**Documentation et bibliothèques**

# Linguistique théorique et documentation automatique : l'exemple de la grammaire des cas

Yves Courrier

Volume 23, numéro 1, mars 1977

URI : https://id.erudit.org/iderudit/1055294ar
DOI : https://doi.org/10.7202/1055294ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

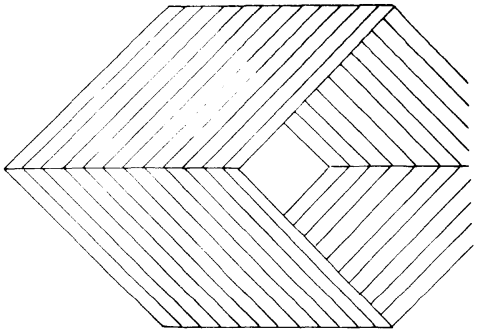
0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Citer cet article

Courrier, Y. (1977). Linguistique théorique et documentation automatique : l'exemple de la grammaire des cas. *Documentation et bibliothèques*, *23*(1), 39–42. https://doi.org/10.7202/1055294ar

# chronique de la recherche



## Linguistique théorique et documentation automatique: l'exemple de la grammaire des cas[1]

L'utilisation de l'ordinateur en documentation n'est certes pas nouvelle: dès la fin des années cinquante, la recherche en ce domaine commençait. Les tâches les plus faciles à automatiser furent rapidement circonscrites: stockage, recherche et diffusion de divers produits documentaires. Une seule étape — celle de l'analyse — résistait cependant à l'automatisation. En effet, il est facile de créer des fichiers, de les exploiter et de produire des bulletins ou des bibliographies variées, ces tâches exigeant de l'ordinateur des capacités pour lesquelles il est bien adapté et pour lesquelles il faut peu de préparation. L'analyse, au contraire, est une opération autrement complexe, si on la définit comme l'opération qui consiste à produire la représentation condensée d'un document. Elle est difficile à automatiser, car il faut rendre un ordinateur capable de simuler le travail d'un analyste qui résume ou indexe un texte en langage naturel. Les tentatives n'ont pourtant pas manqué. Elles peuvent être groupées en trois catégories, selon la méthode qu'elles privilégient. On trouve d'abord les méthodes statistiques où l'on suppose que de simples critères de fréquence suffisent à extraire d'un texte les principaux descripteurs. Malgré la très abondante documentation sur ce sujet, on n'a jamais abouti à des applications opérationnelles. La deuxième catégorie comprend les travaux inspirés des théories linguistiques: grammaires structurales ou transformationnelles ou analyse par chaînes. La troisième catégorie comprend les approches plus pragmatiques où on se préoccupe plus des résultats à atteindre et des contraintes de l'ordinateur que d'*a priori* théoriques.

Parallèlement à ces travaux sur la documentation automatique, la linguistique suivait sa propre évolution: après l'enthousiasme qui accompagna les premières expériences en traduction automatique, on s'est rendu compte que notre connaissance du langage était très sommaire et qu'il fallait développer les recherches théoriques. On a alors assisté à une effervescence remarquable chez les linguistes, facilement discernable dans la quantité — sinon la qualité des publications.

Ces deux courants de la recherche en documentation automatique et en linguistique théorique ont conflué au début de la présente décennie. On a vu la parution de nombreux ouvrages — et articles de périodiques importants — décrivant l'état de la question en linguistique automatique et l'applicabilité à la documentation.

Parmi les ouvrages, on peut citer Coyaud (1972), Spark-Jones et Kay (1973), Bély (1970), Young (1973). Parmi les articles les plus importants, on trouve Gardin (1973), Montgomery (1972). Ces auteurs s'accordent sur un certain nombre de points:

1 Yves Courrier, *Document Analysis, Verbs and Case Grammar*, Pittsburg, University of Pittsburg, 1976, 205 p. (Thèse présentée à la Graduate School of Library and Information Science, en vue de l'obtention du Ph.D., avril 1976).

- le problème de l'indexation automatique est un problème essentiellement linguistique: i.e. rendre un ordinateur capable de dériver d'un texte en langage naturel une représentation condensée est une tâche avant tout linguistique;
- cette tâche met en jeu deux catégories d'information: l'information syntaxique et l'information sémantique;
- la linguistique théorique actuelle propose plusieurs théories assez développées pour décrire la structure syntaxique des phrases;
- l'application sur ordinateur de ces théories est loin d'apporter des résultats satisfaisants: on est capable de réaliser l'analyse de phrases simples seulement. On est incapable d'aboutir à l'analyse syntaxique de phrases complexes et encore moins à l'analyse d'unités linguistiques plus grandes que la phrase;
- les théories linguistiques contemporaines sont tout à fait incapables de rendre compte de phénomènes sémantiques et, bien souvent, sont incompatibles avec certains phénomènes sémantiques connus;
- les différentes méthodes utilisées en linguistique automatique proposent des modèles assez voisins de structures sémantiques;
- ces modèles ne couvrent que des domaines très restreints du savoir et sont encore inapplicables dans la cadre de l'indexation automatique.

Si l'on cherche à tenir compte de ces différentes idées, on s'aperçoit que pour réaliser l'indexation automatique, on a besoin d'une théorie et d'un modèle qui intègrent l'information syntaxique et sémantique. Un tel modèle permettrait d'utiliser pendant l'analyse et de récupérer ensuite et l'information syntaxique et l'information sémantique, c'est-à-dire à la fois les concepts et leurs relations. La grammaire des cas, théorie proposée et rapidement développée vers la même époque, semblait le candidat idéal (Montgomery, 1972, p. 210).

On peut sommairement décrire la théorie de la façon suivante. Conformément aux grammaires transformationnelles, on distingue dans la langue la structure de surface et la structure profonde. La structure de surface est l'ensemble des éléments présents dans la chaîne parlée. La structure profonde est une structure abstraite qui sous-tend toute structure de surface. Pour Chomsky, la structure profonde est produite par un ensemble de règles syntaxiques, et l'interprétation sémantique permet d'y adjoindre les éléments lexicaux d'une langue. Pour les «sémanticiens génératives», on suppose que la structure profonde est avant tout faite d'éléments sémantiques.

La grammaire des cas, qui se place dans cette dernière catégorie, suppose que la structure sémantique est composée d'un prédicat et de plusieurs éléments qui y sont rattachés par des liens bien définis, les cas. Ainsi, la phrase «Paul a donné un livre à Jean» aurait la structure profonde suivante:

PRÉDICAT AGENT OBJET BÉNÉFICIAIRE
(donner) — Paul — Livre — Jean

On voit que le prédicat correspond le plus souvent au verbe de la structure de surface et que les cas ressemblent fort aux notions utilisées dans les grammaires des langues flexionnelles (latin, grec, allemand, russe, etc.) La distinction entre structure de surface et structure profonde permet de ne pas s'arrêter aux phénomènes superficiels par lesquels chaque langue marque les cas. Un premier problème pour la théorie consiste à déterminer la liste exacte des cas. Les différents auteurs ont proposé des listes assez courtes, incluant, outre les cas mentionnés, d'autres cas tels qu'INSTRUMENT, EXPÉRIENCEUR, TEMPS, LIEU, MANIÈRE, ACCOMPAGNEMENT. D'ailleurs, établir une liste n'est qu'une première étape, puisqu'il faut aussi trouver des critères linguistiques précis qui permettent de déterminer à coup sûr le cas d'un élément de la structure de surface.

De plus, une fois la liste des cas arrêtée, on s'aperçoit que les combinaisons possibles entre cas sont assez limitées. On peut ainsi déterminer des catégories de prédicats en fonction de leur structure casuelle. Par exemple, les verbes *prendre, donner, acheter* ont tous besoin des cas d'AGENT, BÉNÉFICIAIRE et OBJET. De même, les verbes *partir, rester, entrer* ont tous besoin d'un AGENT et d'un LIEU.

Outre sa simplicité, la théorie a semblé attirante pour les documentalistes en ce qu'elle s'accordait avec certaines données empiriques obtenues dans l'analyse automatique des textes (Bély et al, 1970; Earl, 1973; Young, 1973).

Cette conjonction d'un besoin pratique et d'une théorie prometteuse déterminait un champ d'investigation assez précis: l'utilisation de la grammaire des cas dans l'indexation automatique.

Le domaine d'investigation est une chose, la méthode en est une autre. L'attitude la plus courante consiste à emprunter à la théorie un cadre conceptuel assez vague et à essayer de développer un algorithme d'analyse automatique. C'est ce qui avait été fait pour les théories structurales et transformationnelles et ce que Young avait fait pour la grammaire des cas.

Malheureusement le plus souvent, on découvre très rapidement les limites du modèle, si bien qu'un auteur a écrit récemment qu'«on sait que tout système ou structure sera finalement insuffisant face au langage, si bien qu'il peut seulement être question de savoir à quel moment il devra être abandonné. La question intéressante... est de savoir jusqu'où il vaut la peine de pousser une approche structurale quelconque avant de recommencer à zéro» (Wilks, 1976, p. 40).

Par contre, puisque la théorie proposée était basée sur un nombre limité de phénomènes linguistiques, on pouvait plutôt l'examiner de plus près et la soumettre à l'épreuve d'un plus grand nombre de données. C'est l'option qui a été choisie dans cette étude.

Comme on l'a vu plus haut, deux aspects de la théorie méritaient d'être précisés; la liste des cas et les structures casuelles des verbes. Le meilleur test pour s'assurer qu'une taxonomie est bien faite est de demander à plusieurs individus de l'appliquer. S'ils aboutissent à des résultats concordants, c'est que la taxonomie est valable. S'ils n'y aboutissent pas, c'est que la taxonomie a été mal expliquée ou appliquée, ou encore qu'elle est insuffisante (pas assez claire ou précise). L'étude expérimentale a effectivement révélé des résultats très divergents, le pourcentage des erreurs dans l'utilisation de la taxonomie atteignant 38.87%.

Pour s'assurer que les erreurs n'étaient pas dues à l'explication ou à l'utilisation de la théorie, on peut faire appel à des méthodes subjectives et objectives. Les deux ont été exploitées et ont conduit à la même conclusion.

La méthode objective utilisait un questionnaire sur l'expérience et la formation des sujets. On étudiait ensuite les corrélations entre les données du questionnaire et les résultats dans l'utilisation de la taxonomie. Le résultat paradoxal a été que plus un sujet avait étudié la linguistique et les documents sur la grammaire des cas, plus il se trompait dans l'utilisation de la taxonomie. On voit donc assez mal comment des explications plus poussées auraient pu améliorer les résultats.

La méthode subjective consistait à demander aux sujets, par un questionnaire ouvert, où ils plaçaient la source des erreurs. L'unanimité a été frappante: la théorie n'est pas assez précise et l'utilisation de la taxonomie laisse trop de place aux interprétations subjectives. La conclusion s'imposait donc d'elle-même.

Quoiqu'ils puissent paraître décevants, ces résultats ne sont pas surprenants. On peut les expliquer assez facilement et en tirer des conclusions pratiques. L'explication du point de vue linguistique est très simple. La linguistique contemporaine recherche des modèles théoriques et ne se préoccupe pas assez de l'utilisation du langage. Depuis Chomsky, on veut s'occuper de la compétence du locuteur, c'est-à-dire des capacités par lesquelles un locuteur

(ou auditeur) abstrait est capable d'énoncer (ou de comprendre) des phrases valides dans une langue donnée. La linguistique n'est pas encore prête à intégrer dans ses théories les utilisations concrètes, les perceptions différentes du langage qu'ont les individus. On peut se demander si des tests tels que ceux qui ont été utilisés dans la présente étude ne devraient pas être plus souvent utilisés en linguistique, car non seulement ils rapprochent la théorie de la réalité, mais ils soulignent aussi les lacunes de la théorie et les aspects susceptibles d'être améliorés.

Du point de vue de la documentation, l'expérience incite aussi à la prudence. Même si une théorie rejoint superficiellement certaines données empiriques — il y a actuellement un grand nombre de théories linguistiques qui s'accordent toutes avec certaines données empiriques — elle n'est pas forcément le modèle idéal et définitif.

> "We believe that our study has shown that there is a great distance between the formulation of a linguistic theory and its application to the diversity of natural language. This is something which linguists are probably aware of, but which information scientists seem to overlook. On the other hand, it does not seem that linguists are very much concerned with the applicability of their theories. This mutual ignorance is not necessarily bad. Scientific research has to proceed methodically, and it ought not to be hindered by the attempt to find an immediate solution to every problem from every angle. But when interdisciplinary research brings together disciplines with the same concern, i.e. understanding of natural language, great caution must be exercised, for the distance between them may be larger than was expected beforehand" (p. 156-157).

Une telle étude ne signifie pas que la grammaire des cas n'est pas applicable à la documentation. En fait, certains concepts de la théorie sont évoqués dans des applications documentaires récentes telles que PRECIS (Austin, 1974) et TITUS (Ducrot, 1973). D'autres études sont en cours en linguistique automatique et les linguistes eux-mêmes continuent à explorer cette voie. Tout n'est pas encore dit, mais encore faut-il que ce qui reste à dire le soit correctement.

**Yves Courrier**
École de bibliothéconomie
Université de Montréal

## Bibliographie

Austin, D. *PRECIS. A Manual of Concept Analysis and Subject Indexing.* London, BNB, 1974.

Bély, N. et al. *Procédures d'analyse sémantique appliquée à la documentation scientifique.* Paris, Gauthier-Villars, 1970.

Coyaud, M. *Linguistique et documentation.* Paris, Larousse, 1972.

Ducrot, J.M. «Le système TITUS II», *Information et documentation,* no 4 (octobre 1973), 3-40.

Earl, L.L., "Use of word government in resolving syntactic and semantic ambiguities", *Information Storage and Retrieval,* vol. 9 (December 1973), 639-664.

Gardin, J.C. "Document analysis and linguistic theory", *The Journal of Documentation,* vol. 29 (June 1973), 137-168.

Montgomery, C.A. "Linguistics and information science", *Journal of the American Society for Information Science,* vol. 23 (May-June 1972), 195-219.

Sparck-Jones, K. and Kay, M. *Linguistics and Information Science,* New York, Academic Press, 1973.

Wilks, Y. *"Natural language understanding systems within the A.I. paradigm: a survey", American Journal of Computational Linguistics,* vol. 1 (1976). Microfiche no 40.

Young, C.E. *Development of Language Analysis Procedures with Application to Automatic Indexing.* Columbus, Ohio State University, 1973.