Meta
Journal des traducteurs
Translators' Journal

META

Automatisation des procédures de travail en terminographie

Pierre Auger, Patrick Drouin et Marie-Claude L'Homme

Volume 36, numéro 1, mars 1991

La terminologie dans le monde : orientations et recherches

URI : https://id.erudit.org/iderudit/001921ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)

Découvrir la revue

Citer cet article

Auger, P., Drouin, P. & L'Homme, M.-C. (1991). Automatisation des procédures de travail en terminographie. *Meta*, *36*(1), 121–127.

AUTOMATISATION DES PROCÉDURES DE TRAVAIL EN TERMINOGRAPHIE[1]

PIERRE AUGER, PATRICK DROUIN ET MARIE-CLAUDE L'HOMME
Université Laval, Québec, Canada

On parle beaucoup, ces dernières années du rôle de plus en plus important que joue l'informatique dans le développement des sciences humaines. Certaines disciplines, plus que d'autres, ont profité de ce moyen de traitement puissant; il s'agit, entre autres, de tout le domaine des sciences du langage et de ses diverses applications qui font appel à la manipulation importante de données. La lexicographie et la terminographie sont à compter parmi ces applications qui ont bénéficié du traitement automatique des données que permettent les ordinateurs. La rencontre de la terminologie appliquée et de l'informatique a créé un champ nouveau de travail qu'on a dénommé la *terminotique*. Fait curieux, la terminologie a été l'une des premières disciplines de la linguistique à avoir utilisé activement l'électronique pour diffuser ces données avec les banques de terminologie, mais elle a tardé à tirer profit de l'informatique pour d'autres types de traitement qui concernent en particulier la chaîne de traitement des recherches terminologiques et terminographiques. Ce n'est qu'au début des années 1980, avec l'implantation et la diffusion de la micro-informatique, qu'on peut retrouver des articles portant sur l'automatisation en ce domaine et qu'on assiste à la naissance de termes comme : *terminologie informatique*, *terminologie assistée par ordinateur* et *terminotique*.

On commence à s'intéresser, avec l'avènement et l'utilisation de plus en plus massive de la micro-informatique, à des aspects divers du travail du terminologue : l'amélioration de banques de terminologie et l'adaptation de celles-ci à la micro-informatique, la création de mini-banques de terminologie, l'utilisation d'outils logiciels existants pour assister le terminologue (et le traducteur), le traitement automatique des nomenclatures et l'informatisation de l'ensemble de la chaîne de travail en terminographie sont devenus des axes importants de recherche.

Ainsi, le terme *terminotique* peut prendre toutes sortes de significations selon l'angle sous lequel on l'aborde, mais il semble qu'actuellement il concerne davantage la partie «appliquée» de la terminologie, c'est-à-dire les étapes de la recherche terminographique, de la collecte de la documentation à la diffusion des dossiers. Voilà précisément ce qui fera l'objet de notre article.

Nous présenterons plus particulièrement un projet d'automatisation des tâches du terminologue en cours actuellement à l'Université Laval. Nous tenterons de présenter les réalisations concrètes à ce jour, en présentant d'abord des exemples et en donnant un aperçu des améliorations possibles ou des pistes les plus prometteuses en ce domaine.

UN PROJET D'AUTOMATISATION DES TÂCHES TERMINOGRAPHIQUES

Les tâches du terminologue, lorsqu'il effectue une recherche terminographique, consistent, en gros, à sélectionner de la documentation (recherche de documents dans un domaine précis, choix de documentation intéressante, familiarisation avec le domaine), à faire le dépouillement de cette documentation (repérage de termes simples ou complexes, choix de contextes), à faire le traitement des termes sélectionnés (analyse des notions,

traitement des synonymes, rédaction des définitions, etc.) et, enfin, à constituer des dossiers terminographiques prêts à être diffusés.

La diffusion des dossiers terminographiques (ou fiches) est depuis longtemps automatisée. En effet, la création des premières banques de données terminologiques remonte au début des années 1960. Cependant, les autres tâches sont encore de nos jours le plus souvent accomplies de façon manuelle par le terminologue, ce qui représente un travail long et fastidieux.

La terminologie, comme discipline appartenant aux sciences de l'information, peut tirer profit des développements de la micro-informatique et notamment des systèmes de gestion de bases de données conventionnels et des nouveaux outils «commerciaux» de traitement du texte (cf. s.g.b.d.-texte). En effet, le traitement terminographique porte essentiellement sur des *données* terminologiques et des *textes* techniques et scientifiques.

L'objectif du projet d'automatisation de la chaîne de travail du terminologue vise, comme son titre l'indique, à faire accomplir, de façon automatique ou interactive, l'ensemble des tâches terminographiques, par un système informatique. Les personnes associées au projet n'ont pas cherché au départ à créer un environnement de travail intégré mais ont plutôt utilisé des outils existants en les adaptant aux besoins de la recherche terminographique. En fait, le projet a privilégié une approche étapiste et graduelle d'automatisation de la chaîne de travail du terminologue.

Une partie du travail d'automatisation a déjà été accomplie dans le cadre du projet. À l'aide d'un corpus textuel portant sur le domaine des industries de la langue[2], on a procédé à l'informatisation de certaines étapes de la recherche terminographique. Même si toutes les étapes de la recherche terminographique n'ont pas encore été informatisées intégralement, faute de temps, un fait encourageant s'est dégagé : on a constaté que toutes les étapes peuvent faire l'objet d'une automatisation.

Nous avons regroupé ci-dessous les étapes déjà automatisées ou en voie de l'être en les faisant correspondre aux étapes de travail traditionnel.

Méthodes traditionnelles	**Méthodes informatiques**
Recherche documentaire	
Sélection optique des documents à dépouiller	1. Le téléchargement ou saisie optique ou manuelle pour constituer le corpus écrit. 2. La mise en forme du corpus par un logiciel de traitement de texte.
Dépouillement	
Choix provisoire de termes	3. La lecture et le traitement du fichier-texte par un logiciel de découpage de mots/termes. 4. La recomposition des termes complexes et de formes composées
Choix définitif de termes	
Établissement de la nomenclature	5. Le marquage des termes à conserver pour la suite de la recherche.

Cueillette de contextes	6. La lecture et le traitement des textes par un logiciel d'indexation, établissement de concordances.
Découpage et sélection de contextes	7. Le découpage simultané des contextes. Délimitation des contextes. 8. L'importation des contextes dans un S.G.B.D.
Traitement	
Traitement des contextes	9. L'analyse sémantique des contextes à l'aide d'un système de gestion des contextes.

Étapes de la recherche terminographique : degré d'automatisation

1. Le téléchargement ou saisie optique ou manuelle pour constituer le corpus écrit.
2. La mise en forme du corpus par un logiciel de traitement de texte.
   (Les étapes 1 et 2 n'ont pas été considérées pour le projet puisque les textes utilisés étaient déjà en format ASCII. Cependant, pour pouvoir mener à bien toutes les étapes subséquentes, une saisie de tous les types de textes est indispensable. Par ailleurs, les logiciels de reconnaissance de caractères fournis avec les lecteurs optiques ne donnent pas, actuellement, toujours des résultats satisfaisants.) Ces deux étapes correspondent à la phase de recherche documentaire dans les méthodologies traditionnelles.
3. La lecture et le traitement du fichier-texte par des logiciels de découpage de mots/termes ou segmenteurs (étape qui correspond à la phase de dépouillement des méthodes traditionnelles).

3.1 Les textes sont, dans un premier temps, balayés par un logiciel de découpage de mots (DAT), un segmenteur primaire[3]. DAT découpe automatiquement en mots toutes les formes comprises entre blancs typographiques, précédées ou suivies par un signe de ponctuation ou un signe diacritique. DAT offre également la possibilité de rattacher les mots découpés à une référence dans le texte comme le numéro de ligne, de paragraphe, etc.

Il résulte de ce premier découpage, une liste de formes, celles qui sont contenues dans le texte.

Mot	**Références**
le	2 2
découpage	2 3
automatique	2 4
de	2 5
textes	2 6
en	2 7
unités	2 8
lexicales	2 9

3.2 La recomposition des termes complexes et de formes composées: tâche accomplie en mode interactif avec SYREX.

Cette deuxième étape consiste à recomposer certains mots découpés par DAT à l'aide d'un logiciel de recomposition des mots/termes composés (SYREX). L'étape

de recomposition paraît assez fastidieuse, parce qu'elle se fait en mode interactif, mais elle est inévitable. Les textes se composent d'un ensemble de mots et de termes complexes que DAT ne peut reconnaître en un premier balayage. SYREX fait défiler, dans une fenêtre, le texte à traiter. Il s'agit, pour l'utilisateur, d'indiquer à SYREX quelles sont les formes complexes qui constituent des termes. Il résulte, de cette étape, deux dictionnaires, un dictionnaire de mots simples (ou de termes simples) et un dictionnaire de mots complexes (ou termes complexes). Ces dictionnaires peuvent être utilisés pour faire effectuer une reconnaissance, automatique cette fois, des termes dans d'autres textes. Le système affichera les termes qui n'auront pas été reconnus.

Dictionnaire de termes simples

auteur
université
laval
de
textes
résumé

Dictionnaire de mots complexes

découpage automatique
unités lexicales
processus de traitement
langues naturelles
information textuelle

4. Le marquage des termes à conserver pour la suite de la recherche ou épuration des dictionnaires (étape qui correspond à l'établissement de la nomenclature): l'utilisateur épure les dictionnaires constitués par SYREX en marquant les formes qu'il souhaite traiter durant les étapes ultérieures. Il peut marquer les formes avec un traitement de texte. Dans le cadre du projet, les formes à conserver ont été marquées à l'aide d'un système de gestion de bases de données textuelles (SATO). Le dictionnaire de formes marquées se présente comme suit:

alphabet	**fréq**	**tot**	**lex**
fr	1	non	accessibilité
fr	29	oui	acquisition
fr	1	oui	adaptation-de-la-langue
fr	1	oui	analyse-de-textes
fr	1	oui	analyse-morphologique
fr	2	possible	assistance
fr	1	oui	assisté

ctl. d'écran : PgDn End —> <— > < Ins

5. La lecture et le traitement par un logiciel de gestion de bases de données textuelles: cette étape est réalisée avec SATO (système de gestion de bases de données textuelles). Il est possible de faire effectuer des recherches dans les textes à partir des dictionnaires constitués avec SYREX. SATO établit des relations entre les formes du dictionnaire et les formes du texte et réalise ce qu'on appelle des concordances. Cette étape correspond en gros au relevé systématique de contextes pour chacun des termes à traiter.

L'utilisateur effectue des recherches dans les textes à dépouiller à l'aide du dictionnaire de formes marquées. Le système recherche dans un ou plusieurs textes toutes les occurrences du terme demandé par l'utilisateur. Il effectue ainsi des

concordances, ce qui donne le résultat suivant:
Concordance pour analyse textuelle
analyse-textuelle
8 *PAGE=17/1/59/7 ... *PAGE=17/1/65/5
travaux effectués à Ottawa et à l'Université de Montréal sur la traduction-assistée-par-ordinateur, les banques-de-terminologie de l'Olf et du Secrétariat d'État, l'Eao, l'**analyse-textuelle**, la synthèse-de-la-parole ont dominé les dernières décennies. Au niveau international, les spécialistes de l'informatique-linguistique ont trouvé une tribune grâce à un organisme désigné par le sigle

6. Le découpage simultané des contextes: chacune des concordances établies par SATO se situe dans un contexte qu'il est possible de délimiter à l'aide d'une routine conçue à cet effet. Les contextes ainsi délimités sont transférés dans les champs d'une base de données, en l'occurrence dBASE III Plus. (Cette étape du travail d'automatisation subira quelques modifications dans la suite du projet.) Cette étape précède le traitement des données sur chaque terme: en fait, il s'agit de transférer l'information dans un S.G.B.D. afin de pouvoir la traiter. Lorsqu'il établit des concordances, l'utilisateur peut délimiter les contextes terminologiques qu'il conservera pour le traitement terminographique subséquent. L'utilisateur précise la longueur des contextes qu'il veut conserver et délimite avec des références (par ex. { et }) la portion du texte qui sera ensuite importée dans un système de gestion de bases de données.
7. L'analyse sémantique des contextes: étape en voie de réalisation. Il s'agit d'extraire l'information contenue dans les contextes récupérés avec SATO et de l'emmagasiner afin de constituer les dossiers terminographiques (rédaction de définitions, de notes, établissement de relations synonymiques, etc.). Un prototype a permis de dégager les composantes essentielles à un éventuel gestionnaire de contextes.

7.1 Module de gestion :
7.1.1 Choix d'un contexte devant figurer dans le dossier terminologique final: un système de gestion de contextes devrait permettre la sélection d'un contexte parmi les contextes terminologiques retenus pour un terme.
7.1.2 Choix d'un contexte pouvant faire office de note: le système doit permettre la sélection de contextes pouvant figurer dans le dossier définitif à titre de notes techniques, encyclopédiques ou métalinguistiques.
7.1.3 Extraction d'éléments d'information à l'intérieur de contextes pour plusieurs fins :
7.1.3.1 Rédaction de définitions: le système doit permettre d'extraire des nombreux contextes disponibles pour un terme les éléments pouvant aider à la rédaction de la définition de la notion.
7.1.3.2 Rédaction de notes: le système doit permettre d'extraire des contextes les éléments pouvant aider à la rédaction de notes techniques, encyclopédiques et métalinguistiques.
7.1.3.3 Gestion de l'arbre de domaine: le système doit permettre l'extraction d'éléments devant figurer dans l'arborescence du domaine.

7.2 Module de rédaction
7.2.1 Utilisation des éléments d'information extraits :
7.2.1.1 Éléments de définition: le système doit permettre la réutilisation des éléments extraits pour la rédaction de définitions.
7.2.1.2 Éléments de notes: le système doit permettre la réutilisation des éléments extraits pour la rédaction de notes techniques, encyclopédiques ou métalinguistiques.

7.2.1.3 Éléments d'arborescence : le système doit comporter un module de gestion de l'arbre de domaine.

7.3 Module de transfert

7.3.1 Importation des contextes sélectionnés à titre de contextes ou de notes dans une mini-banque de terminologie.

7.3.2 Importation des définitions et des notes rédigées dans une mini-banque de terminologie.

7.3.3 Présentation graphique de l'arborescence.

8. Dernière étape qui restera à automatiser : le transfert automatique des données issues du *gestionnaire de contextes* dans une mini-banque de terminologie (s.g.b.d. de type relationnel).

DÉVELOPPEMENTS À VENIR

Les recherches effectuées afin d'automatiser la recherche terminographique ont fait ressortir que la terminographie peut tirer profit de l'informatique et que la gestion des différentes étapes de la recherche poursuit une logique adaptable à la logique des bases de données. Les travaux réalisés dans le cadre du projet de recherche sur l'automatisation des tâches du terminologue sont énormes si on considère que l'automatisation était pratiquement nulle dans ce domaine. Un point majeur se dégage de la description du projet d'automatisation. Toutes les étapes de la recherche terminographique peuvent transiter par l'informatique en rendant inutile toute intervention manuelle et ce, dans un avenir proche.

Il reste que certaines améliorations peuvent être proposées et les pistes de recherche qui figurent ci-dessous apparaissent, dans l'état actuel des recherches, tout à fait réalisables.

1. Saisie optique automatique : la lecture de documents écrits et leur conversion automatique en fichier informatique apparaît comme l'amélioration majeure dans l'automatisation des tâches du terminologue. La saisie automatique de documents imprimés permettrait au terminologue d'utiliser tous les genres d'écrit pour sa recherche. Cette étape est d'autant plus importante qu'elle précède toutes les autres mais elle ne dépend pas du projet directement.
2. Importation des dossiers créés avec le système de gestion de contextes dans une mini-banque terminologique. Dans son état actuel, le système de gestion de contextes permet la création de dossiers non structurés (d'un point de vue terminographique). Les mini-banques de terminologie permettent la création de fiches structurées. Il faudrait prévoir un module de gestion des dossiers du système de gestion de contextes afin de pouvoir les importer dans une mini-banque terminologique existante.
3. Création d'un module d'exclusion, lors de la recomposition des termes complexes, des mots sans intérêt d'un point de vue terminographique (locutions adverbiales, prépositives, mots de la langue générale, etc.). La création d'un tel module accélérerait grandement le travail terminologique en ce que le terminologue n'aurait plus à passer en revue toutes les formes contenues dans les textes dépouillés.
4. Intégration de toutes les étapes déjà automatisées dans un même module de gestion. Dans leur état actuel, les étapes automatisées demandent plusieurs manipulations informatiques de la part de l'utilisateur. La création d'un module commun éviterait au terminologue d'effectuer toutes ces tâches lui-même.

CONCLUSION

Une grande partie de la chaîne de travail en terminographie a été informatisée dans le cadre du projet de recherche et, nous l'avons vu, certaines étapes supplémentaires pourraient être automatisées à plus ou moins long terme.

L'avantage des procédures d'automatisation proposées dans le présent document est qu'elles portent surtout sur des tâches routinières et fastidieuses (transcription de contextes, repérage des occurrences des termes, etc.). De plus, la recherche par ordinateur permet de relever, sans oublis, toutes les occurrences d'un terme donné, ce qui n'est pas toujours le cas lorsque le travail est effectué par un opérateur humain pour toutes sortes de raisons.

Il reste que plusieurs étapes exigent une intervention de la part de l'utilisateur (choix des termes à conserver pour la recherche, extraction des éléments d'information, choix des contextes à conserver pour les fiches de terminologie définitives, etc.). Il semble que l'automatisation complète de ces étapes ne soit pas envisageable dans un avenir rapproché et que la recherche terminographique exigera toujours une intervention du terminologue.

NOTES

1. Le présent article fait le point sur un projet de recherche commencé à l'été 1989 grâce à une subvention de la faculté des Lettres et portant sur l'automatisation des procédures de travail en terminographie. Ce projet se poursuit depuis le début de l'année 1990 grâce à une subvention du ministère des Affaires extérieures du Canada et du Réseau des industries de la langue, dans le cadre des Sommets francophones (ACCT) avec un nouvel intitulé «Méthodologie du travail terminologique assisté par ordinateur».
2. Ce corpus de quelque dizaines de méga-octets comprend essentiellement les fichiers électroniques des publications collectives K-9 et K-10 du CIRB qui rassemblent des textes reliés au domaine des Industries de la langue.
3. Les logiciels de segmentation (DAT) et de recomposition des unités lexicales complexes (SYREX) ont été écrits en Turbo-Pascal par M. Jacques Ladouceur, étudiant au troisième cycle en linguistique à l'Université Laval.

BIBLIOGRAPHIE

Actas de la exposición de linguística informática y de terminológia científico-tecnica (1988), Paris, Union Latine.

AUGER, Pierre (s. d.)(à paraître) : «La terminotique, un volet particulier de l'informatique langagière», texte d'une conférence présentée au Congrès annuel 1988 de la Société des traducteurs du Québec, 7 p.

AUGER, Pierre (1988) : «Le travail du terminologue amélioré et simplifié», *Circuit*, 22, pp. 12-13.

AUGER, Pierre (1989) : «La terminotique et les industries de la langue», Actes du Colloque «Les terminologies spécialisées : approches quantitative et logico-sémantique», Université de Montréal, 13-14 octobre 1988, *Meta*, 34-3, pp. 450-456.

AUGER, Pierre (1989) : «Informatique et terminologie : revue des technologies nouvelles», Actes du Colloque «Terminologie et industries de la langue», Paris, 19-20 janvier 1989, *Meta*, 34-3, pp. 485-492.

DE SCHAETZEN, Caroline (1987) : «S.g.b.d. et terminologie», *Le linguiste (De Taalkundige)*, 33-3.

DE SCHAETZEN, Caroline (1987) : «Terminologie en langues africaines et gestionnaires de données textuelles», *Le langage et l'homme*, 22-2, pp. 172-174.

DE SCHAETZEN, C. et O. MEERT (1987) : «Un outil d'aide à la création et à la gestion de bases de données terminologiques pour les langues africaines», *Le langage et l'homme*, 22-2, pp. 157-165.

PARADIS, C. et P. AUGER (1987) : «La terminotique ou la terminologie à l'ère de l'informatique», *Meta*, 32-2, pp. 102-110.

PERRON, Jean (1989) : «TERMINO : un système de dépouillement terminologique», *Terminogramme*, 54, pp. 3-9.

TERMINOGRAMME (1988) : «Terminologie et informatique», janvier, 46, 32 p.

VAN DIJK (Bureau) (1988) : *Journées d'études «1992 : Marché unique — Marché multilingue : les outils du traitement des langues».*