

CARACTÉRISATION LEXICALE DU VOCABULAIRE TECHNIQUE QUÉBÉCOIS

Hélène CAJOLET-LAGANIÈRE
et Normand MAILLET

Centre d'analyse et de traitement informatisé du français québécois
Université de Sherbrooke, Québec, Canada

Résumé : Rares sont les études faites en français, et encore davantage à partir d'un corpus de textes québécois, sur les particularités de la langue technique. Dans le but de décrire les caractéristiques lexicales de la langue technique québécoise, nous avons constitué un corpus représentatif de ce type particulier de discours. Notre analyse porte sur un corpus de 250 000 mots indexés et lemmatisés, tirés de cent textes extraits de dix domaines techniques reliés aux principales activités socioéconomiques du Québec. Nos conclusions portent sur la richesse lexicale de ce vocabulaire, sur la fréquence, la dispersion et l'usage des catégories grammaticales, sur les structures particulières des unités nominales complexes et sur l'identification d'un vocabulaire général d'orientation technique.

Mots clés :langues de spécialité, vocabulaire technique, base de données textuelles, analyse statistique et richesse lexicale, unités nominales complexes, vocabulaire général d'orientation technique.

CARACTÉRISATION LEXICALE DU VOCABULAIRE TECHNIQUE QUÉBÉCOIS

Dans un article publié dans la revue *Présence francophone,* intitulé « Le français dans les communications spécialisées », Yves Gambier déclarait :

> Depuis les années 1970, le panorama des « langues de spécialité » (au moins en Europe), ne s'est guère diversifié quant aux systèmes linguistiques concernés :

ISBN: 0 08 043 438X

> l'écrasante majorité des études touche l'anglo-américain, plus accessoirement l'allemand, le russe et très épisodiquement l'espagnol. (Gambier 1995)

Selon l'auteur, la définition de l'objet Langue de spécialité (LSP) demeure floue, sans nécessairement être controversée : marquée hier par certaines approches linguistiques (parfois discursives, pragmatiques), elle semble être dépassée aujourd'hui par des analyses de type interactionnel et textuel, par des interrogations touchant la vulgarisation des connaissances scientifiques et techniques. Par ailleurs, on ne peut nier l'importance des LSP : selon les études récentes, les textes technicoscientifiques constituent plus de 60 % de l'ensemble de la production écrite en français. Cela a des conséquences directes sur l'enseignement des LSP, sur la formation des rédacteurs et traducteurs techniques et scientifiques, sur la traduction des textes technicoscientifiques, notamment la traduction automatique des textes, de même que sur la production lexicographique, puisque c'est souvent la proportion des termes techniques qui fait varier la nomenclature des dictionnaires usuels. En outre, la délimitation entre LSP et langue commune ne répond pas à des critères rigoureux, ce qui explique que la qualification et le marquage des termes techniques varient d'un lexicographe à l'autre. Aussi, le peu d'études faites sur les LSP en français, et moins encore à partir d'un corpus de textes québécois, ont motivé notre recherche.

Pour Rotislav Kocourek (1991), la langue de spécialité se définit comme « l'ensemble des phénomènes et des ressources linguistiques qui interagissent lors de la transmission d'information dans un domaine spécifique des sciences et des techniques ». Dans les études effectuées jusqu'à présent, les chercheurs ont parlé de langue scientifique en incluant parfois certaines techniques, ou de langue technicoscientifique, laissant entendre ainsi que les deux possèdent les mêmes caractéristiques. Cependant, la langue technique n'a pas été décrite en elle-même et les domaines analysés jusqu'à maintenant ont plutôt été scientifiques : mathématiques, chimie, physique, sciences naturelles, médecine, astronautique, etc. C'est pourquoi, dans le cadre des recherches du Centre d'analyse et de traitement informatique du français québécois (CATIFQ) portant sur la caractérisation de la langue de spécialité au Québec, deux études distinctes sont menées, l'une sur la langue technique et l'autre sur la langue scientifique, faisant appel à deux corpus de textes différents. Une étude comparative des deux corpus permettra de cerner, selon le cas, les spécificités de ces deux types de discours.

La présente recherche concerne essentiellement la langue technique. Nous croyons que la langue technique a son style, ses registres et ses emplois particuliers tirés en partie de la langue commune, qu'elle comporte des traits caractéristiques qui fondent son autonomie, qu'elle constitue enfin un type de discours homogène. L'objet de notre recherche porte sur l'analyse et la description des traits caractéristiques propres à la langue technique québécoise.

Aux fins de notre recherche, nous appelons langue technique, la langue propre aux spécialités quand on les considère du point de vue des manipulations, des applications pratiques, de la fabrication ou de la production d'objets ou de produits. Nous la distinguons de la langue scientifique, qui est plutôt liée aux opérations intellectuelles supposant toute démarche d'analyse, de recherche, d'induction ou de déduction. Bien que cette distinction ne soit pas tout à fait satisfaisante, nous l'avons retenue dans le cadre de nos travaux en raison de son caractère fonctionnel. En effet, il y a beaucoup d'interrelations (sciences et techniques) dans le développement des technologies contemporaines. Les domaines eux-mêmes sont soumis à

ISBN: 0 08 043 438X

ces interférences et ne constituent donc pas, à eux seuls, un critère sûr de délimitation.

MÉTHODOLOGIE

Dans le but de bien décrire les caractéristiques lexicales de la langue technique québécoise, nous avons constitué un corpus d'analyse composé de cent tranches d'égale longueur, soit 2500 mots, comptés selon une méthodologie qui assure l'uniformité de la dimension de celles-ci et choisies à partir de critères rigoureux garantissant la meilleure représentativité possible des types de discours et des rédacteurs. Notre analyse porte ainsi sur un corpus de 250 000 mots répartis dans 100 textes divers extraits de dix domaines techniques reliés aux principales activités socioéconomiques du Québec et reflétant les principaux types de communication propres à la langue technique. Les domaines choisis sont : aluminium, mines, télécommunication, informatique, environnement, hydroélectricité, pâtes et papier, transport, divers 1 (construction, essai et matériaux, béton, etc.) et divers 2 (hydraulique, laser, géomatique, etc.). Quant aux types de communication, ils se regroupent comme suit : rapports techniques, rapports d'évaluation, rapports d'avant-projet, manuels de formation, monographies, guides d'entretien, guides de construction, cahiers de charges, études techniques, enquêtes, normes et codes de procédures. En outre, nous n'avons retenu que des rédacteurs et rédactrices de langue maternelle française, ayant fait des études primaires et secondaires et, le cas échéant, des études collégiales et universitaires au Québec. Enfin, comme il s'agit d'une étude synchronique, tous les textes ont été rédigés après 1980. Ces critères ont été élaborés pour répondre le mieux possible aux exigences d'homogénéité temporelle, spatiale et rédactionnelle, pour rendre compte de la variation communicationnelle et terminologique à l'intérieur des domaines représentatifs de l'activité technique québécoise.

La préparation des tranches du corpus en vue de l'analyse statistique a nécessité plusieurs étapes de traitement et a été faite conformément à certaines normes de dépouillement et de traitement en usage au CATIFQ de l'Université de Sherbrooke. Cela concerne le tirage au sort des tranches, le comptage des mots, l'indexation, la lemmatisation, le traitement des syntagmes, etc. Pour ces divers aspects relatifs à la méthodologie de traitement du corpus, nous référons le lecteur à l'article « Caractérisation des textes techniques québécois », publié dans le numéro 47 de la revue *Présence francophone*, novembre 1995. Nous avons consacré beaucoup de soin et d'énergie à la constitution de ce corpus; cette banque de textes constitue un corpus unique et original compte tenu de son ampleur (250 000 mots tirés de 100 textes techniques entièrement indexés et lemmatisés), mais aussi compte tenu de sa représentativité des principaux domaines techniques québécois.

Tout le travail de dépouillement, d'indexation et de lemmatisation étant terminé, nous disposons actuellement d'un index de toutes les données de notre corpus. Il contient quelque 8400 vocables simples et quelque 6000 syntagmes nominaux lexicalisés. Nous avons considéré comme lexicalisés les syntagmes figurant dans les banques de terminologie du Québec (*Le grand dictionnaire terminologique du Québec*) et du Canada (*Termium*). Tout cela nous a permis de tirer un certain nombre de conclusions sur la caractérisation de la langue technique québécoise.

Comme l'espace dont nous disposons est très court, nous nous sommes concentrés sur les quatre variables suivantes : l'emploi des substantifs, des verbes, des adjectifs et des adverbes

ISBN: 0 08 043 438X

en -ment. Ces faits de langue, on le sait, constituent des indicateurs fondamentaux de l'étendue et de la richesse lexicale d'un texte.

Au départ, nous avons postulé que les textes sélectionnés de chacun des dix domaines retenus devaient contenir un nombre à peu près égal des faits de langue caractérisant la rédaction technique en français. Aussi, l'emploi par les rédacteurs et rédactrices des différentes variables retenues devrait présenter une distribution aléatoire (courbe normale) entre les dix domaines.

Nous avons relevé, pour chacun de nos textes, le nombre de substantifs, de verbes, d'adjectifs et d'adverbes et nous avons calculé, pour chaque variable, les écarts positifs et négatifs observés dans les domaines à partir de la moyenne théorique attendue pour chacune des variables. Nous avons fait ces calculs à la fois pour les occurrences et pour les vocables. Nous avons en outre calculé l'écart type et le χ^2 afin de voir si les écarts situés de part et d'autre de la moyenne étaient significatifs. Aux fins de notre analyse, nous avons considéré comme des écarts « non significatifs » les données situées à l'intérieur de la moyenne plus ou moins un écart type. Cela devrait regrouper normalement 66 % des données pour chaque variable. Nous avons considéré par ailleurs comme des écarts « significatifs » les données qui présentaient de un à deux écarts types par rapport à la moyenne. Cela devrait normalement regrouper quelque 95 % des données pour chaque variable. Enfin, nous avons considéré comme des écarts « très significatifs » les données qui présentaient plus de deux écarts types par rapport à la moyenne.

RÉSULTATS

Le tableau ci-dessous présente la distribution globale des unités lexicales (occurrences et vocables) dans notre corpus.

Domaine	Rang	N	V
Divers 2	1	25 000	2684
Mines	2	25 000	2548
Transport	3	25 000	2425
Informatique	4	25 000	2423
Pâtes et papier	5	25 000	2392
Environnement	6	25 000	2392
Télécommunication	7	25 000	2370
Divers 1	8	25 000	2311
Aluminium	9	25 000	2239
Hydroélectricité	10	25 000	2034
Total		250 000	8384

Les trois domaines dans lesquels le vocabulaire est le plus étendu (nombre le plus important de vocables) sont Divers 2, Mines et Transport; à l'opposé, les domaines dans lesquels le vocabulaire est le moins étendu (les domaines les plus pauvres en vocables) sont

ISBN: 0 08 043 438X

Hydroélectricité, Aluminium et Divers 1.

Voyons maintenant comment ces unités lexicales se distribuent compte tenu des quatre variables étudiées : substantifs, verbes, adjectifs et adverbes.

Domaines	Vocables	Substantifs	Verbes	Adjectifs	Adverbes
Divers 2	2464	1309	561	502	92
Mines	2286	1180	496	520	90
Transport	2226	1188	517	447	74
Papier	2189	1173	508	438	70
Environnement	2182	1161	465	486	70
Divers 1	2121	1178	463	421	59
Informatique	2111	1049	536	441	85
Télécommunication	2013	1067	461	407	78
Aluminium	2012	1052	481	392	87
Hydroélectricité	1825	957	420	400	48

Le tableau présenté ci-dessus indique le rang qu'occupe chaque domaine quant à la richesse lexicale en fonction des quatre variables étudiées. Les domaines qui dénotent la plus grande richesse lexicale sont Divers 2, Mines et Transport; à l'opposé, ceux qui dénotent la plus faible richesse lexicale sont Hydroélectricité, Aluminium et Télécommunication. Un autre domaine présente certaines particularités. Il s'agit d'Informatique. L'emploi des verbes dans le domaine de l'Informatique est très significatif, et ce, tant pour les occurrences que pour les vocables. Les rédacteurs utilisent beaucoup de verbes et beaucoup de verbes différents. Ce domaine dénote donc une richesse lexicale particulière pour ce qui est de l'emploi des verbes. Par ailleurs, ce domaine présente un écart significatif, mais cette fois négatif, quant à son emploi des substantifs. Il révèle enfin des écarts significatifs positifs pour les adverbes (ce qui s'explique par son emploi très grand de verbes), et négatifs pour les articles (ce qui s'explique encore une fois par le peu de substantifs utilisés). De même, comme nous l'avons mentionné précédemment, les domaines Transport et Papier sont au-dessus de la moyenne théorique, tant pour les occurrences que pour les vocables, quant à leur emploi de verbes et de substantifs (ce qui signifie que les rédacteurs emploient non seulement beaucoup de verbes et de substantifs, mais encore beaucoup de verbes et de substantifs différents dans leurs textes). Ce sont là deux domaines qui présentent une richesse lexicale plus grande que les autres domaines quant à ces deux catégories grammaticales.

Que pouvons-nous déduire de ces résultats?

D'une part, compte tenu des résultats obtenus, on ne peut affirmer que les faits de langue étudiés se distribuent normalement d'un domaine à l'autre. On doit admettre que les écarts notés pour ces quatre variables sont directement liés au domaine technique. De plus, sept domaines sur dix présentent des écarts significatifs, et parfois même très significatifs, pour l'une ou l'autre des variables étudiées.

L'analyse des écarts significatifs relevés dans ces quatre classes grammaticales nous mène à

ISBN: 0 08 043 438X

la conclusion que les variables qui influencent leur emploi relèvent plutôt du style (structures syntaxiques particulières telles que des énoncés télégraphiques, des répétitions, des énumérations, des consignes, etc.) et du caractère formalisé de certains textes (études d'avant-prpjet, normes, etc.) que du domaine technique lui-même. Toutefois, certaines particularités lexicales, comme l'usage des verbes dans des énoncés de programmation du domaine Informatique joue un rôle dans l'emploi spécifique des unités lexicales. En résumé, l'utilisation de structures syntaxiques particulières, l'emploi à valeur nominale de verbes dans des contextes informatiques et les répétitions fréquentes dans certains textes constituent les causes principales observées qui influencent la fréquence d'emploi de l'une ou l'autre de ces quatre classes grammaticales.

Par ailleurs, il est intéressant de noter que si l'on observe les données propres à chaque texte de chacun des domaines, le nombre d'unités lexicales différentes est plus élevé dans les extraits provenant de monographies et autres documents visant une certaine vulgarisation des informations véhiculées. À l'opposé, les textes qui présentent le plus faible taux de richesse lexicale sont surtout des extraits de manuels d'entreprise. Il s'agit essentiellement de documents à l'usage des travailleurs de l'entreprise. Contrairement aux monographies, ces documents ont un but exclusivement pratique, soit la formation des travailleurs quant à la connaissance de nouveaux procédés, de nouvelles machines, etc. Les thèmes étant limités, il peut de ce fait être normal que la terminologie utilisée soit elle aussi limitée et en partie redondante (noms des opérations, des machines, des outils et autres).

CONCLUSION

L'objet de notre communication portait sur la constitution d'un important corpus original composé d'un échantillonnage représentatif de textes techniques tirés des principales activités socioéconomiques au Québec et sur les résultats des analyses faites à partir de quatre variables. Cette étude a révélé plusieurs données intéressantes : d'une part, les quatre variables observées, soit l'emploi des substantifs, verbes, adjectifs et adverbes, ne se distribuent pas d'une manière aléatoire entre les textes des dix domaines techniques retenus. Le type d'ouvrages d'où sont extraits les textes joue un rôle prépondérant dans la richesse lexicale de ces derniers. Les textes à large diffusion, dont le caractère de vulgarisation est plus évident, présentent une richesse lexicale plus grande que les textes plus techniques, dont le contenu est plus circonscrit. Il semble y avoir une corrélation très nette entre le type de texte, vulgarisé ou plus spécifique, et la richesse lexicale des domaines.

L'emploi des quatre classes grammaticales étudiées est influencé également par la taille et le nombre de champs lexico-sémantiques mis en jeu dans chaque domaine. En outre, comme nous l'avons mentionné, certaines particularités syntaxiques et la structure même des textes modifient la fréquence d'emploi des unités lexicales dans le discours technique québécois.

Les analyses subséquentes faites sur les syntagmes de même que les études des unités lexicales par groupe de fréquences, notamment celle des hapax, ont permis de dégager de nombreux autres faits linguistiques. Les mots grammaticaux, par exemple, ont un emploi bien distinct dans les textes techniques en comparaison à d'autres genres de textes, littéraires ou oraux. Les syntagmes nominaux influencent fortement le poids lexical de la classe nominale dans l'ensemble du vocabulaire technique. Nous avons également observé qu'il y a un lien

ISBN: 0 08 043 438X

étroit entre la basse fréquence moyenne d'emploi de ceux-ci et la complexité de leurs structures. Mentionnons enfin que leur distribution est liée, comme c'est le cas pour les unités lexicales simples, à des contraintes thématiques et stylistiques.

Les données recueillies dans cette recherche permettront de pousser des investigations et des comparaisons dans différents autres corpus tant québécois que français, belges et autres. Ces études comparatives permettront, d'une part, de poursuivre la description des traits linguistiques du français québécois et, d'autre part, de mieux cerner les particularités des LSP.

RÉFÉRENCES

GAMBIER, Yves. « Le français dans les communications spécialisées : bilan mitigé », *Présence francophone, langue de spécialité*, numéro 47, Université de Sherbrooke, Québec, 1995, p. 9-36.

KOCOUREK, Rotislav (1991). *La langue française de la technique et de la science*, deuxième édition augmentée, refondue et mise à jour, Oscar Brandstetter Verlag GMBH et CO, Wiesbaden, 327 p.

CAJOLET-LAGANIÈRE, Hélène et Normand MAILLET. « Caractérisation des textes techniques québécois », *Présence francophone, langue de spécialité*, numéro 47, Université de Sherbrooke, Québec, 1995, p. 113-137.

ISBN: 0 08 043 438X