5. L'automatisation de la terminométrie : premiers résultats

Jean Quirion

RÉSUMÉ

L'usage réel des termes intéresse au premier plan les terminologues et les autres spécialistes de la langue. La présente contribution porte sur la mesure de l'implantation terminologique, ou terminométrie. Après avoir défini et situé cette dernière, un logiciel terminométrique est présenté, de concert avec la méthodologie qui le sous-tend. Il s'agit, à notre connaissance, du premier logiciel du genre à voir le jour. Des perspectives sur le développement de la terminométrie closent l'exposé.

Automation of terminometry : first results

Jean Quirion

SUMMARY

The real usage of terms is of prime importance for terminologists and other language specialists. This contribution deals with the measure of terminology implementation, also known as terminometry. First, this concept is defined and situated within the field of terminology. Afterwards, a terminometrics software is presented, along with the measure protocol on which it is built. It is, to our knowledge, the first software of this kind. The discussion ends on ideas of future development.

LA TERMINOMÉTRIE

La terminométrie est la mesure de l'implantation terminologique ou, plus largement, de l'usage terminologique.

L'enquête terminométrique est une recherche qui vise à mesurer l'implantation de tous les termes désignant une notion ou un ensemble de notions. Généralement menée sur un groupement notionnel homogène, tel un sous-domaine, elle est le moyen privilégié d'évaluer l'utilisation de terminologies.

Les enquêtes terminométriques ont jusqu'ici surtout porté sur l'emploi de vocabulaires issus du processus de changement planifié que représentent les aménagements linguistique et terminologique menés par des organisations d'État à vocation linguistique. Mais la terminométrie ne se limite pas à ce contexte, comme nous le verrons plus loin.

SITUATION PAR RAPPORT AUX ÉTAPES DE L'AMÉNAGEMENT TERMINOLOGIQUE

Le changement terminologique planifié se découpe en plusieurs étapes, selon Pierre Auger (1986 : 48) : « […] nous caractériserons l'aménagement terminologique par six fonctions fondamentales : a) fonction recherche, b) fonction normalisation, c) fonction diffusion, d) fonction implantation, e) fonction évaluation et contrôle et f) fonction mise à jour […]. »

C'est vers la fin du processus, lors de l'évaluation et du contrôle, que s'insère la terminométrie. La nouveauté relative de cette dernière s'explique par le fait qu'avant de l'évaluer, il faut laisser passer le temps nécessaire à l'implantation, période dont la durée est difficile à évaluer avec exactitude, mais que l'on situe généralement entre sept et cinquante ans (Quirion, 2003a).

Bien que les organisations étatiques aient traditionnellement occupé l'avant-scène du changement terminologique, elles n'en ont pas l'apanage. Les organisations publiques et parapubliques, les entreprises et les médias sont autant d'institutions susceptibles de recourir à la terminométrie. Le but recherché est identique dans tous les cas : s'assurer que les communications internes et externes de l'organisation sont efficaces, terminologiquement parlant.

UTILITÉ DE LA TERMINOMÉTRIE

La terminométrie apporte un éclairage nouveau quant au sort des termes diffusés. Elle permet de chiffrer l'utilisation relative d'un terme par rapport à ses concurrents, c'est-à-dire aux autres termes qui désignent la même notion. Notons que ces concurrents peuvent être de langues différentes, dans le cas de langues en contact (le finnois et le suédois en Finlande, par exemple, ou encore le catalan et le castillan en Catalogne).

Les résultats d'une enquête terminométrique fournissent des données chiffrées sur l'implantation de chacun des termes étudiés : c'est le coefficient d'implantation. D'une valeur comprise entre 0 et 1 (0 signifiant un usage nul et 1 un usage exclusif du terme pour désigner la notion), le coefficient chiffre précisément l'usage terminologique. À l'échelle de domaines ou de sous-domaines, les coefficients d'implantation tracent le portrait de l'utilisation d'une terminologie. Plusieurs facettes peuvent alors être examinées : la prédominance des termes dans une langue plutôt qu'une autre; l'utilisation des termes qui sont recommandés ou normalisés par rapport à ceux qui sont simplement proposés; la variation terminologique en fonction du type de textes, du vecteur de diffusion, du domaine ou du sous-domaine, etc.

C'est le désir de connaître le succès de l'aménagement terminologique qui a d'abord motivé les recherches en mesure de l'implantation. Il est donc naturel que les retombées les plus saillantes de la terminométrie à ce jour touchent les usages terminologiques sur un territoire donné. Il ne faut cependant pas oublier que l'usage terminologique actuel est le fruit de multiples décisions prises précédemment par des terminologues tout au long du travail terminologique : choix des termes à retenir et à écarter, et création de néologismes, notamment. Il va sans dire que ces multiples décisions fondées sur les méthodes de travail courantes en terminologie. À titre d'exemple, les règles généralement mises de l'avant pour la création de néologismes font la promotion de termes courts, dérivables, et conformes aux règles phonétiques, grammaticales et syntaxiques. Or, bien que ces règles paraissent logiques, elles ne sont pas prouvées. Il est en effet difficile de valider entièrement les méthodes terminologiques sans connaître clairement leurs conséquences. Or, ce sont justement ces effets que met désormais en lumière la terminométrie : une fois connus les résultats de l'implantation terminologique, la recherche de leurs causes peut commencer. Connaître les coefficients d'implantation est bien, mais en connaître la cause est nettement plus utile. On peut observer les caractéristiques des termes bien et mal implantés pour dégager les facteurs qui influencent les choix des locuteurs. Cette étude boucle la boucle des étapes de l'aménagement terminologique selon Auger (1986) : une fois effectués l'évaluation et le contrôle de l'aménagement terminologique, l'ultime phase, celle de la mise à jour des terminologies, peut être enclenchée. Ce qui entraîne comme conséquence que les critères et méthodes de travail qui président aux travaux terminologiques sont remis en question, validés ou remplacés.

TRAVAUX À CE JOUR

D'après nos recherches, une vingtaine d'enquêtes terminométriques ont été menées à ce jour, dans quatre États : Israël, la France, la Catalogne et le Québec. On note immédiatement qu'elles présentent des caractéristiques qui les rendent difficilement comparables. Elles s'attachent à des aspects distincts : aux communications écrites ou orales, aux communications institutionnelles ou individuelles, à l'utilisation des termes ou à leur connaissance, à l'utilisation réelle ou au comportement déclaré, à la fréquence d'emploi relative ou absolue, etc.

La terminométrie est une activité exigeante en temps et en ressources humaines. Outre la création du corpus, plusieurs phases doivent être menées à bien; elles seront exposées plus loin. Afin d'accélérer le processus, un logiciel terminométrique a été créé en collaboration avec le Conseil national de recherches du Canada, au sein du Centre de

recherche en technologies langagières, situé sur le campus de l'Université du Québec en Outaouais, au Canada.

BARÇAH, LOGICIEL TERMINOMÉTRIQUE

Le logiciel terminométrique Barçah a été mis au point en 2005. C'est le premier logiciel du genre. Il est fondé sur le protocole terminométrique que nous avons mis au point (Quirion, 2003a). Ce protocole possède les caractéristiques suivantes : il étudie les communications institutionnelles écrites, dans un domaine donné, dans un territoire ou une organisation, à une époque donnée. Il s'appuie sur les quatre vecteurs de diffusion de la terminologie énoncés par Jean-Claude Corbeil (1980), soit l'Administration, l'économie, l'enseignement et les médias.

EXPLICATION DU PROTOCOLE

Le protocole s'apparente à un sondage, dans la mesure où il est soumis aux mêmes règles d'échantillonnage; il est d'ailleurs possible d'en calculer la marge d'erreur. La première étape consiste à choisir le domaine. La terminométrie, rappelons-le, étant une activité exigeante en temps, il est généralement avisé de choisir un sous-domaine plutôt qu'un domaine en entier, afin de limiter le nombre de termes étudiés. Vient ensuite le choix des notions, dont l'augmentation du nombre entraîne l'accroissement rapide des termes; notre expérience montre qu'une notion est désignée par cinq termes en moyenne, dans le cas de deux langues en contact. C'est la raison principale pour laquelle la plupart des enquêtes terminométriques menées à ce jour ne dépassent pas une ou deux centaines de notions, ce qui peut représenter près d'un millier de termes.

Une fois les notions déterminées, on fait l'inventaire des termes. Afin de dresser le portrait le plus juste du paysage terminologique, on recensera à cette étape l'ensemble des termes qui désignent potentiellement chacune des notions à l'étude. Attention : il ne s'agit pas d'effectuer des travaux terminologiques en bonne et due forme, mais d'exploiter les travaux précédemment effectués dans le domaine. On recourra pour cela à des banques de terminologie nationales ou internationales, comme Teleterm (télécommunications), Inforterm (technologies de l'information) ou Eurodicautom. En Grèce, les termes recueillis seront, sans surprise, grecs, mais on pourra aussi compter des termes concurrents, en anglais par exemple. Notons qu'il intéresse tout autant connaître l'implantation de termes grecs uniquement, si tant est qu'ils ne souffrent pas de concurrence de la part d'autres langues. L'étude subséquente de leurs caractéristiques lèvera le voile sur les facteurs d'implantation.

À cette étape, l'exploitation des banques de terminologie est automatisable. Des ententes conclues avec les responsables de ces banques offrent l'avantage d'un accès aux données brutes; ainsi, des ensembles terminologiques (domaine, sous-domaine, année de création, statut du terme (officialisé ou non), etc.) peuvent être cernés et versés électroniquement dans le logiciel terminométrique. Avantages : nul besoin de saisie, d'où un gain de temps et une diminution du risque d'erreurs.

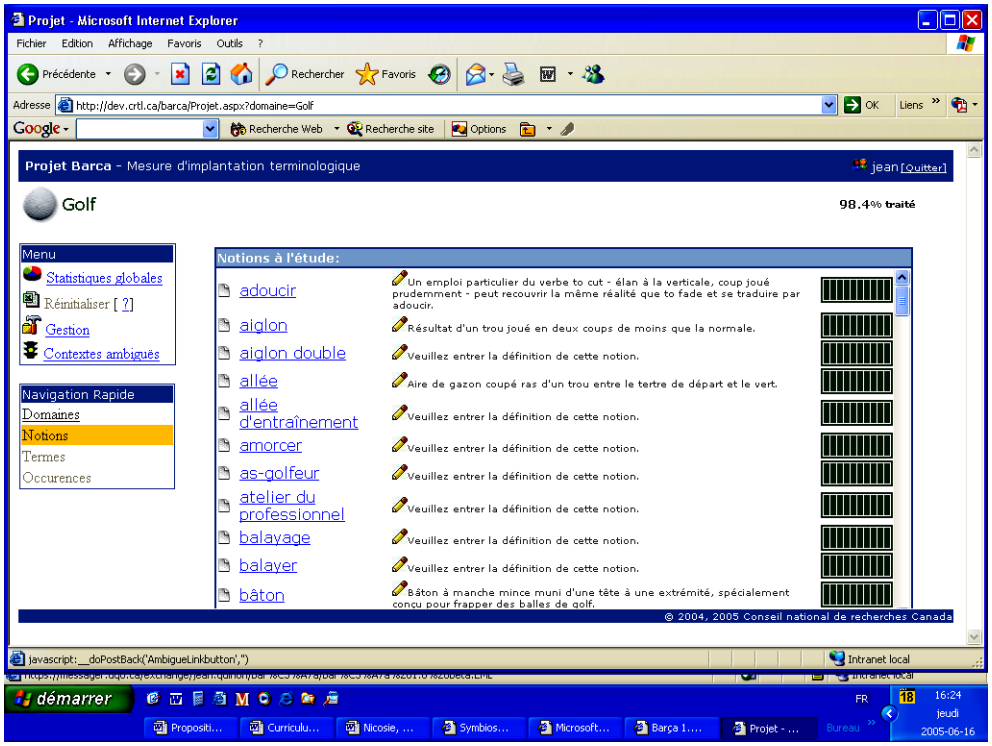

Figure 1. Interface du logiciel Barçah, présentant, au centre gauche les notions à l'étude, au centre leur définition et, à droite, l'état d'avancement de la désambiguïsation.

CRÉATION DU CORPUS

Une fois les termes connus, le corpus dans lequel ils seront recherchés doit être assemblé. Dans le cas de la terminométrie à l'échelle d'une seule organisation, le corpus est formé de l'ensemble des textes produits par celle-ci pendant la période visée. Dans un cas d'aménagement linguistique, les documents produits par les acteurs typiques du domaine constituent le corpus de la période à l'étude. Les vecteurs de l'Administration, de l'économie, de l'enseignement et des médias constituent autant de sous-corpus. Ces documents sont idéalement tirés des sites Internet de ces organisations, étape aisément

automatisable. Sinon, les documents doivent d'abord être obtenus des organisations, puis numérisés.

Une fois le corpus en format électronique, sont retirés les fichiers dans des langues autres que celles étudiées, puis le corpus est indexé.

GESTION DES NOTIONS, TERMES ET CONTEXTES

La suite des opérations tombe en bonne partie sous la coupe du logiciel Barçah. Ce dernier effectue la recherche des termes dans tous les corpus. À titre d'exemple, une enquête sur l'usage de 750 termes dans quatre sous-corpus signifie près de 3 000 interrogations. Celles-ci étant automatiquement effectuées par Barçah, l'utilisateur ne perd pas son temps avec les termes absents du corpus, qui représentent jusqu'à 80 % des termes recherchés (c'est, à tout le moins, le cas dans le domaine des régimes de retraite et rentes et celui des transports au Québec). Certes, une fois l'interrogation effectuée, le logiciel Barçah ne désambiguïse pas automatiquement les contextes : il revient au seul terminologue de déterminer si, par exemple, l'occurrence de *carrefour* relève du domaine des transports qu'il étudie (« lieu où se rencontrent plusieurs voies de communication », et non de la langue générale (« lieu où se rencontrent plusieurs idées ». La désambiguïsation s'effectue dans un environnement convivial, qui présente jusqu'à trois niveaux de contexte, selon les besoins de l'utilisateur : un contexte restreint de deux lignes, un contexte élargi de plusieurs paragraphes et le contexte original, c'est-à-dire le document d'origine, avec sa mise en page. Dans chacun des cas, le terme traité est mis en surbrillance pour assurer son repérage rapide. Le terminologue inscrit, à l'aide de boutons radio, si l'occurrence est valide ou invalide par rapport au domaine étudié, ou si elle est ambiguë.

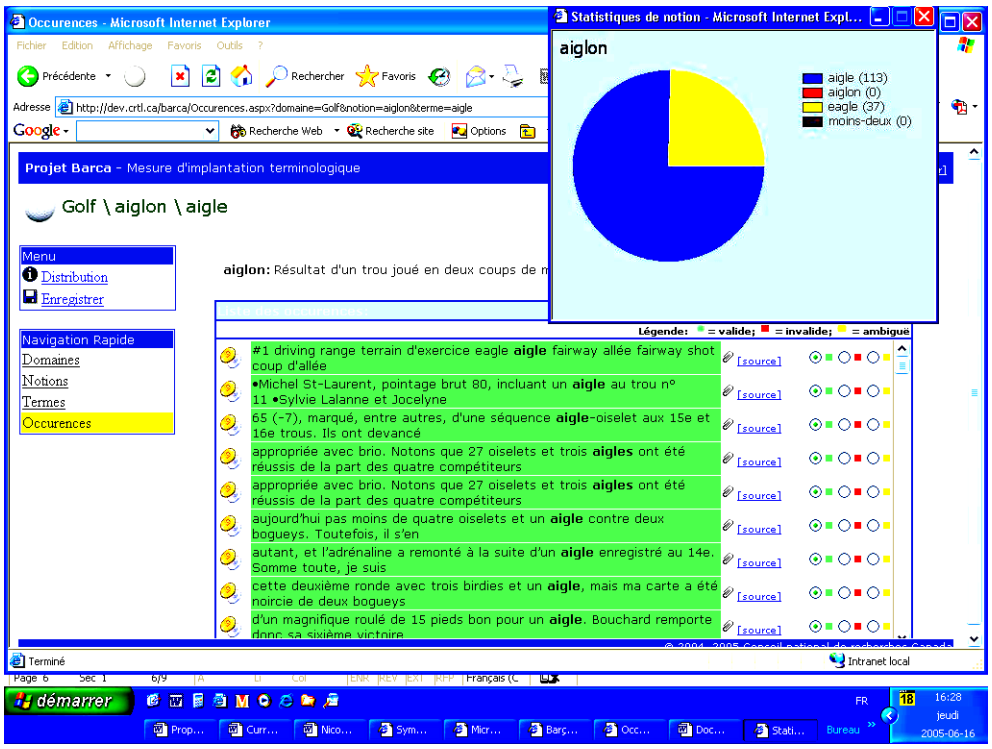

Figure 2. Interface du logiciel Barçah. Au centre, le contexte restreint présentant, en gras, le terme à l'étude. À droite du contexte, le lien [source] affiche le contexte élargi. À droite complètement, les boutons radio utilisés pour valider les contextes. En haut à droite, un aperçu des statistiques d'utilisation générées par le logiciel.

Un autre avantage du logiciel réside dans sa capacité à sélectionner aléatoirement des contextes, ce qui est utile lorsqu'un terme est fortement présent dans un corpus. Supposons que *carrefour* paraît 1 200 fois dans le corpus. Des tests ont démontré qu'il suffit de désambiguïser 100 occurrences choisies aléatoirement pour obtenir le portrait fidèle d'un ensemble plus vaste (Quirion, 2003a). Le logiciel présentera 100 occurrences aléatoires à l'utilisateur. Supposons maintenant que la désambiguïsation manuelle de ces 100 occurrences indique que 33 % des contextes concernent *carrefour* au sens recherché du domaine des transports. Alors, Barçah inscrira automatiquement 400 occurrences (1 200 X 33 %) au crédit du terme.

GÉNÉRATION DES STATISTIQUES D'IMPLANTATION

Une fois la désambiguïsation terminée, Barçah génère automatiquement des statistiques sur l'usage terminologique, c'est le coefficient d'implantation. Ces coefficients chiffrent l'utilisation de chaque terme pris individuellement. Divers regroupements brosseront tantôt le portrait de l'usage par vecteur, par type de termes, par âge du terme, etc., ce qui éclaire notamment les questions suivantes : les termes officialisés s'implantent-ils plus facilement que les termes ordinaires? L'Administration mène-t-elle le bal quant à l'utilisation des termes officialisés? Quel est l'âge moyen des termes les mieux implantés? Évidemment, un intérêt prédominant concerne la connaissance des facteurs influençant l'implantation terminologique, comme il a été discuté précédemment.

UTILITÉ ET UTILISATEURS POTENTIELS

L'intérêt de la terminométrie et l'utilité du logiciel Barçah ne font pas de doute. L'automatisation partielle de la mesure de l'usage entraîne une réduction importante – de l'ordre de 60 % – du temps requis pour un exercice terminométrique. Cette accélération des travaux permet d'envisager des mesures rapprochées, afin de suivre l'évolution de l'implantation terminologique. Car une mesure terminométrique synchronique n'offre qu'un instantané de l'usage terminologique, alors que son évolution, possible par des mesures diachroniques, présente un intérêt encore plus grand. À titre d'exemple, nous collaborons actuellement avec l'Office québécois de la langue française à suivre l'évolution de la terminologie d'un domaine en émergence, celui des nanotechnologies. Par des mesures semestrielles, l'évolution de l'usage sera révélée, dévoilant du même coup le processus d'implantation terminologique, offrant des éléments de réponses aux questions suivantes : quel est le temps requis à un terme pour passer dans l'usage? Quelle est la période critique pendant laquelle on doit en faire la promotion, le cas échéant? Quelle est la durée moyenne de chacune des phases du cycle de vie d'un terme?

Le protocole terminométrique offre également la possibilité, inexploitée à ce jour, de procéder à des mesures sur l'ensemble d'un domaine. Une telle enquête couvrirait un grand nombre de notions, donc de termes. Afin de réduire la quantité requise de travail, l'intension de l'enquête serait diminuée au profit de son extension, générant des coefficients d'implantation à la marge d'erreur augmentée. En revanche, le portrait global de l'usage terminologique d'une sphère d'activité serait dressé, et non seulement celui d'un sous-domaine restreint. Il ne faut pas perdre de vue qu'un tel allégement procédural, s'il entraîne des gains de rapidité, livre corollairement des résultats moins fiables.

Il est aisé de déterminer les utilisateurs potentiels d'un logiciel terminométrique. Parmi eux figurent, au premier chef, les organismes étatiques à vocation linguistique. La terminométrie offre à ces organismes des résultats valides et fiables de leurs efforts d'aménagement terminologique et linguistique. Il va de soi que toute organisation soucieuse de connaître les usages terminologiques de ses membres, afin d'améliorer la qualité de sa terminologie et de sa communication, trouvera également profit à effectuer une enquête de ce genre.

Développé par David Nadeau au Centre de recherche en technologies langagières du Canada, le logiciel est doté d'une interface Web, ce qui décentralise son utilisation[1]. Barçah a obtenu la mention d'honneur en mars dernier au 7e gala des Mérites du français dans les technologies de l'information, organisé par l'Office québécois de la langue française en collaboration avec la Fédération de l'informatique du Québec. Les Mérites du français couronnent les réalisations exemplaires d'entreprises et d'organismes en matière d'utilisation et de promotion du français dans les technologies de l'information.

PERSPECTIVES

Les perspectives de recherche et de développement sont nombreuses. Des affinements doivent constamment être apportés au logiciel, entre autres des améliorations à l'interface et des affinements méthodologiques. Mais le véritable défi réside dans la désambiguïsation automatique des contextes, jusqu'ici traités manuellement. Pour ce faire, les chercheurs du Centre de recherche en technologies langagières bénéficient, grâce à Barçah, d'un bassin de dizaines de milliers de contextes désambiguïsés manuellement dans divers domaines : régime de retraite et rentes, chaussure, golf, nanotechnologies. Ces contextes constituent une matière première nécessaire à l'apprentissage machine. La désambiguïsation automatique réduirait considérablement le temps requis pour procéder à la terminométrie, favorisant des mesures d'implantation plus fréquentes. Plus largement, le projet fera progresser les connaissances en désambiguïsation.

RÉFÉRENCES

AUGER, P. « Francisation et terminologie : l'aménagement terminologique », dans Guy Rondeau et Juan Carlos Sager (réd.), *Termia 84 : terminologie et coopération internationale : la terminologie, outil indispensable au transfert des technologies. Colloque*

[1] Le logiciel présente les caractéristiques suivantes : langage de programmation : C# (côté serveur) et ASP.NET (interface Web);
logiciel d'indexation : Coveo Enterprise Search
base de données : Microsoft SQL Server
Serveur : machine Windows Server 2003

international de terminologie, Luxembourg, 27-29 août 1984, [Québec], Girsterm, 1986, p. 47-55.

CORBEIL, J.-C. *L'aménagement linguistique du Québec*, coll. Langue et société, 3, Montréal, Guérin, 1980, 154 p.

QUIRION, J. *La mesure de l'implantation terminologique : proposition d'un protocole. Étude terminométrique du domaine des transports au Québec*, coll. Langues et sociétés, n° 40, Montréal, Office québécois de la langue française, 2003a, 225 pages.

QUIRION, J. «Methodology for the Design of a Standard Research Protocol for Measuring Terminology Usage», *Terminology*, 9, 1, 2003b, p. 29-49.

Jean Quirion
Directeur du Département d'études langagières
Chercheur associé au Centre de recherche en technologies langagières
Université du Québec en Outaouais
C.P. 1250, succursale Hull
Gatineau (Québec) J8X 3X7
Canada