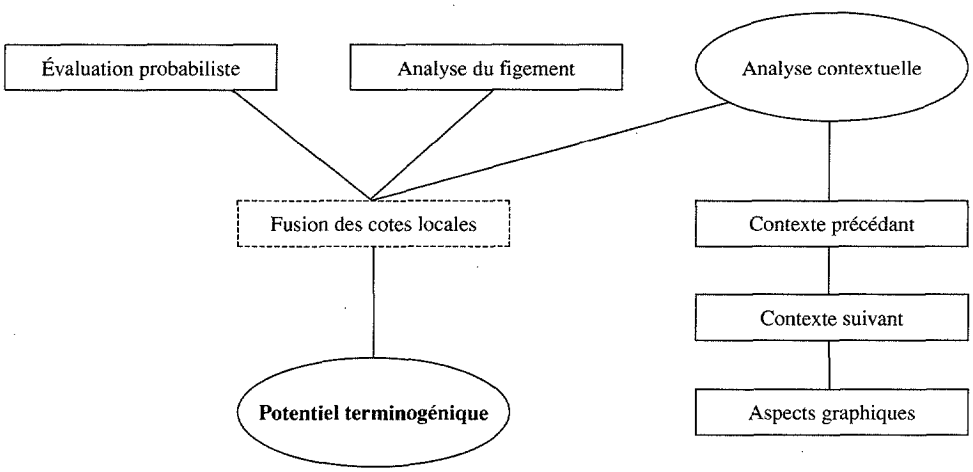

Figure 1 :
Calcul du potentiel terminogénique d'un segment

4. CONCLUSION

La qualité des résultats obtenus à l'aide d'une technique hybride constitue son principal avantage. L'épuration des résultats à l'aide de techniques linguistiques permet une réduction considérable du bruit par rapport aux méthodes purement linguistiques ou statistiques. La majorité des candidats recensés présentent un intérêt pour le travail du langagier indépendamment du domaine de spécialité qui fait l'objet du traitement. Cette méthodologie s'adapte donc facilement aux conditions de travail ponctuelles auxquelles est confronté le terminologue en situation de production.

Le terminologue peut donc utiliser les résultats dans le cadre d'une démarche traditionnelle, automatique, ou assistée par ordinateur. Il demeure important, dans l'état d'avancement des travaux en terminotique, de ne viser que l'automatisation de certaines parties de la chaîne de travail. Ce fractionnement permettra d'automatiser avec beaucoup plus de précision l'ensemble des tâches du terminologue. L'automatisation des étapes purement mécaniques du travail terminologique est primordiale si nous voulons un jour envisager une méthodologie de travail entièrement automatisée.

Il est essentiel de reconnaître que, pour plusieurs années encore, le recours à l'expertise du langagier dans le cadre d'une démarche automatique est essentiel. Il est tout aussi important que le langagier se rende compte de l'aspect inévitable du recours à la machine et du gain inestimable en temps et en précision qui en découle.

Notes

* Cet article est issu d'une communication présentée par l'auteur aux IV^{es} Journées scientifiques du réseau «Lexicologie, terminologie, traduction» de l'AUPELF-UREF (Lyon, France, 28, 29, 30 septembre 1995).

1. Nous tenons à remercier le *Conseil de recherche en sciences humaines* (CRSH) du Canada pour sa contribution à cette recherche. Nous remercions aussi Marie-Claude L'Homme pour sa lecture attentive de ce texte et ses nombreux commentaires.

2. Nous nommons *bruit* tout résultat indésirable pour le travail terminologique.
3. Cette technique s'oppose à celle de Daille qui s'intéresse à l'attirance et les lexèmes d'un point de vue lexical.
4. Ainsi, l'analyse du terme *système de gestion de base de données* fait appel à trois applications successives de la matrice terminogénique **nom + de + nom.**
5. Nous nous rapprochons ici de ce que fait Bourigault (1992, 1993, 1994a et 1994b). Ce dernier adopte une méthodologie qui se fonde sur le dépistage des *frontières de termes*. Ces dernières se distinguent de notre description du non-terme en ce qu'elles sont identifiées textuellement, en cours d'analyse, et qu'elles permettent ainsi le dépistage des candidats. Pour sa part, notre description du non-terme est utilisée pour améliorer les résultats des techniques statistiques qui ont procédé au repérage des candidats.
6. C'est le cas avec *base de données* qui donne naissance à *base de données à référence spatiale*.

RÉFÉRENCES

AUGER, P. (1979) : «La syntagmatique terminologique, typologie des syntagmes et limite des modèles en structure complexe», *Table ronde sur le problème 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 Fourtheenth International Conference on Computational Linguistics (COLING 92)*, Nantes, pp. 977-981.

BOURIGAULT, D. (1993) : «Analyse syntaxique locale pour le repérage de termes complexes dans un texte», *T.A.L.*, 34 (2), pp. 105-118.

BOURIGAULT, D. (1994a) : *Un logiciel d'extraction de terminologie. Application à l'acquisition de connaissances à partir de textes*, Thèse de l'École des Hautes Études en Sciences Sociales, 352 p.

BOURIGAULT, D. (1994b) : «Acquisition automatique des termes complexes en français et en anglais, approche comparative», P. Bouillon et D. Estival (Eds), *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*, 2 et 3 décembre, Genève, ISSCO, pp. 29-43.

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

CHOUEKA, Y. (1988) : «Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in a Large Textual Database», *Actes de colloque du RIAO 88*, Cambridge, CUP, 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, 4 (1), pp. 34-39.

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

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

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*, 21-24 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», P. Bouillon et D. Estival (Eds), *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*, 2 et 3 décembre, Genève, ISSCO, pp. 18-28.

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

LADOUCEUR, J. et P. DROUIN (1995) : «Une analyse terminométrique pour le repérage automatique des descripteurs complexes dans les textes de spécialité», *Meta*, Montréal, Presses de l'Université de Montréal.

LAUER, M. (1994) : «Conceptual Association for Compound Noun Analysis», *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Student Session*, Juin, Las Cruces, 6 p.

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

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

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

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

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

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