

Recherche d'une synergie entre développement linguistique informatisé et systèmes experts : importance de la terminologie

Richard Parent

Volume 34, numéro 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/003788ar>

DOI : <https://doi.org/10.7202/003788ar>

[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

Parent, R. (1989). Recherche d'une synergie entre développement linguistique informatisé et systèmes experts : importance de la terminologie. *Meta*, 34(3), 611-614. <https://doi.org/10.7202/003788ar>

RECHERCHE D'UNE SYNERGIE ENTRE DÉVELOPPEMENT LINGUISTIQUE INFORMATISÉ ET SYSTÈMES EXPERTS : IMPORTANCE DE LA TERMINOLOGIE

RICHARD PARENT
*Ministère des communications
Québec, Canada*

L'émergence des industries de la langue interpelle les chercheurs en linguistique pour un développement de leurs connaissances sous certains formats informatisés visant à un certain degré d'automatisme dans l'analyse du contenu de textes. À l'heure actuelle, cette évolution est encore jeune, mais nous verrons dans cet article comment des automatismes basés sur un faible niveau de connaissances linguistiques peuvent déjà être exploités de façon rentable.

Le développement linguistique informatisé fera appel aux recherches menées en intelligence artificielle particulièrement dans le domaine connu sous le nom de «traitement de langue naturelle». Un autre domaine de recherche en intelligence artificielle, celui des «systèmes experts», semble fortement lié à la langue puisqu'elle constitue le moyen principal d'expression des connaissances.

Pour développer un système expert, la tâche principale est d'acquérir les connaissances constituant l'expertise concernée. Où trouve-t-on cette connaissance? Auprès des experts bien sûr, mais encore? Au moyen d'entrevues longuement transcrites et analysées, constitue la réponse la plus courante. Le problème est que c'est une approche fort coûteuse du temps de l'expert, de l'analyste (ou cogniticien). Les textes déjà rédigés par les experts sont une source actuellement négligée de connaissances. Il s'agit encore du support le plus complet des connaissances, et elles sont accessibles à moindre coût. Leur mise en exploitation est une longue route, et voici une histoire de premiers pas dans la gestion de bases de données textuelles.

Nous insisterons dans ce cadre sur les rapports entre les méthodes d'analyse et terminologie et celles utilisées en acquisition et modélisation des connaissances.

MODÉLISATION DES CONNAISSANCES

Deux grandes approches sont utilisées en modélisation des connaissances. L'approche dominante est «déductive», c'est-à-dire que par entrevue (ou introspection) on cherche à découvrir les grandes notions de l'expertise. Cette approche correspond au travail onomasiologique du terminologue, soit la déduction à partir de connaissances générales et l'exploration interne à un groupe de notions. L'expérience révèle que cette approche est coûteuse et difficile, et qu'elle résulte généralement en un bricolage généralisé de vocabulaires fragiles.

L'autre approche utilisée en modélisation des connaissances est dite inductive. Elle consiste habituellement à fournir des exemples à un système qui les compile et procède à l'induction de connaissances rendant compte des variations observées dans le groupe

d'exemples. En pratique, cette méthode s'applique rarement parce que la circonscription du domaine est trop exigeante et limitative. Cette approche correspond au travail sémasiologique du terminologue. Une adaptation de cette approche inductive à l'exploitation de textes s'avère prometteuse : sous contrôle humain abstrait, un ensemble de données terminologiques et lexico-sémantiques peuvent être rendues directement utilisables dans l'élaboration d'un domaine à systématiser. Cet ensemble est alors collé aux usages courants et au vocabulaire familier pour l'utilisateur dans sa tâche.

Avant de présenter trois exemples d'application d'une approche plus sémasiologique dans le développement de systèmes d'information, une comparaison sera faite entre terminologie et systèmes experts.

TERMINOLOGIE ET SYSTÈMES EXPERTS

Les réalisations en terminologie et en systèmes experts s'articulent autour de structures notionnelles pour établir des distinctions entre concepts et classes de concepts au moyen de :

- ◆ la définition logique, basée sur les traits essentiels ;
- ◆ la description, basée sur les traits fréquents ;
- ◆ les schémas de relations opératoires entre notions (par exemple, un agent utilise un instrument pour transformer un objet).

Dans l'élaboration d'un système notionnel, le terminologue combine deux approches, onomasiologie et sémasiologie, dans un mouvement dialectique :

- ◆ en onomasiologie, il fait l'analyse des tâches ou des problèmes réels, il essaie d'identifier les concepts impliqués, et seulement ensuite, leurs correspondances lexicales ; c'est une sorte d'analyse conceptuelle épurée, l'étude d'une logique ;
- ◆ en sémasiologie, il étudie la langue dans des textes, il essaie d'identifier les concepts impliqués dans le lexique ; c'est une analyse de la langue guidée par des critères conceptuels, l'étude d'un vocabulaire.

Tel que déjà noté, à l'onomasiologie correspond l'approche déductive en construction de systèmes experts (au moyen d'un travail conceptuel sur des abstractions d'objets à propriétés évaluées et de règles «si... alors»), et à la sémasiologie correspond l'approche inductive reposant parfois sur la compilation d'une banque de cas, mais plus souvent, sur l'exploitation d'une base de données textuelles.

En terminologie et en systèmes experts, il y a un but normatif important d'uniformiser le vocabulaire à utiliser, ce but ayant une plus haute priorité en terminologie où le domaine de connaissances est généralement plus vaste. Dans le cas des systèmes experts, une norme déjà sanctionnée existe plus souvent, et la priorité va plutôt dans l'intégration du vocabulaire avec le monde cognitif de l'utilisateur, tel que déterminé par la langue, les exigences de la tâche, l'interface du système et les variations individuelles.

Le résultat en terminologie est une liste de termes sur fiches destinée à être consultée par des humains, alors qu'avec un système expert, le résultat est une base de connaissances utilisable par des «moteurs» ou (algorithmes complexes) d'inférence, d'analyse, ou d'appariement : ces connaissances sont, d'une certaine manière, actives en vertu des schémas de relations opératoires, des règles de production et d'autres mécanismes automatiques comme l'«héritage des propriétés» (au sens où un chêne hérite de propriétés générales d'un arbre).

Enfin, entre les deux termes de la présente comparaison, il y a des interdépendances :

- ◆ l'usage de systèmes experts en terminologie est illustré par la présentation, dans ce même colloque, d'un système d'aide au dépouillement terminologique (Pierre Plante et Jean Perron) ;

- ◆ dans la construction de systèmes experts, il y a lieu d'emprunter aux méthodes terminologiques afin de dépasser le bricolage courant de vocabulaire ; comme le dit Alain Rey, «La maîtrise du réel présuppose la maîtrise d'un ensemble de noms pour l'appréhender et le manipuler».

Trois exemples :

Divers projets pilotes menés au gouvernement du Québec peuvent servir à exemplifier ces hypothèses théoriques du rapprochement souhaitable entre terminologie et systèmes experts.

Ces exemples montrent bien la rentabilité d'une approche plus fortement sémasiologique dans le développement de systèmes experts, et en fait, de systèmes d'information en général. La réalisation de ces projets dépend d'une collaboration entre du personnel régulier du gouvernement et des chercheurs du Centre d'analyse de textes par ordinateur de l'Université du Québec à Montréal. Les logiciels utilisés (SATO et Déredec-EXPERT) ont été développés par ce centre de recherche universitaire.

a) Conseil du Trésor :

Une base de données textuelles a été créée qui comprend le «Répertoire des politiques administratives», composé de 145 documents et comptant 280 000 mots. L'exploitation de cette base de données textuelles a conduit à :

- ◆ créer des lexiques des expressions représentatives dans ce domaine de l'administration publique en vue d'indexer les nouveaux documents ;
- ◆ constituer un index des mots-clés, ou thésaurus, à partir d'un jugement humain sur les lexiques (choix de ce qui paraît pertinent, construction des expressions complexes) ;
- ◆ extraire des définitions dans le texte sur la base de régularités d'écriture ;
- ◆ générer des éditions spéciales en formats particuliers peu accessibles sur logiciels de traitement de textes en usage ;
- ◆ rechercher des concordances à diverses fins particulières d'analyses couramment effectuées lors de la modification d'un élément ou l'autre de la réglementation par exemple, ainsi que dans l'élaboration d'un système expert d'aide à l'attribution de contrats (ce qui est le volet de la réglementation qui occasionne le plus de consultations du Répertoire des politiques administratives).

b) Ministère de l'Environnement :

Pour accélérer la modélisation des connaissances d'un système expert en évaluation environnementale (ou «étude d'impacts»), une base de textes pertinents a été créée. Elle comprend 12 000 pages de rapports, avis, lois, analyses, lettres, comptes rendus, etc. L'analyse des concordances d'un concept, de son environnement textuel dans un tel corpus représentatif, a permis une productivité accrue et une meilleure validité dans la démarche de modélisation des connaissances. Les experts et cognitivistes progressent depuis avec plus d'assurance quant au caractère systématique de leur connaissance du domaine.

c) Gestion du personnel :

L'indexation de quelque 110 conventions collectives de travail, représentant plus de 20 000 pages de textes, a conduit, dans la même foulée, à développer une nouvelle méthodologie, plus sémasiologique, dans l'élaboration d'un thésaurus. La nouvelle façon de procéder à l'indexation, en procédant toujours en contexte grâce au mécanisme des concordances.

Ces trois exemples illustrent une nouvelle approche générale qui utilise ces pistes que sont les signes (mots) pour atteindre les concepts.

VERS UNE APPROCHE PLUS SÉMASIOLOGIQUE

Ces expériences montrent qu'il importe d'augmenter la conscience du niveau terminologique pour la conception et le développement de systèmes experts. Il est rentable de supporter une démarche d'analyse de textes au moyen d'outils simples déjà disponibles, avec des automatismes de niveau modeste laissant place à l'intervention du jugement humain en tout point crucial. Pour les textes volumineux, la connaissance du lexique est utile à la construction de thésaurus, à l'annotation de textes, à l'identification des concepts et relations en usage dans un domaine que l'on cherche à modéliser ou systématiser.

L'utilisation de ces techniques d'exploitation d'une base de données textuelles occasionnera peut-être des changements méthodologiques. L'économie relative entre onomasiologie et sémasiologie s'en trouve modifiée : en premier lieu, les coûts de sémasiologie sont largement abaissés, mais aussi dans un projet, le terminologue, l'indexeur, ou le cogniticien peut effectuer une itération beaucoup plus rapide dans son cycle naturel entre onomasiologie et sémasiologie : on peut se demander si cette rapidité accrue de répétition cyclique ne pourrait pas conduire à une meilleure intégration entre les deux ? Ne souhaite-t-on pas moins se creuser la tête sur trop peu d'évidence, en allant facilement vérifier plus fréquemment des questions qui surgissent en phase onomasiologique ?

L'exploitation de bases de données textuelles offre un potentiel élevé de contribution à la productivité dans le travail administratif, dans le travail des médias, en recherche, etc.