Méthodologie de la constitution du corpus

Une étude sur l'implantation de termes officiels suppose bien évidemment que soit constitué tout d'abord un corpus de termes qui servira de base à cette étude. Et lorsqu'on se propose d'observer ou de mesurer l'implantation de ces termes dans l'usage écrit, à tel ou tel niveau de communication, on sera conduit à constituer également un corpus de textes à dépouiller.

Les quelques remarques qui vont suivre présentent la démarche de chacune des équipes ayant participé à l'enquête sur l'implantation des termes officiels mise en place par la Délégation générale à la langue française. Elles font apparaître dans les méthodes utilisées pour la constitution des corpus une assez grande diversité liée à la situation particulière des vocabulaires étudiés et aux orientations propres à chaque recherche.

1 La constitution des corpus de termes

Nous distinguerons le *corpus de base*, parfois trop vaste pour permettre de réaliser une étude d'implantation, et le *corpus utile* constitué par une sélection restreinte des termes du corpus de base.

S'agissant d'études sur des termes officiels, le corpus de base est constitué généralement par les listes de termes proposées par les commissions ministérielles de terminologie et publiées au *Journal officiel*.

L'équipe de l'Université de Toulouse Le Mirail, chargée de l'enquête sur le vocabulaire de la télédétection aérospatiale, a élargi son corpus de base en complétant les listes de termes officiels par un dictionnaire spécialisé dans le domaine et un ouvrage de référence. D'autre part l'étude du vocabulaire de la métallurgie de transformation appliquée à l'industrie aéronautique (également réalisée à Toulouse) présente une situation particulière. Aucun arrêté ministériel ni aucun recueil normé n'étant consacrés à ce domaine technique, l'équipe responsable de cette étude a dû constituer son propre corpus. Elle a réuni une documentation scientifique et technique concernant le savoir de base et les connaissances plus approfondies du domaine, et elle a fait appel à des spécialistes pour sélectionner les termes en fonction de leur représentativité et du degré de spécialisation des personnes qui devaient être interrogées.

Le nombre des termes officiels constituant le corpus de base du

vocabulaire de la santé et de la médecine (36 termes) et du vocabulaire de l'informatique (136 termes) étant relativement peu important, ces corpus ont pu être utilisés directement pour réaliser l'étude d'implantation. Par contre, pour les autres domaines, le corpus de base a été jugé trop vaste, et un corpus utile plus restreint a été constitué.

La sélection des termes a été réalisée selon différents critères.

L'équipe de l'Université de Rouen (vocabulaire du génie génétique) note dans son rapport: «... il nous a fallu procéder à une typification de ce corpus et porter notre attention sur les formes posant problème et sur les points sensibles. L'objectif principal de l'arrêté étant de limiter les emprunts qui sont faits à l'anglais, le problème du contact des langues nous a guidés pour une sélection des unités» (Gaudin et Guespin 1993: 11).

Dans les études concernant les domaines de l'audiovisuel, de la publicité et de la télédétection aérospatiale, on a cherché plus particulièrement à réunir un ensemble homogène en fonction des objectifs poursuivis.

Ainsi, dans l'étude que nous avons réalisée sur le vocabulaire de l'audiovisuel et de la publicité, nous nous sommes attaché principalement à observer la réalité de l'usage, dans la communication technique (plus spécialement au niveau de la communication didactique et de vulgarisation) et dans la langue courante. Après avoir procédé à différentes éliminations dans le corpus de base dont nous disposions (termes français étudiés par la commission, termes anglais et équivalents français ayant la même forme graphique, etc.), nous avons donc essayé de constituer un corpus restreint comprenant des termes utilisés aussi bien dans les langues techniques que dans la langue courante. Pour ce faire, nous avons utilisé des documents de référence, en particulier des dictionnaires d'usage, et nous avons soumis le corpus à des spécialistes du domaine au cours de l'enquête préparatoire (Chansou 1993: 6).

L'objectif retenu par Josiane Rouges-Martinez pour étudier l'implantation des termes dans le domaine de la télédétection aérospatiale était tout différent. L'étude reposait non pas sur une recherche d'attestations dans des textes, mais sur les déclarations d'utilisation de certains termes faites par des spécialistes en réponse à un questionnaire. Il s'agissait de proposer un ensemble de termes représentatifs de la discipline à deux groupes: d'une part des spécialistes intervenant dans les activités de formation et de recherche de l'université et du CNRS, et d'autre part des spécialistes œuvrant à l'élaboration ou à la commercialisation des produits de la télédétection. Les termes sélectionnés devaient donc se situer à un certain niveau de communication scientifique et technique et concerner les pratiques langagières des deux groupes d'interlocuteurs. On citera par exemple un des critères figurant dans la grille de sélection soumise à un informateur spécialiste du domaine: «Cette grille devrait permettre de définir un corpus comportant: [...] c) des termes dont l'usage est en débat dans la discipline; nous savons qu'à un moment donné dans une discipline, des débats entre spécialistes s'instaurent autour de termes quand l'opération de désignation d'une réalité ne paraît plus parfaitement adéquate en fonction de l'évolution des connaissances» (Rouges-Martinez 1992: 3). Un tel critère de sélection est précisément défini en fonction des orientations données à l'enquête.

On notera enfin que l'équipe de l'Université Rennes II, pour réaliser une étude complémentaire sur l'implantation des termes officiels dans les dictionnaires d'informatique, a retenu un échantillon aléatoire de 15 termes dans l'ensemble des termes de l'arrêté ministériel.

2 La constitution des corpus de textes

Mises à part les deux études réalisées à Toulouse, les études d'implantation reposent sur une recherche d'attestations dans des textes. Toutes les équipes, pour réunir un corpus de textes, ont eu recours aux mêmes principes méthodologiques de base. Mais chaque démarche présente des particularités qui font apparaître des éléments d'une méthodologie concernant plus précisément les recherches sur l'implantation terminologique.

Pour constituer son corpus de textes, Philippe Thoiron tient compte tout d'abord de la dimension diachronique. Les deux arrêtés qui servent de base à l'étude du vocabulaire de la santé et de la médecine ont été publiés en 1975 et en 1978. Deux tranches de temps sont distinguées en fonction de la date de la parution des textes: textes publiés entre 1975 et 1984 d'une part, entre 1985 et 1992 d'autre part. Le corpus, par ailleurs, doit être équilibré pour représenter les diverses spécialités concernées. Enfin il est constitué en fonction d'un choix au niveau du degré de spécialisation des textes. Il comprend des articles de pointe, des ouvrages de vulgarisation et des manuels pédagogiques, et se situe donc au niveau d'une communication spécialisée (Thoiron 1993: 3). Le corpus ainsi constitué est important; il contient 249 titres. On peut penser qu'une recherche d'attestations dans la presse grand public aurait été peu productive, compte tenu des conditions d'emploi

des termes des arrêtés, et qu'elle aurait porté atteinte à la cohérence de cette étude d'implantation.

L'équipe de Rouen retient à la fois une approche diachronique et une approche synchronique. L'arrêté relatif à la terminologie du génie génétique a été publié en 1990. Le corpus comprend des textes publiés en 1987 et en 1988, et des textes contemporains de l'arrêté. On peut ainsi considérer les unités déjà implantées avant la publication de l'arrêté. «Les enquêtes, note le *Rapport final*, porteront utilement sur la partie problématique du corpus, la dimension diachronique permettant d'évacuer des problèmes d'usage que le temps a réglés, la comparaison en synchronie permettant de repérer les conflits...» (Gaudin et Guespin 1993: 17, 18). Le corpus comprend des dictionnaires techniques, des articles de presse (presse grand public, revues de vulgarisation, revues d'interface), des ouvrages spécialisés et des documents universitaires (polycopiés de cours et thèses). On notera que la presse grand public est peu représentée dans ce corpus étant donné le caractère spécialisé des termes du génie génétique. Les sources ont été sélectionnées en fonction de leur rôle «glottopolitique» (ouvrages de référence normatifs et non-normatifs), du lectorat visé, du caractère normatif des situations d'utilisation (supports pédagogiques formels et informels). Le rapport souligne à ce sujet l'importance, dans l'étude des pratiques langagières, de la fonction pédagogique des enseignants/chercheurs: «... l'écriture d'une thèse constitue l'occasion d'un effort maximum de francisation demandé en milieu universitaire. Par contre l'attitude adoptée en cours dépend du seul sentiment des enseignants. [...] ces deux types de support permettent d'évaluer l'effort de francisation effective en milieu universitaire...» (*op. cit.*: pp. 16, 17).

En ce qui concerne l'étude sur le vocabulaire de l'audiovisuel et de la publicité, nous avons adopté une approche synchronique. Nous nous proposions en effet de rendre compte de la réalité de l'usage dans la période strictement contemporaine, de donner en quelque sorte un instantané de l'usage. L'étude ayant été réalisée dans le courant de l'année 1992, nous avons retenu la période de 1989 à 1992, période relativement étendue, mais qui permettait une assez large observation. Nous avons l'intention, dans un deuxième temps, de rassembler un corpus complémentaire avec des textes des années 1993 et 1994, de façon à observer dans une perspective diachronique d'éventuelles évolutions de l'usage, spécialement pour des formes «sensibles» telles que *sponsoring/parrainage*, *mailing/publipostage*, etc. Les textes ont été choisis avec l'aide de professionnels de la communication, de telle façon que soient représentés d'une manière à peu près équilibrée les différents niveaux de communication observés dans l'enquête: manuels techniques et ouvrages de vulgarisation, manuels et documents pédagogiques, presse spécialisée grand public, presse d'information. Le vocabulaire de l'audiovisuel et de la publicité étant largement répandu dans l'usage courant, les textes de la presse grand public tiennent ici une place importante dans le corpus. Des dictionnaires généraux ont été également retenus comme témoins de l'usage le plus habituel. Par ailleurs, l'audiovisuel faisant l'objet de diverses réglementations, nous avons réuni un ensemble de textes produits par le Conseil supérieur de l'audiovisuel, textes qui nous permettent d'observer les usages «officiels». Ainsi la composition du corpus reflète dans une certaine mesure la place tenue par l'audiovisuel et la publicité dans la société d'aujourd'hui.

L'équipe de Rennes, s'appuyant sur les travaux d'un groupe d'étudiants et disposant d'importants moyens d'informatique, a pu constituer des corpus de textes très étendus. Dans le but de rechercher des attestations des termes «officiels», un premier corpus a été rassemblé au cours d'une recherche aléatoire menée par des étudiants dans le cadre d'activités habituelles de formation (constitution de répertoires et d'index documentaires). Les étudiants ignoraient qu'ils participaient à cette recherche. Ils avaient pour consigne de relever tous les termes d'informatique. Les sources étaient librement choisies et couvraient un vaste ensemble de situations de communication: catalogues, devis, brochures, articles, nomenclature, cours, notes de service, etc., à l'exception des dictionnaires. Un deuxième corpus a été constitué en vue d'étudier les degrés d'implantation des termes français dans des discours ou documents se rapportant à l'informatique. Le champ de la recherche était très large: le corpus était défini comme un ensemble de documents écrits et sonores produits par toute personne dont l'informatique constitue un objet d'étude, une activité, un métier. Divers critères ont été retenus pour assurer la représentativité du corpus, et des informateurs ont été invités à participer à la recherche des documents. On a considéré les secteurs d'application (production, promotion, vente, maintenance de matériels et de logiciels, information dans la presse spécialisée et non spécialisée, formation, etc.), les types de public, les degrés de formalisme. Le corpus brut ainsi rassemblé comprenait 502 documents. À partir de ce corpus brut a été déterminé un

corpus efficace correspondant à un échantillon représentatif des énoncés.

Conclusion

Au terme de cette étude sur la méthodologie de la constitution des corpus, on constate que les démarches des équipes ayant participé à l'enquête présentent de sensibles différences.

Ces différences sont évidemment explicables.

Chaque approche méthodologique est fonction des caractéristiques du domaine étudié, du degré de spécialisation des termes proposés dans les arrêtés ministériels, des situations dans lesquelles ils sont employés. Il est clair que la constitution d'un corpus de textes pour mener une enquête sur l'implantation des termes du génie génétique, domaine très spécialisé, et le choix de textes pour une étude sur le vocabulaire de l'audiovisuel et de la publicité, vocabulaire largement répandu, supposent des démarches différentes. La délimitation et l'étendue du corpus dépendent d'autre part des moyens matériels et humains dont on peut disposer; l'étude «lourde« réalisée par l'équipe de Rennes se distingue, de ce point de vue, des recherches forcément plus limitées menées par d'autres équipes.

Enfin les différences observées dans les démarches méthodologiques des diverses équipes sont liées aux orientations de chaque recherche et aux objectifs poursuivis, objectifs définis à partir des choix personnels des responsables de la recherche. Et l'on voit que chaque enquête a été menée dans une perspective particulière.

Aussi paraît-il difficile, au stade actuel de nos travaux, de dégager de véritables convergences dans les méthodes, du moins en ce qui concerne la constitution des corpus.

Mais l'examen des diverses approches permet de développer une réflexion utile, d'enrichir la méthodologie et d'affiner des outils pour d'autres études d'implantation terminologique.

Michel Chansou,
Laboratoire «Lexicométrie et textes politiques», Institut national de la langue française, CNRS, Saint-Cloud, France.

Sources

Chansou (Michel), 1993: *Évaluation d'une action de politique linguistique. Les travaux de la commission ministérielle de terminologie de l'audiovisuel et de la publicité. Rapport de recherche* + Annexe. Délégation générale à la langue française.

Gasquet (Evelyne), 1992: Villebrun (Isabelle), *Rapport final du programme de recherche. L'implantation terminologique dans le domaine de la métallurgie*, Université Toulouse Le Mirail.

Gaudin (François) et Guespin (Louis), 1993: *Rapport final. Enquête sur l'impact des arrêtés terminologiques. Domaine: Génie génétique*, Université de Rouen.

Gouadec (Daniel), dir., 1993: *Étude d'implantation des termes officiels de l'informatique*, Université de Rennes II.

Rouges-Martinez (Josiane), 1992: *Rapport final du programme de recherche. L'implantation terminologique dans le domaine de la télédétection aérospatiale*, Université Toulouse Le Mirail.

Thoiron (Ph.), Iwaz (J.) et Zaouche (N.), 1993: *Résultats de l'enquête d'implantation des termes de santé et de médecine*, Université Lumière Lyon II.