Meta
Journal des traducteurs
Translators' Journal

META

Terminologie et banques de données d'information scientifique et technique

Jean-Claude Corbeil

Volume 36, numéro 1, mars 1991

La terminologie dans le monde : orientations et recherches

URI : https://id.erudit.org/iderudit/002370ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)

Découvrir la revue

Citer cet article

Corbeil, J.-C. (1991). Terminologie et banques de données d'information scientifique et technique. *Meta*, *36*(1), 128–134.

TERMINOLOGIE ET BANQUES DE DONNÉES D'INFORMATION SCIENTIFIQUE ET TECHNIQUE

JEAN-CLAUDE CORBEIL
Conseil de la langue française du Québec, Montréal, Canada

Nous sommes ici au confluent de deux domaines qui n'ont cessé de se développer depuis les vingt dernières années, le plus souvent séparément, selon des logiques propres déterminées par des objectifs particuliers, la terminologie et les diverses formes de diffusion de l'information scientifique et technique.

Deux grands courants contemporains sont à l'origine du développement de la terminologie.

D'une part, de nombreux facteurs ont considérablement augmenté l'obligation de disposer de terminologies sûres dans des secteurs de plus en plus nombreux et variés. Les facteurs les plus marquants sont: l'intensification des contacts entre les langues européennes à l'occasion de la constitution des grands ensembles politiques et économiques, les exigences de la communication scientifique, technique, administrative et commerciale en vocabulaire de spécialité, l'augmentation du volume de la traduction et de l'interprétation dans tous les croisements de langues, enfin l'essor de l'enseignement fonctionnel des langues dont la partie la plus embarrassante pour les professeurs est le vocabulaire de la spécialité des étudiants adultes réunis pour un apprentissage d'une langue en fonction de leurs besoins professionnels immédiats.

D'autre part, beaucoup de pays se retrouvent devant la tâche d'actualiser le vocabulaire de leurs langues nationales en vue de leur utilisation dans des domaines nouveaux, pour lesquels elles ne sont pas adaptées, ou doivent sans cesse assurer le développement des terminologies de spécialités en langue nationale pour faire face à la concurrence des langues dominantes. Citons comme exemple du premier cas les travaux de terminologie qu'il a fallu mener à bien au Rwanda pour faire fonctionner l'enseignement primaire dans la langue nationale du pays, ou comme exemple du second cas la situation du français au Québec face à l'anglais ou du catalan en Espagne face à l'espagnol. L'élaboration et la mise en place d'un plan d'aménagement linguistique implique nécessairement un volet terminologique, et donc une instrumentalisation linguistique adéquate, notamment pour assurer la standardisation du lexique, la normalisation des terminologies et la création des néologismes.

On comprend alors pourquoi une chose aussi ésotérique en apparence que la terminologie soit devenue une condition *sine qua non* de l'usage des langues dans ce nouveau monde multilingue qui se met en place autour de nous.

Le développement des banques de données d'information scientifique et technique s'est accéléré avec l'usage de l'informatique et de la micro-informatique. Ces banques se constituent dans tous les domaines et dans toutes les directions. On peut distinguer comme grandes catégories: les banques de terminologie, les bases de données bibliographiques, les bases de données textuelles (structurées ou plein texte) et les systèmes experts.

Ce développement rend nécessaire et ouvre de nouvelles avenues de recherches en vue de la mise au point d'instruments de traitement des données ainsi réunies en langue naturelle, tout particulièrement en vue de l'analyse automatique des textes pour en déceler le contenu d'information, ce qui est fondamental en documentation scientifique et technique.

Il nous apparaît aujourd'hui opportun de réfléchir sur les relations entre terminologie et banques de données. Nous nous proposons ici d'explorer ce thème, au moins pour identifier les zones de contact entre ces deux champs.

LA TERMINOLOGIE

Le domaine de la terminologie se subdivise en trois grandes sections :

a) Une méthode de travail, aujourd'hui bien rodée et uniformisée, qui permet d'identifier la dénomination des notions, grâce à un double mouvement d'analyse, analyse des champs notionnels avec identification précise des notions qu'ils comportent dans le but de préciser les choses à nommer et leurs relations entre elles d'une part; d'autre part, une recherche des termes désignant les notions par dépouillement de textes spécialisés et appréciation de chaque terme par rapport à la notion et par rapport à l'usage le plus juste chez les spécialistes. La méthode terminologique allie donc une démarche onomasiologique, pour l'analyse du réel, et une démarche sémasiologique, pour l'analyse des textes spécialisés.

b) Ces travaux aboutissent à la constitution de recueils de termes, plus ou moins extensifs selon les besoins auxquels ils sont destinés, en une ou plusieurs langues, colligés sur le support de base qu'est la fiche de terminologie, recueils susceptibles d'être publiés dans des formats très divers : lexique, dictionnaire, vocabulaire, ou destinés à des emplois très disparates : rédaction de textes, publicité, préparation de manuels, traduction et interprétation, enseignement des langues et des spécialités, etc.

c) La multiplication des fiches de terminologie a conduit à la conception et à l'organisation des banques de terminologie, systèmes informatisés de gestion et de diffusion des terminologies.

Les rapports entre terminologie et banque de données d'information scientifique et technique peuvent donc se situer sur ces trois plans : méthode de travail, lexiques spécialisés et gestion informatisée de l'information.

LES BANQUES DE TERMINOLOGIE

Les banques de terminologie contiennent des termes spécialisés, en une ou plusieurs langues, avec définition, souvent avec des contextes, parfois avec des notes sur l'usage d'un terme, le tout accompagné des sources bibliographiques de chaque renseignement.

D'une manière globale, on peut dire que les banques de terminologie poursuivent deux objectifs principaux : réunir en un seul lieu les résultats de la recherche terminologique, qu'elle soit ponctuelle ou systématique; rendre cette information accessible aux usagers, d'une manière efficace, rapide et conviviale.

L'élément clé d'une banque de terminologie est le terme normalisé, c'est-à-dire le terme que les spécialistes d'une discipline considèrent comme le plus juste pour désigner une notion de leur spécialité. Autour de cet élément clé gravitent les synonymes, les parasynonymes, les termes fautifs, les termes apparentés, qui sont inscrits dans la banque avec renvoi au terme spécialisé. Dans plusieurs banques, on ajoute les équivalences dans une ou plusieurs autres langues, avec plus ou moins de rigueur ou d'exigence en ce qui a trait à la qualité du rapport d'équivalence, notamment lorsqu'il s'agit de prendre position entre plusieurs équivalents possibles.

Pour chaque terme, il est donc nécessaire d'indiquer à quelle spécialité il appartient, puisque ce renseignement détermine, d'une certaine façon, sa définition. On le fait en indexant chacun à l'aide de descripteurs qui identifient le domaine et le sous-domaine de la spécialité à laquelle il appartient. S'il est en usage dans plusieurs spécialités, il y aura autant de fiches que de lieux d'appartenance. Ces descripteurs sont organisés et hiérarchisés par champs de spécialité selon une structure en forme d'arbre qui va du plus général au plus particulier, chaque niveau supérieur incluant les termes des niveaux inférieurs. Il s'établit alors une dynamique entre développement de la banque et complexification des arbres de domaines : au fur et à mesure que les termes d'un domaine s'accumulent, la structure de l'arbre correspondant se complexifie en étendue jusqu'à couvrir la totalité de la spécialité en un seul schéma d'organisation notionnelle. Cependant, la recherche au travers de cette structure de classification ne peut se faire, en général, que verticalement, du plus général au plus particulier et réciproquement, mais pas horizontalement, d'une branche à une autre de l'arbre et, encore plus difficilement, d'un arbre à un autre.

Le repérage de l'information terminologique peut alors s'effectuer de deux manières, selon les besoins de l'utilisateur: soit à partir du mot, avec renvoi, le cas échéant, au terme normalisé, soit à partir du domaine de spécialité, avec possibilité de raffiner la question grâce à la structure de l'arbre d'indexation.

La structure d'indexation des termes, de même que les logiciels de gestion, varient d'une banque à l'autre. Les banques sont donc, actuellement, peu compatibles entre elles et chacune correspond à un mode d'emploi avec lequel l'utilisateur doit se familiariser.

LES BASES DE DONNÉES BIBLIOGRAPHIQUES

Une base de données bibliographiques vise, comme son nom l'indique, à fournir aux chercheurs des références bibliographiques classées par auteur ou par sujets, accessibles rapidement grâce à un système de repérage le plus précis et le plus performant possible, correspondant le mieux au mode normal de recherche des utilisateurs auxquels cette base est destinée.

La technique de la référence bibliographique ne pose aucun embarras: elle est parfaitement au point et seuls quelques détails mineurs sont encore discutés, par exemple la place de la date de publication, immédiatement à la suite du nom de l'auteur ou à la fin de l'article, après la mention de l'éditeur.

En général, chaque référence comprend un résumé du texte cité. Même si cet élément est déjà plus délicat à réaliser, il ne présente pas d'autre difficulté que l'habileté de l'analyste à bien rendre compte du texte. La valeur du résumé dépend aussi de sa longueur, encore trop souvent décidée en fonction des contraintes du système informatique.

Le plus difficile, dans une base de données bibliographiques, est l'identification du sujet dont le texte traite. On pourrait même ajouter que la difficulté grandit quand le sujet traité est susceptible d'intéresser plusieurs spécialistes de disciplines différentes. L'identification du sujet se fait le plus souvent à l'aide de descripteurs, choisis en fonction de leur capacité à évoquer un domaine de spécialité chez le plus grand nombre de spécialistes de ce domaine. En général, on laisse à l'analyste de chaque texte le soin d'identifier la ou les spécialités dont traite le texte et celui de choisir lui-même les descripteurs à partir d'une liste pré-établie ou en suivant sa propre inspiration.

L'élément le plus faible des bases de données bibliographiques est donc le choix des descripteurs. Il s'agit surtout d'un problème de rigueur d'analyse de la relation notion-terme et d'un problème de synonymie ou de parasynonymie, à la fois au niveau de la base elle-même comme ensemble et au niveau des choix individuels de chaque

analyste. Par exemple, y a-t-il une différence entre éducation et enseignement et laquelle? Si l'auteur du texte confond éducation et enseignement, par exemple au sujet de l'immersion, faut-il ne pas en tenir compte dans l'indexation de son texte? Donc, d'une base à l'autre, les termes d'indexation peuvent ne pas correspondre aux mêmes notions, d'un analyste à l'autre, dans la même base, le choix peut varier selon l'idiolecte de chacun. Ceci rend difficile le repérage de l'information par l'usager, d'autant qu'il a lui-même son propre vocabulaire, et rend périlleuse l'intégration des bases en grands ensembles sectoriels.

On voit immédiatement les rapports avec la terminologie. La terminologie peut ou pourrait fournir les termes d'indexation à un bon niveau de normalisation, avec définition des termes et structure de renvoi dans les cas des synonymes ou des termes apparentés. La méthode terminologique pourrait servir à former les analystes aux problèmes de la synonymie ou du repérage des unités terminologiques et les initier à la pratique du terme normalisé par une saine connaissance de son rôle. Enfin, la mise au point des arbres de domaines intéresse aussi bien les terminologues que les documentalistes et les cogniticiens, comme nous le verrons par la suite. Les uns et les autres gagneraient à collaborer sur ce point puisqu'en principe une même spécialité ne peut pas donner lieu à des arbres de domaine contradictoires ou trop divergents.

LES BASES DE DONNÉES TEXTUELLES ET L'ANALYSE AUTOMATIQUE DES TEXTES

Aux éléments d'une base de données bibliographiques, la base de données textuelles ajoute les textes originaux eux-mêmes, l'objectif primaire étant de fournir aux chercheurs les textes qui correspondent à leurs besoins d'information.

La diffusion du texte ne nous intéresse pas ici, quoiqu'il soit toujours primordial pour le terminologue d'avoir accès à des textes originaux. Mais ce n'est qu'un aspect du travail de documentation. Le plus intéressant est d'explorer les rapports entre banque de textes et terminologie, dans l'un et l'autre sens.

Voyons d'abord comment un terminologue peut tirer parti d'une banque de textes.

Une banque des textes d'une même spécialité est une sorte de super-texte. Le terminologue peut parcourir ce super-texte pour en extraire le vocabulaire et créer ainsi une liste de mots qui sera la première étape du repérage de la nomenclature des termes d'une spécialité en supprimant de cette liste les mots qui ne sont pas des termes. Ce tri pourrait se faire plus ou moins automatiquement, par exemple en incluant dans la procédure d'interrogation une liste de mots à ne pas retenir. On peut également obtenir des renseignements sur le statut terminologique des mots en exploitant les signes typographiques comme les guillemets, les caractères italiques ou gras, les parenthèses, dont l'emploi par le rédacteur correspond à des jugements sur le terme, selon qu'il le considère comme étranger, ou néologique, ou synonyme d'un autre, ou équivalent à la périphrase précédente. Le terminologue peut aussi interroger la banque de textes pour trouver et choisir des contextes d'utilisation des mots qu'il a retenus et procéder au travail de définition du terme à partir de ses emplois. Il peut même retracer dans les textes des éléments de définition, ou des définitions formulées par phrase du type «On entend par...» ou les mots entre parenthèses à la suite d'une périphrase, etc.

Les banques de textes ont suscité l'idée de l'analyse automatique des textes, comme instrument plus sophistiqué d'examen du contenu. Les descripteurs des banques bibliographiques sont, en définitive, comme nous l'avons vu plus haut, des indices rudimentaires et subjectifs du contenu d'un texte, incapables de révéler de quoi il traite vraiment. Des équipes essaient aujourd'hui de construire des logiciels qui pourraient repérer plus adéquatement le contenu d'un texte.

Dans l'analyse automatique, le vocabulaire joue un rôle important, puisqu'un sujet traité dans un texte correspond à un ensemble de termes. La relation avec la terminologie est évidente, soit comme méthode de travail, soit comme répertoire de termes. D'un autre côté, le découpage d'un texte en mots est facile lorsqu'il s'agit de mots simples, entre deux blancs, mais devient très compliqué dans le cas des termes complexes, du type *pomme de terre*, renvoyant à une seule et même notion. Des terminologues essaient de mettre au point des logiciels qui pourraient isoler automatiquement les termes complexes dans un texte. L'approche la plus évoluée intègre les indices syntaxiques dans la stratégie de repérage. Ces travaux intéresseront certainement les spécialistes de l'analyse automatique des textes.

LES SYSTÈMES EXPERTS

Les systèmes experts sont à l'avant-garde de la recherche en traitement de l'information scientifique et technique. L'objectif est ambitieux: permettre à un utilisateur de procéder à des opérations intellectuelles complexes en manipulant l'information spécialisée réunie dans une banque de données à l'aide d'un système informatique conçu à cette fin.

Prenons comme exemple un système expert d'analyse de l'adaptation des plantes alimentaires à un environnement de culture: telle plante pourra-t-elle croître d'une manière rentable dans tel pays? Dans le système, on introduit d'un côté les caractéristiques climatiques, pédologiques et entomologiques du pays, région par région; de l'autre, les conditions indispensables à la croissance d'une plante, selon une échelle de rendement économique de sa culture; ceci, évidemment, dans la mesure où ces deux ensembles de données sont disponibles. On construit un logiciel de traitement et de comparaison des deux ensembles de données, appelé moteur d'inférence, dont la fonction est de soutenir le raisonnement de l'utilisateur en répondant correctement aux questions qu'il pose au système, dans le vocabulaire qui est le sien, par exemple: le sol de telle région contient-il les sels minéraux requis pour la croissance de telle plante? Malgré son caractère élémentaire, cet exemple permet de comprendre ce qu'est un système expert.

Comme on le voit, un système expert est constitué de deux éléments:

a) une base de connaissances spécialisées, correspondant à un domaine de recherche, incluant le plus souvent des spécialités multiples en interrelation. Ces connaissances sont inscrites en langue naturelle, avec recours à une terminologie aussi exacte que possible.

b) un système informatique de traitement de l'information, qui permet à l'utilisateur d'interroger le système à sa manière et au système de lui répondre.

Les travaux dans ce domaine sont très récents. Une distinction s'établit peu à peu entre système de simulation et système expert proprement dit.

Dans le premier cas, celui des systèmes de simulation, la collecte des données tend à être exhaustive et les données elles-mêmes tendent à être d'un haut niveau de fiabilité. Le système est destiné à des utilisateurs du même niveau de compétence que ses concepteurs, ce qui entraîne des effets importants à la fois sur la manière d'identifier et de structurer les données, et sur la conception du logiciel de traitement.

Dans le second cas, celui des systèmes experts, la collecte des données correspond à l'état des connaissances du moment caractérisé par des zones d'ombre et un niveau de fiabilité variable. Le système est destiné à des utilisateurs qui ne sont pas toujours du même niveau de compétence que ses concepteurs. D'où la nécessité de guider le cheminement de l'interrogation au moyen d'un logiciel très convivial. Il devient alors impératif de tenir compte de l'écart possible entre le vocabulaire plus ou moins précis de l'utilisateur par rapport à celui du concepteur et de prévoir les mécanismes d'équivalence

qui amèneront l'utilisateur aux données qui l'intéressent, malgré ses imprécisions terminologiques et en partant de ses propres connaissances.

On voit que la langue naturelle est employée aussi bien pour la conception que pour l'utilisation du système. Le vocabulaire devient alors la composante linguistique la plus importante, comme élément essentiel de l'interaction personne/machine.

À l'évidence, la terminologie peut être très utile à la réalisation d'un système expert.

La constitution de la base de connaissances suppose qu'on interroge les experts du domaine et qu'on dépouille la documentation écrite. Dans les deux cas, l'objectif est le même: identifier l'ensemble des connaissances que le système doit contenir pour fonctionner efficacement, selon l'état des connaissances — que la documentation écrite est susceptible de bien refléter —, et selon les besoins des chercheurs, — ce que les entrevues avec les chercheurs peuvent indiquer. Les deux démarches conduisent à la création d'un modèle abstrait explicitant l'organisation hiérarchisée des connaissances, opération qui s'apparente de près à la construction des arbres de domaines en terminologie. Cette partie de la méthode terminologique peut donc être adaptée aux systèmes experts.

La base de connaissances et l'usage de la langue naturelle ont en commun le recours à une terminologie, soit pour l'indexation des connaissances par des termes, lors de la création de la base, termes qui permettront de les retrouver et de les manipuler, soit pour la formulation des questions que l'utilisateur posera au système. Cette terminologie est plus ou moins écartelée entre celle de l'expert et celle de l'utilisateur. Il est donc nécessaire de procéder à une analyse terminologique précise, avec comme objectif l'identification des termes normalisés et l'organisation autour d'eux des termes apparentés, y compris les termes fautifs ou d'usage discutable. En principe, on peut penser que les termes normalisés serviront à la modélisation des connaissances et à leur enregistrement dans le système informatique, alors que le réseau de renvoi des termes apparentés vers le terme normalisé facilitera l'interrogation du système par chaque utilisateur, indépendamment de l'exactitude de sa terminologie personnelle. La limite est atteinte quand une notion n'est pas la même dans le système et dans la tête de l'utilisateur. Encore ici, les relations avec la terminologie sont évidentes, soit que la méthode terminologique permette de réunir le lexique requis pour l'organisation du système expert, soit que la terminologie soit déjà disponible dans des publications ou des banques de termes.

CONCLUSION

De cet examen rapide des relations entre terminologie et information scientifique et technique, on peut tirer quelques observations.

La terminologie peut vraiment servir à autre chose qu'à produire des lexiques. Elle constitue aujourd'hui une démarche intellectuelle rigoureuse qui permet d'appréhender des univers notionnels en maintenant l'articulation entre observation du vocabulaire et modélisation des notions dans des ensembles hiérarchisés des relations complexes entre éléments. En ce sens, elle est une discipline qu'on ne peut négliger chaque fois qu'il s'agit d'analyse notionnelle et de dénomination des notions.

Les rapports entre elle et l'information scientifique et technique sont si évidents et si prometteurs qu'il faut qu'au plus tôt, des relations de collaboration s'établissent entre ces deux champs, surtout par une mise en contact des spécialistes, malgré les barrières universitaires et les méfiances d'écoles.

Enfin, on peut regretter que les logiciels en usage dans les vieux systèmes d'information, comme les banques de terminologie ou les bases de données bibliographiques,

soient si démodés par rapport aux approches actuelles en intelligence artificielle. Il faudra trouver les astuces propres à récupérer cet acquis en améliorant les systèmes informatiques à l'aide des approches récentes.

La langue naturelle revient de plus en plus au centre du traitement de l'information scientifique et technique. Il faudra bien que linguistes et terminologues s'y impliquent et que les spécialistes des autres disciplines concernées découvrent les vertus de ces sciences du langage et apprennent à y avoir recours.