Meta
Journal des traducteurs
Translators' Journal

META

Le syntagme nominal, une nouvelle approche des bases de données textuelles

Richard Bouché

Volume 34, Number 3, septembre 1989

1. Actes du Colloque Les terminologies spécialisées : Approches quantitative et logico-sémantique et 2. Actes du Colloque Terminologie et Industries de la langue

URI: https://id.erudit.org/iderudit/002106ar
DOI: https://doi.org/10.7202/002106ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Bouché, R. (1989). Le syntagme nominal, une nouvelle approche des bases de données textuelles. *Meta*, *34*(3), 428–434. https://doi.org/10.7202/002106ar

LE SYNTAGME NOMINAL, UNE NOUVELLE APPROCHE DES BASES DE DONNÉES TEXTUELLES

RICHARD BOUCHÉ
École Nationale Supérieure de Bibliothécaires, Lyon, France

I. INTRODUCTION

Dans son exposé concernant les relations entre terminologie et lexique[1], M. Le Guern a mis en évidence l'importance de la prise en compte des phénomènes de référence à la réalité extra-linguistique. Cet appel à la référence, alors qu'il constitue le fondement d'un certain nombre de procédures qui au bout du compte permettent à l'utilisateur de trouver une information pertinente à sa recherche, est presque toujours négligé ou au plus considéré comme mineur au moment de la conception des outils permettant la gestion et le traitement de l'information dans les bibliothèques, les centres de documentation ou les bases de données. Or nous allons essayer de montrer que ces considérations sur la valeur référentielle des descripteurs documentaires, non seulement conduisent à proposer de nouveaux outils, et en particulier permettent une approche prometteuse dans la conception de bases de données textuelles, mais aussi pourraient être un moyen de résoudre les difficultés rencontrées par l'utilisateur final dans ses recherches dans les bases de données bibliographiques ou les catalogues en ligne.

Une panorama rapide des travaux portant sur l'indexation automatique et une analyse des conditions de fonctionnement des systèmes actuels vont nous permettre de préciser l'importance du mécanisme de référence. Dans une seconde partie, nous montrerons comment la prise en compte de la valeur référentielle de certaines parties du discours peut conduire à une nouvelle conception d'une base de données textuelles.

II. INDEXATION AUTOMATIQUE

Le traitement du document écrit en vue de l'intégrer de façon automatique à une base de données textuelles consiste à identifier dans le texte les éléments qui en donnent une représentation du contenu.

On peut envisager diverses approches :

A. LES MÉTHODES STATISTIQUES

Elles consistent à définir un modèle probabiliste des occurrences des mots dans un document considéré à l'intérieur d'une collection bien définie. L'objectif est de déduire du comportement statistique de ces mots leur caractère pertinent pour participer à la description du contenu.

Les difficultés rencontrées sont très liées au principe de la méthode employée. Les résultats ne peuvent être considérés comme donnant une bonne approximation des phénomènes analysés que si les fréquences observées sont suffisamment importantes. Or, c'est un phénomène bien connu, dans les textes les mots de basse fréquence et notamment des hapax sont fréquents. Il est donc indispensable de regrouper au maximum les formes sui-

vant des critères de variation morphologique. Ce problème se pose encore plus dans les langues à morphologie riche comme le français, que pour d'autres comme l'anglais par exemple.

On aboutit d'ailleurs à un blocage complet si on veut prendre en compte, non pas des mots isolés, mais des expressions plus complexes.

Enfin, la propriété de pertinence identifiée à propos d'un mot donné est liée plus ou moins à la collection échantillon considérée. L'évolution du fonds documentaire conduit alors fatalement à un changement de l'indice de pertinence au cours du temps. L'indexation n'est plus la même. Que faire alors des documents déjà intégrés à la base ?

B. LES MÉTHODES FONDÉES SUR UNE APPROCHE «INTELLIGENCE ARTIFICIELLE»

Il s'agit, ici, des techniques de l'intelligence artificielle appliquées au traitement automatique des langues. La propriété que l'on cherche à identifier est l'appartenance plus ou moins directe (pouvant résulter d'un mécanisme d'inférence) du mot à un réseau sémantique complexe défini a priori. D'une certaine façon, on peut considérer ces méthodes comme un développement de la technique classique des mots-clés. La représentation du domaine de la connaissance, assez rudimentaire dans le cas d'un thesaurus, est très détaillée grâce à de multiples relations dans les réseaux sémantiques.

L'analyse d'un document en entrée procède, en fait, par un filtrage à travers le réseau de représentation des connaissances posé à l'avance. Ce caractère a priori du procédé pose le problème de l'évolution du domaine couvert et de la non détection d'informations nouvelles.

En réalité, cette approche est plutôt utilisée pour l'analyse des questions d'un utilisateur en recherche d'information (voir ci-dessous).

C. L'APPROCHE LINGUISTIQUE

L'identification des parties du discours pertinentes se fait ici sur des critères linguistiques.

Le mot a longtemps été considéré comme l'unité principale. Et la catégorie grammaticale à laquelle il appartient constitue le premier critère de sélection. Pour le français, des considérations prenant en compte l'observation des résultats d'une indexation manuelle conduisent à ne considérer que les catégories «nom»et «adjectif». Cela fait quand même beaucoup de formes qui restent pour décrire un document. En essayant de combiner l'approche statistique avec l'approche linguistique on arrive à obtenir, en bout du compte, un nombre raisonnable de termes.

Cette approche, suivie par le groupe SYDO Lyon[2] a été à l'origine de la réflexion sur le statut linguistique du mot-clé.

L'expérimentation[3] avait été menée sur un corpus de dossiers médicaux contenant les comptes rendus des visites des médecins d'un service de cardiologie aux patients hospitalisés. Les résultats portant sur l'identification de mots isolés étaient catastrophiques[4]. Par contre ils se révelaient nettement bons si on prenait en compte des expressions linguistiques identifiées par un automate et permettant de repérer des formes du type : Nom-Préposition-Nom, Adjectif-Nom, Nom-Adjectif, Nom-Préposition-Déterminant-Nom, etc.

La réflexion sur de tels résultats a mis en évidence la confusion qui règne en matière d'indexation automatique entre langue et discours. Le lexique, en tant que composante de la langue, ne contient que des éléments qui sont des propriétés c'est-à-dire des prédicats. Le mot est donc un prédicat et il ne peut pas, considéré isolément, faire référence à un objet de la réalité extra-linguistique de l'auteur du document. Il ne peut pas exprimer «ce dont parle le document». Il ne peut donc pas être descripteur.

Comme le dit Michel Le Guern: *La finalité du descripteur exclut qu'on puisse l'envisager en faisant abstraction de la valeur référentielle de ses occurrences dans le corpus. Les mots de la langue, en tant qu'ils sont mots de la langue, ne signifient que des attributs, et non des substances, tant qu'ils ne sont pas mis en œuvre dans le discours. Le descripteur, quant à lui signifie une entité au sens de la philosophie d'Aristote. Le descripteur ne peut donc pas être considéré, à l'instar des mots de la langue comme un symbole sans référence.*[5]

On admettra donc que la plus petite unité du discours porteuse d'une valeur référentielle est le syntagme nominal. C'est elle qu'il importe d'identifier dans le document. Le groupe SYDO s'est donc attaché à définir la grammaire correspondante (qui sera explicitée ci-dessous), les outils permettant d'identifier de telles unités et de les représenter au sein d'une base de données afin de permettre la recherche ultérieure de l'information.[6]

III. VALEUR RÉFÉRENTIELLE ET SYSTEMES D'INFORMATIONS

Essayons de définir un modèle général des systèmes d'informations qui nous préoccupent (bibliothèque, centre de documentation, base de données bibliographiques, base de données textuelles).[7] Nous allons les considérer comme un canal complexe de communication (figure 1). Les messages en entrée sont émis par un grand nombre d'émetteurs. La communication entre certains d'entre eux et un utilisateur du système est réalisée grâce aux propriétés de mémorisation du canal et à l'aiguillage qu'il possède.

Fig. 1. Modélisation du système d'information.

Suivant les cas, la fonction d'aiguillage est remplie par un langage classificatoire ou un langage documentaire.

1. CLASSIFICATION

Une classification est un système symbolique hiérarchisé fondé sur des propriétés allant du plus général au particulier. Comme tout système symbolique un langage classificatoire est purement arbitraire[8]. Le résultat est un ensemble de classes disjointes par définition. Chacune d'elles est dotée d'un indice de classification qui sert à décrire le contenu de l'ouvrage concerné[9]. On se trouve ici dans un système formel à valeur référentielle nulle.

2. THESAURUS

Le traitement des articles de périodiques au contenu généralement plus spécialisé que celui des livres ne peut être fait de façon satisfaisante à l'aide d'une classification. Un système combinatoire reposant sur des mots clés est donc utilisé pour définir des classes (non disjointes, cette fois) de références bibliographiques. Le mécanisme d'accès aux références repose alors sur des opérateurs ensemblistes classiques. L'ensemble des mots clés muni d'un certain nombre de relations sémantiques (synonymie, hyperonymie, hypo-

nymie, association) constitue le langage documentaire ou thesaurus. Cette construction est un système symbolique aussi arbitraire que le précédent. On peut même dire que la multiplicité des choix qui doivent être faits au moment de sa conception lui donne un caractère beaucoup moins cohérent qu'une classification[10]. Comme l'a montré Michel Le Guern la valeur référentielle d'un tel langage est aussi nulle (voir ci-dessus).

3. MISE EN ŒUVRE DE MÉCANISMES DE RÉFÉRENCE À LA RÉALITÉ EXTRA-LINGUISTIQUE

Dans la réalité du fonctionnement de ces systèmes d'information, la description du contenu d'un document ne se réduit heureusement pas à un indice de classification, ou à une liste de mots-clés.

La notice bibliographique qui accompagne en principe chaque document comporte d'autres informations comme le titre, l'auteur, quelques fois un résumé, etc. Ces éléments sont très porteurs de référence à la réalité extra-linguistique et permettent à l'utilisateur de faire un tri final beaucoup plus pertinent que celui qui résulte de l'usage exclusif d'une classification ou d'un thesaurus.

Dans le cas du livre, et aussi, dans une moindre mesure, dans le cas des périodiques, l'accès direct au document (déjà localisé sur les rayonnages grâce à l'indice de classification ou par le nom de la revue), permet le fonctionnement d'un autre mécanisme de référence par le moyen de la consultation de la table des matières, de l'introduction, de la préface, etc.

Il importe de noter le rôle très important joué par le bibliothécaire ou le documentaliste qui connait bien son fonds et qui intervient dans le fonctionnement général du système grâce à sa mémoire et à son aiguillage personnels. En réalité, grâce à sa pratique quotidienne, il a réussi à associer au système symbolique utilisé les liens référentiels nécessaires. Ce ne peut être le cas de l'utilisateur plus ou moins occasionnel qui a maintenant à sa portée, par la télématique, les bases de données documentaires et les catalogues des bibliothèques. Les solutions actuelles consistent à ajouter des éléments d'aide qui sont autant d'appels à la réalité extra-linguistique. Par exemple :

- le recours au texte plein par l'intermédiaire d'opérateurs de troncature ou de proximité[11],
- la consultation des titres et des résumés permettant à l'utilisateur d'indiquer au système, parmi les documents qui lui sont signalés, ceux qu'il juge pertinents. Au dispositif, ensuite, de relancer une recherche en tenant compte de cette information.

Cette analyse de la situation confirme donc les hypothèses de Michel Le Guern et du groupe SYDO sur l'importance du phénomène de référence à la réalité extra-linguistique. En reprenant le schéma de la figure 1 on peut donc dire que l'objectif du concepteur d'un tel système d'informations est de favoriser au maximum la conservation de la référence, à travers le canal de communication. Le perfectionnement et la complexification du système symbolique pour rendre plus précises et plus adaptées les relations utilisées sont voués à l'échec, si les mécanismes de référence sont éliminés.

En particulier, on peut s'interroger sur les recherches en cours et qui ont pour objectif de partir d'une requête en langue naturelle pour la transformer (indépendamment de l'utilisateur) en une question adaptée au système d'informations. Le mécanisme de référence est alors complètement court-circuité. Seul peut être mise en œuvre par la machine le système symbolique qui lui a été fourni. Il faut ajouter que cette démarche suppose que l'utilisateur sait correctement poser sa question même en langue naturelle. Très souvent, ce n'est pas le cas.[12] En tout état de cause, cette analyse met un évidence l'importance des phénomènes linguistiques.

IV. BASES DE DONNÉES TEXTUELLES

Les considérations précédentes ont été appliquées par SYDO Lyon au cas des bases de données textuelles.

A. MODÈLE LINGUISTIQUE

L'objectif d'un tel modèle est de permettre l'identification des syntagmes nominaux, tout en mettant bien en évidence à travers la structure syntaxique reconnue les phénomènes de transition entre les mots qui appartiennent au lexique et qui sont des prédicats, et les syntagmes qui pointent sur des objets de la réalité extra-linguistique.

Le modèle conçu a pour objectifs :

- permettre l'identification des syntagmes nominaux,
- déterminer la structure de ces syntagmes en mettant en évidence les relations entre ses constituants. Ceci permet le stockage d'une représentation du syntagme nominal facilitant la recherche d'informations.
- bien montrer le mécanisme de passage partant des mots (prédicats fonctionnant dans une logique intensionnelle) et arrivant à l'unité à valeur référentielle (le syntagme nominal, dans le cadre d'une logique extensionnelle).

La grammaire de reconnaissance du syntagme nominal s'articule autour de trois niveaux : [13] (fig. 2)

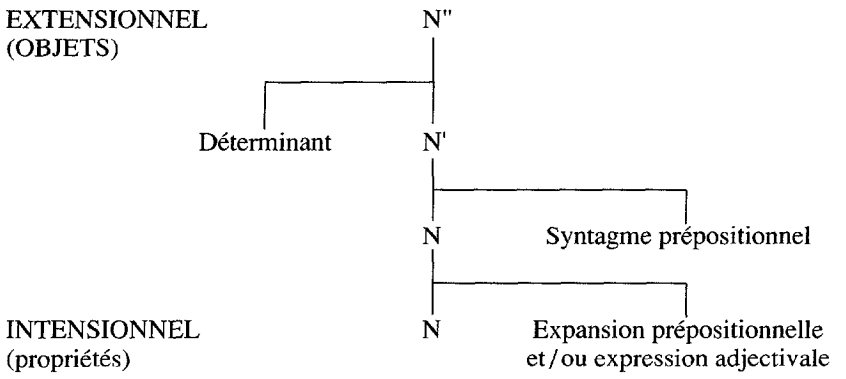

Figure 2. Grammaire du syntagme nominal

N : nous sommes au niveau purement intensionnel. Les unités considérées sont des prédicats simples (le nom) ou complexes (les propriétés du nom sont modifiées par des éléments adjectivaux ou des expansions prépositionnelles).

N' : c'est la transition entre l'intensionnel et l'extensionnel. La prise en considération de l'univers du discours considéré, en particulier par l'intervention de syntagmes prépositionnels introduisant des éléments dont on peut présupposer l'existence, définit une classe d'objets de la réalité extra-linguistique. N' est toujours un prédicat mais il est maintenant lié à cette classe d'objets.

N" : L'opération de fermeture, au moyen d'un quantificateur, sélectionne un élément précis dans la classe précédente. Il y a maintenant référence à un objet de la réalité extra-linguistique.

Le modèle linguistique est en cours d'extension et doit permettre la prise en compte du verbe et des relations anaphoriques.[14, 15]

B. RECHERCHE D'INFORMATIONS

Le processus de recherche sera évidement centré sur la mise en œuvre du mécanisme de référence à la réalité extra-linguistique. Une analyse identique à celle du document permet d'identifier dans la question exprimée en langue naturelle un ou plusieurs syntagmes, leurs composantes et les relations qui les lient. Tous ces éléments sont autant de points d'entrée dans la base de données. L'hypothèse faite est qu'au moins l'un d'entre eux permet d'amorcer un cheminement dans la base en fournissant une liste de syntagmes nominaux ayant la même caractéristique (même composante impliquée dans une même relation). En fonction de leur valeur référentielle, les«bons syntagmes» sont identifiés par l'utilisateur. Celui*ci* peut alors utiliser d'autres composantes figurant dans ces syntagmes et auxquelles il n'avait pas pensé pour poser d'autres questions de plus en plus en accord avec ce qu'il veut et ainsi mieux cerner sa recherche.

C. INTÉGRATION DES DOCUMENTS DANS LA BASE DE DONNÉES

Le résultat de l'analyse du document n'est donc pas une simple liste de syntagmes nominaux. Les relations observées sont aussi mémorisées dans la base et servent au processus de recherche. Ce sont :

- les relations entre constituants d'un syntagme nominal, le noyau, les modifieurs et les emboîtements de syntagmes qui traduisent des relations d'hyponymie ou d'hyperonymie,
- les relations entre syntagmes par l'intermédiaire du verbe. L'hypothèse faite ici est que dans la majorité des cas le syntagme sujet est le thème de la proposition,
- les relations anaphoriques[16], principalement celles où il y a coréférence. On peut penser, en effet que le mécanisme de référence se manifeste ici de façon forte et que si des syntagmes sont plus que d'autres à retenir ce sont ceux qui sont impliqués dans ce type de relation.

V. CONCLUSION

L'approche linguistique des problèmes d'indexation automatique, en posant la question du statut linguistique du mot-clé, a montré l'importance du phénomène de référence à la réalité extra-linguistique. Elle oblige à reconsidérer les systèmes d'informations documentaires sous un autre angle que le perfectionnement du système de classification ou du thesaurus. Ce qui paraît indispensable, c'est de développer tous les outils facilitant la référence à la réalité extra-linguistique de l'utilisateur.

Notes

1. Le Guern (M.) Sur les relations entre terminologie et lexique. in «Deuxièmes entretiens du Centre Jacques Cartier», colloque sur les terminologies spécialisées, Département de linguistique et de philologie de l'université de Montréal, Canada, 13 et 14 octobre 1988.
2. Le groupe SYDO s'est constitué en 1975 et rassemble le Centre de Recherches Linguistiques et Sémiologiques de l'Université Auguste et Louis Lumière (Lyon 2) et le Laboratoire d'Informatique Documentaire de l'Université Claude Bernard (Lyon 1). La composante grenobloise SYDOG est représentée par le Centre de Recherches en Informatique pour les Sciences Sociales de l'Université de Grenoble 2.
3. Lainé (S.), Extraction et sélection de descripteurs complexes dans un ensemble de textes pour leur indexation automatique. in Thèse de doctorat-ingénieur en mathématiques (informatique), Université Claude Bernard-Lyon 1, 1982.
4. Les résultats exprimés en termes de bruit (nombre d'unités non pertinentes sélectionnées/nombre total d'unités sélectionnées) et de silence (nombre d'unités non sélectionnées mais pertinentes/nombre total d'unités sélectionnées) donnaient pour les mots isolés un bruit et un silence de l'ordre de 70%. Pour les expressions ces mêmes valeurs étaient de l'ordre de 20%.

5. Le Guern (M.) Les descripteurs d'un système documentaire: essai de définition. in Bès (G.C.), Fauchère (P.M.), Lagueunière (F.), actes du colloque«Traitement automatique des langues naturelles et systèmes documentaires», Condenser, supplément 1, Université de Clermont Ferrand, 1982.
6. Berrendonner (A.) Grammaire pour un analyseur: aspects morphologiques, in Document de travail du groupe SYDO.
Metzger (J.P.) Syntagmes nominaux et information textuelle-reconnaissance automatique et représentation. in Thèse de doctorat ès sciences Université Claude Bernard-Lyon 1, 1988.
7. Bouché (R.) Sciences de l'information : sciences de la mise en forme. in Infomédiatique, annales de l'École Nationale Supérieure de Bibliothécaires. Cercle de la librairie, Paris, 1988, pp. 11-18.
8. Quoi de plus arbitraire que la division systématique en 10 sous ensembles de la CDU ?
9. Maniez (J.) Les langages documentaires et classificatoires. Les éditions d'Organisation, Paris, 1987.
10. Bouché (R.) *Opus cit.*
11. On peut voir là une tentative bien rudimentaire de recherche de syntagmes nominaux.
12. Tague (J.), Schultz (R.) Some measures and procedures for evaluation of the user interface in an information retrieval system. in 11th international conference on research and development in information retrieval. Grenoble France june 13-15 1988. Presses universitaires de Grenoble.
13. Berrendonner (A.), *Opus cit.*
Metzger (J.P.), *Opus cit.*
14. Batache (M.), Contribution à l'analyse automatique du verbe français, in Mémoire de DEA Conception de Systèmes d'Informations Spécialisées, Universités Lumière-Lyon 2 et Claude Bernard Lyon 1.
15. Vidalenc (I.) Traitement automatique des anaphores pronominales en français, Thèse de l'Université Lumière-Lyon 2, à paraître en 1988.
16. Vidalenc (I.) *Opus cit.*