Meta
Journal des traducteurs
Translators' Journal

META

La terminotique et les industries de la langue

Pierre Auger

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/001922ar
DOI: https://doi.org/10.7202/001922ar

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

Auger, P. (1989). La terminotique et les industries de la langue. *Meta*, *34*(3), 450–456. https://doi.org/10.7202/001922ar

LA TERMINOTIQUE ET LES INDUSTRIES DE LA LANGUE

PIERRE AUGER
Université Laval, Québec, Canada

J'ai choisi de traiter de la terminotique (ou terminologie automatique), cette nouvelle composante lourde de la terminographie moderne, en la situant à la fois dans le courant de l'informatique linguistique contemporaine et des produits «industriels» qui sont en amont ou en dérivent, pour nous approcher du concept plus englobant d'Industries de la langue (I.D.L.L.).

Je m'arrêterai d'abord à présenter brièvement le concept très large que recouvrent les I.D.L.L. Ce concept a été défini lors du Premier sommet de la francophonie de Paris (1986), de la façon suivante: «[Les industries de la langue sont celles qui] fabriquent et commercialisent des automates qui manipulent interprètent, génèrent le langage humain, aussi bien sous sa forme écrite que sous sa forme parlée» (Rapport de synthèse: Industries de la langue 1986: 86). C'est aux Français Bernard Cassen et Jean-François Dégremont que l'on doit les premières définitions du concept (Cassen: 1986, Cassen et Dégremont: 1986). De façon plus précise, on peut encore définir les industries de la langue comme «des activités de développement, de production et de commercialisation des nouvelles technologies de l'information (NTI) qui font appel à la fois à l'informatique (ordinateurs et logiciels) et aux résultats de l'étude systématique des langues. Elles développent des produits (machines) capables de traiter des informations linguistiques et susceptibles de communiquer ces informations entre eux et également avec les humains.» (Rapport de synthèse: Industries de la langue 1986: 86). Le linguiste André Abbou, directeur adjoint du Réseau des industries de la langue, nous en donne une définition plus synthétique dans son ouvrage récent sur les Industries de la langue (Les applications industrielles du traitement de la langue par les machines): «Nous désignerons ici par industries de la langue tous les produits, techniques, activités ou services qui appellent un traitement automatique de la langue naturelle» (Abbou 1987: 30).

Parlant des I.D.L.L., William Baranes (1986: 74) écrit que «(Or) cette manipulation «machinique» [l'auteur fait reférence au texte sous forme électronique] nécessite des produits industriels. On assiste donc à la naissance de véritables «industries de la langue», regroupant d'une part les artisanats traditionnels du langage (traduction, dictionnaires, etc.) qui s'industrialisent grâce aux moyens informatiques, et, sous la poussée de la demande, de l'autre part, un ensemble d'activités nouvelles, au confluent de l'informatique et de la linguistique (traitement automatique du langage naturel, synthèse et reconnaissance de la parole, etc.)».

Les produits issus des industries de la langue peuvent commodément être classés en trois grandes catégories (Auger: 1988):

a) Les outils de développement de la langue

Il s'agit essentiellement d'outils internes de recherche et de développement qui sont mis à la disposition des linguistes, des terminologues et des traducteurs, et qui sont nécessaires à la modernisation de la langue et à sa nécessaire adaptation à l'évolution technologique. On peut classer dans cette catégorie les travaux de recherhe linguistique qui

comprennent tous les instruments physiques, conceptuels (algorithmes) ou les logiciels utiles aux recherches terminologiques, linguistiques et sémantiques : analyse de texte, reconnaissance de termes complexes, indexation, modélisation de processus, etc. On peut rattacher à ce groupe les Réseaux de terminologie et les Banques de données linguistiques et terminologiques qui sont d'ores et déjà des outils très performants d'accès à l'information de type linguistique.

b) Les outils d'utilisation de la connaissance linguistique

On situe dans cette catégorie les outils qui permettent de traduire dans des applications de la vie courante les résultats de la recherche dans les différentes sphères de l'informatique linguistique. Les systèmes de traduction automatique sont à classer dans cette seconde catégorie. Ils constituent un des domaines les plus anciens de la recherche en intelligence artificielle et ce n'est que maintenant, tout juste après trente ans de recherche, que l'on commence à obtenir des résultats intéressants dans ce domaine. On peut encore citer, dans la même catégorie, les banques de données textuelles, les systèmes d'interprétation du langage naturel, les systèmes de contrôle vocal pour les machines, les systèmes experts et les didacticiels.

c) Les outils de réalisation de produits linguistiques

La quasi totalité des textes produits aujourd'hui transitent, à un moment donné de leur existence, par un ordinateur. C'est donc dire l'importance que prend aujourd'hui l'informatique à orientation textuelle. Parmi ces produits qui visent la réalisation de produits linguistiques ou langagiers (le texte, sous toutes ses formes), mentionnons les logiciels de traitement de texte, les progiciels grammaticaux ou les correcteurs orthographiques et stylistiques, les logiciels d'éditique, les logiciels d'indexation ou de génération de textes, les dictionnaires et les thésaurus sur disque compact (CD-ROM) pour ne mentionner que les produits les plus courants. On peut donc voir l'importance de IDLL comme domaine de recherche dans le monde contemporain, domaine carrefour pour tout le progrès concernant les sciences de l'information et les nouvelles technologies qui sont mises en œuvre (N.T.I. = Nouvelles technologies de l'information).

La terminotique devient ainsi un concept de recherche particulier de l'informatique linguistique, celui du traitement automatique du terme, mais surtout un concept pragmatique étroitement relié à ce qu'on appelle aujourd'hui l'industrialisation de la langue, un concept générateur de produits performants de traitement de l'écrit.

Avant d'aborder pareil sujet, il convient de dire que le concept de terminotique repose avant tout sur les progrès récents accomplis par l'informatique d'orientation textuelle depuis dix ou quinze ans. Conçus à l'origine pour le traitement de données numériques, les ordinateurs se sont peu à peu adaptés au traitement de données textuelles (les caractères). Une deuxième constatation s'impose également qui nous fait observer que l'on doit beaucoup à la sociabilisation de l'informatique et à sa portabilité, qui ont permis à des linguistes ou des langagiers (comme on les appelle maintenant au Québec) d'accéder à l'informatique et d'utiliser directement ce moyen pour leurs recherches et même de l'adapter plus efficacement à leurs besoins. Ces deux constatations constituent en quelque sorte le fondement historique du vaste domaine de l'informatique langagière.

Pour conclure cette introduction, disons également que la terminologie, comme de nombreuses disciplines reposant sur le traitement de l'information écrite a évolué naturellement dans le sens d'un rapprochement constant avec l'informatique et ses moyens. En effet, la terminographie (comme la lexicographie) repose sur un traitement extensif et intensif de l'information textuelle : — extensif parce qu'elle traite du texte sous toutes ses

formes, prenons pour exemple la diversité des informations nécessaires à l'accomplissement d'un travail terminologique et les tâches de traitement qui s'y rattachent; — intensif: songeons seulement au volume considérable de cette documentation et aux lourdes tâches, souvent répétitives, qui attendent le terminologue dans sa recherche, on se rapproche ici sensiblement d'un idéal méthodologique d'exhaustivité que permet seul l'automate (Paradis et Auger: 1987, de Schaetzen: 1987).

Si l'on examine de près le processus méthodologique lié à la confection d'un dictionnaire terminologique du début des travaux jusqu'à sa mise en marché, quatre grandes catégories de tâches peuvent être distinguées: 1) l'accomplissement de tâches préliminaires qui sont avant tout d'ordre documentaire (exploration du domaine, sélection et enregistrement des sources) 2) l'accomplissement de tâches proprement terminographiques liées au traitement d'écrits spécialisés (mise en forme et traitement d'un corpus de textes), 3) le stockage et le traitement des données recueillies (structuration de bases de données et utilisation de bases de données et 4) l'édition et la diffusion du produit final sous diverses formes. Il est maintenant facile d'imaginer que chacune de ces quatre catégories de tâches traditionnellement conduites de main d'homme peut, séparemment, faire l'objet d'un traitement automatique, et cela à des degrés divers d'automatisation. Une projection vers un avenir peut-être pas très lointain permet d'imaginer sans doute l'automatisation de l'ensemble de la chaîne de traitement terminographique jusqu'à l'obtention du produit final, quelque soit sa forme (imprimée ou électronique). Voyons maintenant ce que peut apporter l'informatique comme assistance aux tâches terminographiques.

A. L'INFORMATIQUE COMME SOUTIEN AUX TACHES TERMINOGRAPHIQUES

A.1 La phase documentaire: On peut avoir recours à la télématique pour la téléconsultation et le téléchargement à partir de banques de données bibliographiques, textuelles et terminologiques, ce qui peut raccourcir de beaucoup l'étape de l'établissement du corpus de travail (sources, documents de reférence, éléments de connaissance etc.).

A.2 La mise en forme et le traitement du corpus de textes spécialisés (sur deux langues, par exemple):

- le téléchargement ou saisie optique ou manuelle pour constituer le corpus écrit;
- la mise en forme du corpus saisi par un logiciel de traitement de texte;
- la lecture et le traitement du fichier-texte par un logiciel de découpage de mots/termes et de termes complexes qui permet l'établissement automatique ou semi-automatique (mode interactif) de la nomenclature);
- la lecture et le traitement par un logiciel d'indexation (découpage des contextes, comptage des formes, fréquence, repérage des descripteurs);
- le découpage simultané des contextes;
- l'analyse des contextes en regard des descripteurs pour la rédaction assistée des définitions.

A.3 L'élaboration d'une base de données:

- la structuration et le stockage de l'information avec un s.g.b.d;
- le traitement de l'information (ajouts, retraits, modifications, tris);
- la fusion des bases de données, anglaise et française, par exemple, à partir d'un thésaurus;
- l'édition de la base de données et la mise en forme des données d'un fichier-texte.

A.4 L'édition de la base de données :

- le transfert sur un système d'éditique pour en arriver à un prêt-à-publier électronique ;
- la publication sous différents supports (disquettes, CD-ROM, vidéo-disques, WORM, imprimés divers).

La liste un peu longue qui vient d'être livrée, sans être très affinée du point de vue du traitement informatique, suggère qu'à toutes les étapes du travail terminologique, une forme d'automatisation, variable en intensité, est possible en recourant ou non à l'interface humain. L'idéal serait la réunion en un progiciel facilement utilisable d'un ensemble de programmes-outils pour l'automatisation de chacune des phases de travail. Certains progiciels de terminotique ont été développés ces dernières années, qui permettent de constituer des bases de données terminologiques relationelles et leur édition et qui, de plus, sont adaptés au traitement terminologique systématique en tenant compte de la structuration notionelle du domaine traité. Signalons, parmi les plus connus, le système BATEM de J. Baudot en développement à l'Université de Montréal et le système MICROCEZEAU de Cézeauterm (Université de Clermont-Ferrand) qui fonctionnent sur la norme IBM-PC. Le dernier est disponible commercialement en plusieurs versions, la plus récente étant programmée en langage DBASE3+. Notons toutefois que le dernier progiciel est avant tout un s.g.b.d. adapté au traitement terminologique et qu'il ne permet pas l'automatisation complète de la chaîne de travail terminologique.

B. LES PROBLEMES RENCONTRÉS (CF. SCÉNARIO EXPOSÉ EN A)

B.1 PROBLEMES GÉNÉRAUX

Si les outils informatiques d'analyse et de traitement de l'écrit (spécialisé, dans le cas qui nous intéresse) existent aujourd'hui (p.ex. Ink Tools, Wordcruncher, Alps Tools qui sont des outils commercialisés, etc.), ou ils ne sont pas assemblés en progiciels complets, ou ils sont trop difficiles d'utilisation pour les non-initiés (songeons aux lemmatiseurs en particulier et aux divers logiciels de découpage ou d'indexation de mots), de plus, ils nécessitent encore, pour une bonne part, une intervention humaine constante pour valider les choix. Prenons par exemple les découpeurs de termes complexes qui ne peuvent fonctionner en mode complètement autonome (i.e. sans intervention humaine) et qui reposent sur l'élaboration et l'enrichissement constant d'un dictionnaire de reférence où les choix sont fournis à la machine par l'opérateur. Ces outils sont également diversement performants quant à leur aptitude à reconnaître, traiter ou analyser la chaîne de caractères du français ou encore sa syntaxe. Problème plus général encore, ces progiciels construits autour de s.g.b.d. ne fonctionnent pas en langage naturel et nécessitent pour l'opérateur l'utilisation d'une syntaxe souvent très complexe pour effectuer des tâches spéciales, d'où la nécessité de recourir à l'élaboration de menus fastidieux et de routines d'automatisation.

Comme remède à ces problèmes généraux, il faut d'abord compter sur l'«industrialisation du français» et le développement de sa capacité de dialoguer avec les ordinateurs (c'est déjà fait pour la langue anglaise) et par ricochet, le développement d'une informatique en langue française naturelle. Songeons, à titre d'exemple, aux limites d'accès à l'information dans les banques de terminologie, limites imposées par leur langage d'interrogation, à leur peu d' «intelligence» à répondre aux besoins des utilisateurs. L'avenir en ce domaine réside très certainement dans le développement de progiciels reposant sur l'intelligence artificielle fondant des systèmes experts langagiers et capables de manipuler efficacement le langage humain. Les systèmes de T.A.O., à titre d'exemple, se situent très exactement dans cette problématique de développement.

B.2 PROBLEMES SPÉCIFIQUES :

Le problème le plus sérieux, à notre avis, est la disponibilité des textes spécialisés en version électronique. Les systèmes de lecture optique pour la saisie automatique des textes écrits ne comblent que de façon très partielle cette lacune. Les systèmes de saisie optique de caractères se situant dans une gamme de prix abordable sont souvent limités quant à la variété de caractères qu'ils peuvent reconnaître ou sont paresseux pour «apprendre» ou sont peu fiables quant au taux d'erreur qu'ils produisent. De toute façon, les textes saisis par lecture optique nécessitent dans les phases ultérieures de traitement de nombreuses manipulations (suppressions de toutes sortes, formatage, encodage) difficiles à éxécuter par des novices et qui sont coûteuses en temps. C'est ici que des bases de données textuelles réunissant des textes spécialisés récents en langue française et couvrant un large éventail de domaines deviendraient des outils de travail extraordinairement puissants pour le terminoticien. On pourrait alors songer au téléchargement à partir de ces banques et même arriver à constituer des corpus de de dépouillement équilibrés et représentatifs.

D'un autre point de vue, on pourrait songer à utiliser de la même façon les dictionnaires sous leur forme électronique (tous les dictionnaires importants existent sous cette forme avant d'être publiés), ces «hypertextes» traités par des logiciels d'indexation pourraient se révèler être de précieux aides pour le terminologue (recherches de contextes, rédaction de définitions etc.). À plus forte raison encore, les dictionnaires et les encyclopédies existant sous forme de CD-ROM (malheureusement presqu'exclusivement en langue anglaise). Enfin, il y a les banques de terminologie qui, pour la plupart, offrent des facilités pour constituer des fichiers à partir du matériel extrait lors de l'interrogation. Pour tous ces exemples toutefois, la question de propriété intellectuelle et celle des droits d'auteur constitue un problème non encore résolu.

C. PROSPECTIVES ET VOIES D'AVENIR.

Faisons maintenant table rase des problèmes que nous venons d'identifier et considérons qu'ils sont d'ores et déjà réglés. Dans ce scénario futuriste, le terminologue a accès à de gigantesques bases de données (ou de connaissances) textuelles, il extrait par télé-chargement les éléments de son corpus de ces bases de données, le fait dépouiller automatiquement sans avoir eu à saisir le texte manuellement au préalable, établit automatiquement sa nomenclature de travail, fait découper les termes-entrées par la machine, en extrait des descripteurs sémantiques qui serviront ultérieurement à rédiger des définitions en mode assisté, classe, trie, fusionne les bases de données et les édite avec un minimum d'intervention de sa part. Son poste de travail multi-tâches lui permet tout en poursuivant son travail d'accéder instantanément à des banques de données terminologiques ou documentaires, d'échanger avec ces systèmes et de télécharger de l'information dans son propre système de terminotique qu'elle soit textuelle ou graphique (une illustration, par exemple), d'utiliser une abondante documentation sur CD-ROM, enfin, doté de fonctions bureautiques avancées «intelligentes» , ce poste de travail lui permet de contrôler lui-même et en tout temps l'élaboration de son produit et de le mener à terme dans les meilleures conditions. Nous avons là, en fait, toutes les composantes d'un système expert de terminotique capable de prendre lui-même certaines décisions et de les faire corroborer au besoin par le terminologue, qui se réserve alors le rôle valorisant de contrôler les faits et gestes de son automate. De plus, les problèmes de mise à jour des dictionnaires terminologiques sont aplanis dans un tel scénario. Cet horizon idyllique est peut-être plus près de nous que nous ne l'imaginons, attendons ce que nous réservent les prochaines années en ce domaine.

Pour donner un aperçu des nouveaux outils de traitement du langage que l'informatique met déjà à notre disposition et qui seront perfectionnés encore dans les prochaines années pour servir le terminologue, mentionnons :

- Les systèmes intelligents de reconnaissance de caractères ;
- Les analyseurs syntaxiques (parsers) et les lemmatiseurs capables de décortiquer un texte en unités et de les classer grammaticalement ;
- Les découpeurs de mots / termes et de termes complexes ;
- Les analyseurs sémantiques ;
- Les logiciels d'indexation ;
- Les logiciels d'interrogation en langue naturelle ;
- Les nouvelles technologies de diffusion de l'information terminologique (CD-ROM, vidéo-disques, disques WORM etc.) ;
- et enfin, des banques de terminologie «intelligentes».

En terminant cet exposé, j'aimerais insister sur la nécessité qu'il y a pour les langagiers de notre génération de se frotter à toute cette nouvelle technologie d'industrialisation de la langue. Les développements en ce domaine nous concernent tous, que nous soyons des concepteurs ou des utilisateurs de l'informatique. Régulièrement, à l'occasion de colloques réunissant des professionnels de l'information (traducteurs, rédacteurs, linguistes etc.), des tenants de la productivité reviennent sans cesse à la charge pour affirmer que la rentabilité est souvent le plus grand obstacle au développement de services linguistiques. Nous avons ici, je pense, une solution dans une intégration forte des activités langagières à l'informatique et les moyens nouveaux qu'elle met à notre disposition. Conscient des besoins urgents des professions langagières en ce domaine, le département de langues et linguistique de l'Université Laval a choisi de former ses étudiants linguistes, terminologues et traducteurs, selon cette orientation. Les cursus prévoient pour chaque programme une initiation poussée à l'informatique de texte, et à l'informatique linguistique dans certains cas, selon une approche très pratique. Ce virage de notre institution a été rendu possible grâce à l'accord IBM/Laval conclu en 1985. Cette orientation sera maintenue et même intensifiée au-delà de la fin de l'accord prévu pour le mois de décembre 1988. Notre objectif ultime est de former de futurs praticiens langagiers prêts pour le marché du travail et capables de manipuler l'informatique pour leurs besoins professionnels. Il s'agit également d'une orientation fondamentale nouvelle de la recherche en linguistique à Laval, entérinée par le CIRB en 1987, qui entend bien contribuer au développement des Industries de la langue au Québec tout en rentabilisant la recherche universitaire.

RÉFÉRENCES

ABBOU, André, MEYER, Thierry, LEFAUCHER, Isabelle (1987) : *Les industries de la langue. Les applications industrielles du traitement de la langue par les machines*, Paris, Éd. Daicadif, 400 p.

AUGER, Pierre (1988) : «L'industrialisation de la langue française et son maintien comme grande langue véhiculaire de la science et de la technique», 56e Congrès de l'ACFAS, Moncton, 9 — 13 mai 1988, s.p. [texte dactylographié].

BARANES, William. «Les industries de la langue», *Qui vive international*, n° 4, pp. 74-75.

CASSEN, Bernard (1986) : «Un nouveau front pour le français et la langue de l'Europe, LES INDUSTRIES DE LA LANGUE. ENJEUX POUR L'EUROPE», Actes du colloque de Tours, Tours, 28 fév. — 1er mars 1986, Encrages, n° 16, nov. 1986, pp. 12-14.

CASSEN, Bernard, DEGREMONT, Jean-François (1986) : «Bilan de la Mission «industries de la langue» au 31 juillet 1986, LES INDUSTRIES DE LA LANGUE. ENJEUX POUR L'EUROPE. Actes du colloque de Tours, Tours, 28 fév. — 1er mars 1986, *Encrages*, n° 16, nov. 1986, pp. 148-151.

DEGREMONT, Jean-François (1986) : «Au croisement de l'informatique et de la linguistique, LES INDUSTRIES DE LA LANGUE. ENJEUX POUR L'EUROPE. Actes du colloque de Tours, Tours, 28 fév. — 1er mars 1986, *Encrages*, n°16, nov. 1986, pp. 22-46.

DE SCHAETZEN, Caroline (1987) : «S.g.b.d. et terminologie», *Le linguiste — De Taalkundige*, vol. 33-3.
«Document de synthèse : Industries de la langue, Sommet de Québec. Deuxième Conférence des chefs d'État et de Gouvernement des pays ayant en commun l'usage du français», Québec, 2 — 4 sept. 1987, *Documents de conférence*, pp. 173-180.