

TABLE-RONDE SUR L'ORGANISATION DES *CORPORA*

Maria Fernanda Bacelar do Nascimento
Centro de Linguística da Universidade de Lisboa (CLUL)

Le noyau central du Corpus de Référence du Portugais Contemporain (CRPC) a été constitué par le *corpus* oral recueilli pendant les années 70 pour l'établissement du Portugais Fondamental (sur le modèle du Français Fondamental). En 1987, année où ont été publiés les volumes contenant les documents et la description des méthodes suivies dans le projet Portugais Fondamental, et où le *corpus* oral commençait à être utilisé pour l'analyse syntaxique (selon les principes théoriques et méthodologiques du Groupe Aixois de Recherches en Syntaxe - GARS), nous avons conçu la constitution d'un grand *corpus* de langue portugaise orale et écrite, *corpus* dont la dimension est aujourd'hui de plus de 80 millions de mots.

Dans le dessin initial du *corpus* et en ce qui concernait sa dimension, nous avons proposé que le *corpus* serait le plus grand possible, c'est-à-dire, nous l'avons envisagé comme un *corpus* ouvert, continuellement actualisé.

Du point de vue des proportions entre les textes, nous avons proposé que les transcriptions de l'oral n'auraient ni moins de 500 mots ni plus de 1000 mots chacune (chaque échantillon du Portugais Fondamental avait 500 mots). En ce qui concerne l'écrit, nous n'avons pas limité la dimension des textes. Nous avons, en tout cas, décidé que le *corpus* ne serait pas un *corpus* textuel (en ce sens qu'il n'aurait pas forcément des textes complets) mais qu'il serait un *corpus* de référence (en ce sens qu'il aurait des échantillons de chaque œuvre ou document représenté). L'idée était d'inclure 30 à 40 pour cent de chaque ouvrage: il nous semblait en effet plus prudent de ne pas informatiser les textes

complets, étant donné l'absence de législation sur les *corpora* qui puisse protéger soit les auteurs de textes, soit les auteurs de *corpora*.

Nous avons établi, pour chaque type de texte, des critères de sélection des échantillons. Par exemple, pour les échantillons des œuvres littéraires on a toujours sélectionné des parties complètes (chapitre, poème, etc.) du début, du milieu et de la fin de l'œuvre. Dans les livres didactiques, nous avons surtout écarté les tableaux, les schémas, les exercices.

En ce qui concerne le contenu, nos critères tenaient compte de quelques objectifs essentiels. Tout d'abord, les fonctions du *corpus* devaient être multiples et, donc, servir une grande variété d'utilisateurs; il devait pouvoir représenter le portugais contemporain "moyen", c'est-à-dire, où les styles particuliers seraient estompés. Il devait donc être non seulement très grand, mais aussi le plus diversifié possible, avec une représentation significative du langage plus courant (l'oral informel, les écrits informatifs, les ouvrages de divulgation). Il serait, ainsi, un *corpus* utile pour les auteurs d'ouvrages de référence (dictionnaires, grammaires, manuels) et pour un assez grand nombre d'études. Mais, d'autre part, il devait pouvoir servir à l'extraction d'information sur les langues de spécialité et donc, il devait contenir des sous-*corpora* spécialisés (techniques, scientifiques, etc.) de dimensions significatives.

Ce *corpus* devait aussi permettre des études contrastives entre les variétés nationales et régionales du portugais, c'est-à-dire, du Brésil, des cinq pays de l'Afrique lusophone et de quelques régions asiatiques comme Macau, Goa ou Timor-Leste. On avait aussi prévu que les textes devaient avoir été originellement écrits en portugais et devaient être datés du XX^{ème} siècle.

Les **supports** seraient: les livres, les périodiques, des documents publiés ou pas, des lettres, des dépliants et les transcriptions de l'oral.

Pour le **genre**, on a simplement prévu pour l'écrit une distinction entre textes de

fiction (romans, narratives, poésie, etc.) et textes informatifs.

Comme vous voyez, nous avons utilisé, dans la sélection des échantillons, des critères externes. Même l'idée de diversification que nous avons introduite est associée à une notion qui est, dans les typologies textuelles de *corpora*, très controversée, quoique assez utilisée: je parle de la notion de thème (ou topique). Nous n'avons pas établi une liste limitée de thèmes, mais nous avons prévu pour l'oral et pour l'écrit, en général, une grande diversification (de la politique aux arts, de l'agriculture à l'économie, de l'informatique à la mode), sans aucune restriction. La liste de thèmes serait donc ouverte et nous essayerons d'assurer la présence des thèmes les plus communs.

Les textes sont bien identifiés; des informations codifiées sur les documents rendent compte des références bibliographiques, des sources primaires utilisées, de la responsabilité éditoriale et, dans le cas du *corpus* oral, des responsables de la transcription. Il y a aussi des renseignements sur le format électronique et sur l'état du document, en ce qui concerne son utilisation (s'il a été codifié, vérifié, etc.).

Pour ce qui est de la représentation logique du texte, nous avons commencé par établir nos propres normes pour le traitement informatique: notations spécifiques pour chaque classe d'information comme, par exemple, les italiques, les gras, les citations, etc.

Nous avons, pourtant, modifié plusieurs aspects au cours de ces dernières années: aussi bien le dessin initial du *corpus*, que les décisions prises à propos de la chronologie, des pourcentages, de la représentation logique du texte. D'une part, la pratique nous a prouvé que quelques-unes étaient infaisables, d'autre part, l'extraordinaire développement de la linguistique de *corpus* et la création de plusieurs projets internationaux a permis aux auteurs de *corpus* qui, comme nous, pendant les années 70 et 80 étaient assez isolés, d'avoir accès à

beaucoup d'information donnée par les auteurs plus expérimentés. Il faut dire que l'idée d'une harmonisation, ou même d'une normalisation de critères de constitution et de traitement informatique de *corpora*, tout en préservant un assez grand degré de liberté, s'est répandue pendant les dernières années, ce qui, entre d'autres bénéfices, va rendre possible la réalisation d'études contrastives et l'utilisation d'outils d'analyse communs et assez sophistiqués.

J'ai dit que quelques-unes de nos décisions de début n'étaient pas faisables et je veux préciser. Tout d'abord, il faut dire que faire un *corpus* électronique de grandes dimensions est un projet très, très, cher et qui demande des équipes interdisciplinaires de linguistes et d'informaticiens à plein temps et qui exige aussi, beaucoup de temps pour l'organisation (négociation avec les fournisseurs de données, rédaction de contrats, etc.). Et je ne parle que de la constitution du *corpus*, de son organisation et classification et de la construction des outils informatiques indispensables aux exploitations courantes. Je ne parle pas de l'exploitation elle-même.

Nous avons vite compris que c'était un peu illusoire de croire qu'on pourrait satisfaire tous les utilisateurs. Il y a, chaque fois, une demande imprévue et même tout à fait inespérée. Pour commencer, nous avons été obligés de changer la datation du *corpus*. Le Dictionnaire de l'Académie des Sciences de Lisbonne a été notre premier utilisateur pour ses citations et pour la validation de sa nomenclature. Or la direction de ce dictionnaire avait prévue l'inclusion d'auteurs de la 2^{ème} partie du XIX^{ème} siècle et, pour répondre à cette exigence, nous avons introduit des oeuvres littéraires portugaises et brésiliennes de cette période. D'autre part, quand les utilisateurs prétendent réaliser des études contrastives, ils demandent des sous-*corpora* que nous ne possédons pas ou qui n'obéissent pas aux caractéristiques de taille, de format, etc, envisagées.

Mais c'est surtout la difficulté d'obtenir les données d'après le dessin initial, les coûts de l'informatisation manuelle ou semi-automatique, aussi bien que de la

conversion, ce qui nous a fait adopter un autre schéma. Actuellement, nous avons établi un réseau de fournisseurs de données électroniques (des institutions publiques et privées qui ont accepté de nous fournir des documents électroniques). Nous acceptons tout ce qu'ils nous donnent et nous constituons un grand corpus général. Si, par exemple, les directeurs des journaux nous autorisent à copier via internet des numéros complets, nous les gardons. Si quelques éditeurs de livres littéraires ou techniques nous offrent les ouvrages complets, nous les introduisons tels quels dans le *corpus*. Après, pour l'exploitation, et selon les objectifs de chaque projet de recherche, nous dessinons un *corpus* qui est à chaque fois extrait du grand *corpus* général.

Par exemple, pour notre Dictionnaire des Combinatoires du Portugais nous avons extrait un *corpus* oral et écrit de près de 13 millions de mots. Il y a actuellement des étudiants et des professeurs qui mènent aussi des recherches sur les associations lexicales et qui utilisent notre programme; mais ils se proposent de l'utiliser sur des *corpora* de spécialité, par exemple, juridique ou politique, que nous constituons à partir du *corpus* général pour qu'ils puissent les comparer avec le dictionnaire de combinatoires.

Il faut encore dire que notre participation aux projets européens (NERC, PP-PAROLE et LE-PAROLE) a été assez fructueuse pour nous. En ce qui concerne la représentation textuelle, l'utilisation du SGML a montré combien notre représentation était faible et comment ce standard international simple et formellement assez rigoureux permet une caractérisation et représentation plus exhaustive des documents; de même, son association à des outils informatiques internationaux adéquats va permettre un très grand nombre d'applications.

Pour le moment, nous regrettons ne pas avoir les moyens pour agrandir le *corpus* oral qui est resté très petit, car le recueil des données, le travail de transcription et d'édition et le traitement automatique de ce genre de *corpus* est vraiment très cher. Nous avons, en tout cas, une partie de ce *corpus* sur

CD-ROM, ce qui peut rendre service à beaucoup d'utilisateurs.

En ce qui concerne l'organisation des données, en général, nous regrettons de ne pas avoir les moyens financiers pour soutenir des équipes qui pourraient étudier et concrétiser quelques propositions qui ont été faites pour la classification des textes selon des critères internes, c'est-à-dire, des critères linguistiques. Je parle, en particulier, des propositions de Halliday et Martin, des analyses factorielles de Biber et Finegan, où des propositions de J. Sinclair. D'après ce dernier auteur, une classification rigoureuse devrait tenir compte d'une combinaison équilibrée des deux types de critères, les externes et les internes, qui se reflètent mutuellement. Pour nous, c'est un projet utopique, tout au moins dans un court délai. Je crois, par exemple, que les auteurs du *corpus* du GARS, qui a justement été analysé au fur et à mesure qu'il était constitué, sont très bien placés pour montrer quels seront les phénomènes linguistiques du français parlé à observer pour la caractérisation des textes en vue d'un classement rigoureux, ayant recours aux critères linguistiques. Je pense que ce travail pourrait constituer, surtout pour les *corpora* oraux des langues romanes, la base d'une typologie textuelle de référence.