Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie

Texte préparé par Elizabeth Marshman, janvier 2003

Ce document présente un court résumé des critères de sélection et de gestion de textes pour des travaux sur corpus. La section 1 présente quelques renseignements de base. La section 2 décrit les discussions théoriques à propos de certains critères de sélection de textes. La section 3 liste les critères proposés pour ses travaux sur corpus de l'équipe ÉCLECTIK, la composante « terminologie » de l'Observatoire de linguistique Sens-Texte (OLST) de l'Université de Montréal, et la façon de consigner les informations pertinents dans la base de données utilisée pour la gestion des corpus. La section 4 décrit la procédure pour la suppression des textes du corpus. Finalement, la section 5 contient les références des ouvrages cités, ainsi que d'autres ouvrages consultés et qui serviraient de ressources supplémentaires.

1. Travaux sur corpus : Un court résumé

1.1 Définition d'un corpus

« A collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis » (Francis 1982 cité dans Francis 1992 : 17).

1.2 Contextes du travail sur corpus

- lexicographie
- terminologie
- traduction (et traduction automatique)
- traductologie
- enseignement de la traduction
- linguistique comparée
- traitement automatique de langage (TAL) (par exemple, études d'usage ou de fréquence d'éléments linguistiques pour développer des méthodes de classification de documents, pour la synthèse automatique, etc.)
- mise au point de bases de données linguistiques
- ...

Puisque la linguistique de corpus est mieux établie dans le domaine de la langue générale que dans les langues de spécialité, beaucoup de ce qui s'écrit sur le travail sur corpus traite des corpus de langue générale. (Les exceptions dignes de mention sont par exemple Ahmad (1995), Bowker (1996), Meyer et Mackintosh (1996) et Pearson (1998).) Pour la conception de travaux sur corpus de langues spécialisées, il est souvent nécessaire de reprendre et puis d'adapter les procédures établies pour des projets portant sur la langue générale.

1.3 Critères importants pour l'utilité du corpus : constitution et équilibre

« The advice on corpus creation is to agree the smallest set of criteria that can be justified in the circumstances, so that the number of different documents is as small as possible » (Sinclair 1991 : 20).

Les questions qui reviennent dans la conception de tout corpus comprennent : le type de corpus, l'adéquation pour le projet visé, la possibilité de réutiliser ou d''interchanger ces corpus, la taille, la représentativité (c'est-à-dire, la variété de textes, d'auteurs, de sources, etc.), l'utilisation de textes complets ou d'échantillons, l'annotation et la gestion méticuleuse. Bien sûr, certains de ces critères sont difficiles à équilibrer entre eux et représentent des difficultés dans la construction du corpus.

Selon les sources consultées, les clés seraient de bien définir les attentes, les critères et les procédures à utiliser avant de commencer le processus de construction du corpus, et de bien décrire ces derniers dans toute présentation des résultats de la recherche (Collins et Peters 1987 : 106). On conseille aussi (Pearson 1998 : 51, Biber 1993 : 256, cité dans Pearson 1998 : 58) un processus cyclique, avec une réévaluation et ajustement des critères en cours de route, si nécessaire.

1.3.1 Types de corpus

Pearson (1998 : 44-48) décrit plusieurs types de corpus selon les applications possible de chaque type. Par exemple, dans les travaux de l'équipe ÉCLECTIK, on aura affaire surtout à des corpus dits « spécialisés » ou « de spécialité », à savoir, des corpus contenant des textes traitant d'un sujet lié à un domaine de la connaissance comme la médecine, le droit ou l'informatique. Pearson, cependant, ne fait pas une distinction de base entre des corpus *ouverts* (auxquels on ajoute constamment de nouveaux textes) et *fermés* (qui restent « stables »). Les corpus ouverts, bien que plus à jour, impliquent un entretien constant et méticuleux. Ceci les rend plus difficile à gérer pour des projets comme ceux de l'équipe.

Certains chercheurs favorisent l'utilisation d'un corpus *témoin* dans leurs travaux pour tenir à jour le contenu du corpus. Ce corpus moniteur contient des textes entiers, dont on prélève éventuellement des échantillons pour intégration au corpus principal (1987 : 106, 107). Comme McEnery et Wilson (2000) indiquent, cependant, un corpus moniteur peut fournir des données qualitatives sur de nouveaux mots ou usages, mais sont moins utiles pour des données qualitatives. Donc, l'utilité d'un corpus moniteur varie selon le projet en cours.

1.3.2 Adéquation pour le projet

Il est clair que la qualité des résultats d'un travail sur corpus dépend en grande partie de la qualité du corpus. Ceci implique divers facteurs :

- que le domaine des textes dans le corpus soit bien défini et délimité ;
- que les textes soient assez représentatifs pour appuyer les conclusions qu'on en tire ;
- que l'organisation, l'annotation, et le contenu du corpus favorisent son exploitation.

Ceci nécessite la spécialisation des techniques de construction des corpus selon le projet en question, ainsi qu'une bonne planification de ce processus.

1.3.3 La réutilisabilité / l'interchangeabilité

Ce critère, qui à première vue pourrait sembler aller à l'encontre de celui qui le précède, doit être équilibré avec ce dernier.

Bowker (1996 : 36) souligne l'importance d'un système de gestion polyvalent et complet afin de faciliter l'utilisation, le partage, et la modification de corpus. Il doit être possible d'adapter le corpus à des besoins de projets autres que celui pour lequel le corpus a été créé. Cette nouvelle application pourrait être similaire ou complètement différent de l'application originale.

Ce critère est mentionné dans la plupart des sources consultées ; ceci est logique, étant donné l'investissement de temps et d'effort nécessaire pour construire un corpus bien équilibré. Concrètement, cette réutilisabilité se manifesterait par une indexation

méticuleuse et complète des textes, ainsi que la documentation des procédures de recherche utilisées pour les trouver. Ceci permettrait de créer des sous-corpus selon divers critères avec un certain degré de fiabilité.

1.3.4 La taille

Il n'y a malheureusement pas de consensus sur la taille souhaitable pour un corpus, ni en langue générale, ni en langue spécialisée. Le seul point certain semble être qu'un corpus général doit être beaucoup plus grand qu'un corpus spécialisé. Pearson (1998 : 56-7) indique que, la plupart du temps, la taille d'un corpus spécialisé est déterminée selon l'intuition des chercheurs, le caractère des textes inclus et leur domaine. Elle donne comme chiffre moyen un million de mots. Cependant, la taille est souvent limitée par la disponibilité des textes et par des questions de droits d'auteur.

1.3.5 La représentativité et l'équilibre

Les critères de la représentativité et d'équilibre incluent entre autres la variété d'auteurs, de textes, de sources, de genres, de situations communicatives et de niveaux de spécialisation …

La représentativité d'un corpus est sa capacité de fonctionner comme une base fiable pour des généralisations sur une langue particulière (générale ou spécialisée). Ceci implique à la fois une taille et une variété de textes adéquates.

Comme le souligne Sinclair (1991 : 17), l'équilibre et la variété du contenu d'un corpus sont importants pour neutraliser les particularités des auteurs, sauf si le chercheur désire les analyser. Mais cette variété devrait s'appliquer non seulement aux auteurs, mais aussi aux sous-domaines d'un champ sous étude, aux sources, aux niveaux de spécialisation et aux situations communicatives des textes, et à plusieurs autres éléments des textes. (Bien sûr, il y a aussi la possibilité de délimiter un corpus par un de ces facteurs.)

Il y a un certain débat sur les critères d'équilibre de corpus. Certains chercheurs essaient de refléter les proportions de textes *produits* dans le domaine sous étude dans leurs corpus, d'autres les proportions de textes *reçus*. La façon de déterminer ces proportions n'est jamais claire et nette. Étant donné la difficulté de créer un corpus parfaitement équilibré, il est généralement considéré que le chercheur devrait établir ses buts en matière d'équilibre au début du projet (les modifiant si nécessaire au cours du travail) et décrire clairement ses procédures dans le rapport des résultats.

1.3.6 Échantillonnage

Sinclair (1991), Pearson (1998 : 60) et Meyer et Mackintosh (1996) favorisent l'utilisation de documents entiers dans leurs corpus, surtout lorsqu'il s'agit de recherches terminologiques ou conceptuelles. Cependant, ceci soulève le problème d'équivalence de représentation des textes dans le corpus, et remet en question la pratique bien établie d'échantillonnage tel qu'utilisée pour des corpus de langue générale.

Si on décide d'avoir recours à l'échantillonnage, il sera nécessaire de bien définir la procédure à suivre avant de commencer. L' utilisation de parties du début, du milieu et de la fin des textes est à conseiller.

D'un point de vue plus large, lorsqu'on recherche des textes dans un très grand répertoire de textes possibles, il peut être souhaitable de prendre un échantillon au hasard suivant un système bien défini (par exemple, en prenant chaque 10e ou 20e texte de la liste). Ceci peut prémunir le chercheur contre un biais dans la sélection des textes. Si un tel système est utilisé, il devrait être noté.

1.3.7 L'annotation

1.3.7.1 Étiquetage des catégories grammaticales

Pour les projets de l'équipe ÉCLECTIK, l'étiquetage des catégories grammaticales est souvent — sinon toujours — nécessaire. Ainsi, il sera également nécessaire de noter le logiciel ou le jeu d'étiquettes utilisé pour faire l'étiquetage de chaque document. Il sera aussi souhaitable de suivre les conseils de Sinclair (1991) — et la pratique actuelle — et de conserver une copie du texte sans étiquetage dans un répertoire séparé.

1.3.7.2 Éléments paratextuels

Selon Meyer et Mackintosh (1996), entre autres, il peut être utile d'annoter les textes du corpus pour indiquer des attributs du texte, des titres et sous-titres, etc. Si on utilise un format de fichier tel que SGML, HTML ou XML, ces éléments sont en grande partie déjà indiqués.

Sinclair (1991 : 20-21) décrit le Text Encoding Initiative (voir TEI 1994) pour normaliser les conventions pour le stockage des textes, incluant des critères de l'information sur les textes et un codage (balisage) qui est facile à séparer du texte même et à convertir vers un autre système d'identification et de classification de divers codes (par exemple, police, formatage, références, etc). Dans l'intérêt de rendre les corpus interchangeables, si on décide d'annoter les textes des corpus, il faudrait utiliser une norme telle que le TEI / Corpus Encoding Standard (CES) ou une autre méthode d'annotation répandue.

1.3.8 La gestion des corpus de langue générale

« *It is advantageous to keep detailed records of the material so that the documents can be identified on grounds other than those which are selected as formative in the corpus* » (Sinclair 1991 : 20).

Quelques critères utilisés pour la classification comprennent :

- la taille ;
- la délimitation du domaine ;
- la sélection de type de corpus : ouvert/fermé ; synchronique/diachronique ;
- la sélection de textes *appropriés au but de la recherche* :
    - variété d'auteurs
    - variété de sources
    - variété de buts des textes (« purposes » (Sinclair 1991 : 13))
    - longueur comparable de textes

- date des textes
- niveau de spécialisation : approprié au but de la recherche, uniforme, auteur et destinataire établis
- langue originale du texte, langue maternelle de l'auteur.

Pour un corpus lexicographique général, Sinclair (1991 : 20) indique plusieurs critères qui pourraient être utilisés pour le tri :

- longueur de texte ;
- fiction/non;
- support (livre/revue/journal) ;
- formel/informel ;
- âge de l'auteur ;
- sexe de l'auteur ;
- nationalité et/ou langue maternelle de l'auteur.

Kaye (1987 :151) souligne l'importance de maintenir la liste de documents dans le corpus sous un format qui permet l'ajout de catégories et sous-catégories, de trier des documents selon divers critères, de faire des ajouts ou modifications globaux à l'index si nécessaire.

1.3.9 Difficultés dans la construction et gestion des corpus

Collins et Peters (1987 : 106) soulignent les difficultés qui peuvent survenir lors de la classification de textes selon le type de discours (informatif, instructif, discursif).

Kytö et Rissannen (1987 : 171-2) confirment la difficulté de bien classifier le style d'un document, et dans leurs travaux ont choisi d'indiquer le type de texte, la relation avec un ouvrage étranger (s'il y a lieu), la relation entre les participants (auteur/destinataire), le niveau de formalité, le statut socio-éducationnel de l'auteur, etc. et de laisser les lecteurs tenir leurs propres conclusions. Cette approche pourrait être adaptée pour les besoins de l'équipe ÉCLETIK (voir la section 3).

En ce qui concerne le problème de classification des textes, Biber et Finegan (1986) ont fondé leur typologie sur les caractéristiques des textes en les combinant entre-elles, créant ainsi une classification multidimensionnelle. Ils différencient ainsi le genre de texte (déterminé par facteurs externes, par exemple le médium ; lié au but de l'auteur) du type de texte (déterminé par les facteurs linguistiques internes mentionnés ci-dessus ; lié à la forme du texte). Les textes dans le corpus ont été par la suite regroupés en utilisant ces critères pour identifier des similarités.

2. Sélection de textes

Pour des fins de terminologie, Meyer et Mackintosh (1996) rappellent le besoin d'un corpus qui consiste en un grand nombre de documents, qui contiennent un grand nombre de termes avec définitions ou contextes explicatifs ou définitoires.

Les critères de sélection de textes forment aussi les éléments important pour la gestion du corpus.

2.1.1 Nom de fichier

Un peu comme Kytö (1987), Meyer (TRA6905, hiver 2000) conseille l'utilisation d'un système de noms de fichier qui servira au tri ultérieur des textes selon divers critères. Le système est défini selon les critères de sélection des textes. (Par exemple, un système indiquant la langue, le domaine, la source, la date, le niveau de spécialisation ou toute autre information pertinente pourrait être envisagé.)

2.1.2 Langue

Le plus souvent, on préfère des textes originaux (et non pas des traductions) écrits dans la langue maternelle de l'auteur. Cependant, ceci peut être difficile à vérifier ; de plus un texte écrit dans une langue seconde peut néanmoins être fiable et pertinent. Il s'agit de bien définir la politique à utiliser dès le début de la construction du corpus, et de noter ces renseignements pour chaque texte.

2.1.3 Genre du texte

Ahmad (1995) indique différents genres de textes, tels que : revues spécialisés, vulgarisation scientifique, manuels techniques, manuels scolaires et universitaires, journaux, et publicités comme des sources de textes pour des travaux en langue spécialisée.

2.1.4 Niveau de spécialisation

Meyer et Mackintosh (1996) favorisent l'inclusion de textes pédagogiques, avancés, vulgarisés, de différents genres (pour donner accès à des variantes de termes ou synonymes utilisés à différents niveaux). Pearson, par contre, se limitait surtout à des situations communicatives entre experts et entre experts et un public spécifique visé pour son projet (1998 : 60).

Il serait de toute évidence possible de laisser ce critère sous-entendu et d'utiliser les deux champs suivants pour déterminer le niveau de spécialisation du texte.

2.1.5 Qualifications/expertise de l'auteur

Il est généralement convenu (par exemple, Pearson 1998 : 60) qu'il faudrait utiliser des textes provenant d'auteurs reconnus dans leur domaine, par exemple par une institution ou par ses collègues. Un point souligné par Pearson et également par Meyer (2000) est la nécessité de vérifier non seulement les qualifications d'un auteur dans un domaine, mais aussi qu'il écrit bel et bien dans son domaine d'expertise.

L'auteur constitue aussi la première partie de la situation communicative décrite par Pearson (1998 : 35-39). Elle divise les auteurs de textes spécialisés dans les catégories suivants, *expert* (expert), *relative expert* (semi-expert), et *teacher* (enseignant). Pour des fins d'études terminologiques, il n'est pas réellement nécessaire d'avoir recours à une classification pour des auteurs qualifiés de *laypeople (non-experts)*, et des textes écrits par des non-experts sont généralement à éviter dans le contexte d'un travail terminologique.

2.1.6 Destinataire

Généralement, les textes sont aussi différenciés selon le public visé (*expert*, *initiate*, *uninitiated*, *student*), ce qui constitue le deuxième élément de la situation communicative décrite par Pearson.

Ceci est souvent fait selon la source, mais aussi peut se faire selon certains critères intratextuels (par exemple, la densité de la terminologie, explications des termes, etc.) (Brekke 1999). Cependant, il n'est pas toujours très facile à déterminer.

2.1.7 Auteur

Il est essentiel que les textes inclus dans le corpus proviennent d'une variété d'auteurs, pour minimiser l'impact de particularités d'auteurs individuels (cf. entre autres, Sinclair (1991) et Meyer et Mackintosh (1996)) Il serait aussi souhaitable d'inclure des auteurs de différentes écoles, étant donné que les étudiants ont tendance à utiliser la terminologie enseignée par leurs professeurs.

Selon le projet, on pourrait envisager d'ajouter des critères pour indiquer le pays d'origine du texte ou de l'auteur, et, éventuellement, sa langue maternelle et pays d'origine.

2.1.8 Référence

Cette information permet d'identifier complètement le texte, de vérifier la version du texte utilisé et d'établir sa provenance, mais aussi d'éviter d'utiliser trop de textes de la même source. Ce sera aussi important pour des questions de droits d'auteur. Le titre du texte peut également révéler des notions et termes clés dans le texte.

La référence sera notée dans un format normalisé, pour éviter la nécessité de retravailler toutes les références lors de leur utilisation ultérieur. Il existe plusieurs formats possibles qui pourraient être utilisés selon la préférence des chercheurs.

Pour des textes provenant d'Internet, Meyer (2000) a suggéré un système d'ajouter de l'URL à l'intérieur de balises dans le texte, surtout lorsque l'outil utilisé pour l'analyse permet d'exclure le contenu de certains types de balises de l'analyse (comme c'est le cas avec WordSmith Tools[1]). Dans certains cas, cette pratique pourrait fausser des analyses, mais en revanche permet d'assurer que la source du texte sera toujours identifiable.

2.1.9 Date

Ce critère découle du projet de recherche et du type de corpus à construire (synchronique, diachronique, etc.). Pour la plupart des travaux en terminologie, les textes choisis sont assez récents ou portent sur des sujets pertinents actuellement.

Cependant, ces critères varient beaucoup, dépendant du projet et du but de la recherche. Bowker (1996) indique que pour des traducteurs, des textes légèrement plus anciens

[1] Un suite d'outils pour l'analyse linguistique conçue par Mike Scott de l'University of Liverpool [http://www.lexically.net/wordsmith/].

peuvent être utiles pour trouver des descriptions de concepts ou comme ressources pour des traductions contemporaines de textes plus anciens. De plus, Meyer et Mackintosh (1996) indiquent que, pour des terminologues, des ressources qui datent de l'introduction d'un concept donné peuvent contenir beaucoup plus de contextes descriptifs et explicatifs sur ce concept, comparativement aux textes qui décrivent un concept ou un terme bien établi.

2.1.10 Format

En ce qui concerne l'exploitation facile des textes avec divers outils d'analyse, ainsi que le stockage de ces documents, il est généralement souhaitable d'enregistrer les fichiers en format ASCII. Cependant, pour des projets qui impliquent une analyse de la structure du document ou d'autres éléments liés au formatage, ceci pourrait entraîner une perte d'informations. L'utilisation du format HTML ou XML pourrait alors être une meilleure option.

2.1.11 Autres questions

2.1.11.1 Autorisation en ce qui concerne les droits d'auteur

La permission de reproduire des textes pose souvent problème dans le domaine des recherches sur corpus. Une stratégie consiste à définire une politique d'utilisation (assurance contre le piratage et l'utilisation abusif) et d'en informer les auteurs des textes en demandant la permission (Sinclair 1991 : 15). Dans tous les cas, on exige de noter la référence complète pour tout texte ajouté au corpus.

2.1.11.2 « Typicality »

Ce qui est nouveau, brillant, qui fait fureur, n'est souvent pas représentatif. Sinclair (1991 : 17) note que, bien qu'il soit permis d'inclure des textes de ce genre dans un corpus, ils devraient être complétés par un grand nombre de textes plus « neutres ».

3. Procédures pour l'équipe ÉCLECTIK

Cette section décrit la procédure suggérée pour les travaux sur corpus spécialisés par l'équipe ÉCLECTIK. Les renseignements sur chaque nouveau texte ajouté au corpus de l'équipe sont inscrits dans une base de données conçue à cette fin.

3.1 Recherches

Les recherches de textes sont faites à l'aide de moteurs de recherche sur Internet, de bases de données, et d'autres ressources. Il sera important de garder la trace des recherches réalisés au moyen des techniques propres à ces ressources. Ceci permettra d'éviter le dédoublement du travail et d'en revoir si nécessaire la méthode. De plus, cela facilite la constitution de sous-corpus.

Les mots-clés de la recherche pourraient servir aussi d'aide à identifier le sujet des documents.

Pour chaque recherche faite, un code identificateur unique composé de la date et le numéro de la recherche, le moteur et le(s) mot(s)-clé(s) sont notés sur le formulaire Access associé à la table « Recherches » de la base de données «Gestion de corpus »

(corpus_management.mdb). S'il y a lieu, la portée de la recherche (par exemple, le nombre de pages de résultats, le système d'échantillonnage des résultats consultés) est également notée dans le champ approprié.

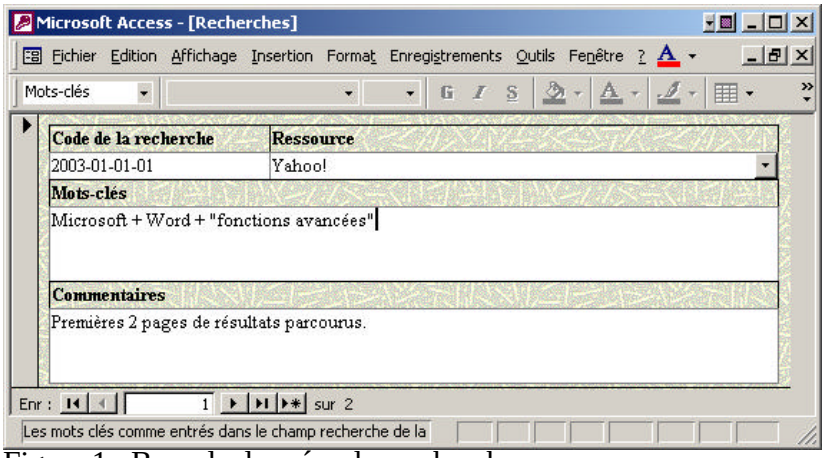

Figure 1 : Base de données des recherches

3.2 Critères de sélection de textes

Une fois des textes trouvés, on applique plusieurs critères de sélection pour décider s'ils doivent être ajoutés dans le corpus. Les résultats de cette analyse sont insérés dans le formulaire Access pour chaque texte choisi. Si le formulaire s'avère incomplet ou n'offre pas le choix désiré, on peut ajouter l'option nécessaire dans le fichier et avertir les autres membres de l'équipe du changement apporté. S'il s'agit d'un cas isolé, la base accepte pour l'instant les éléments qui n'apparaissent pas dans la liste des choix.

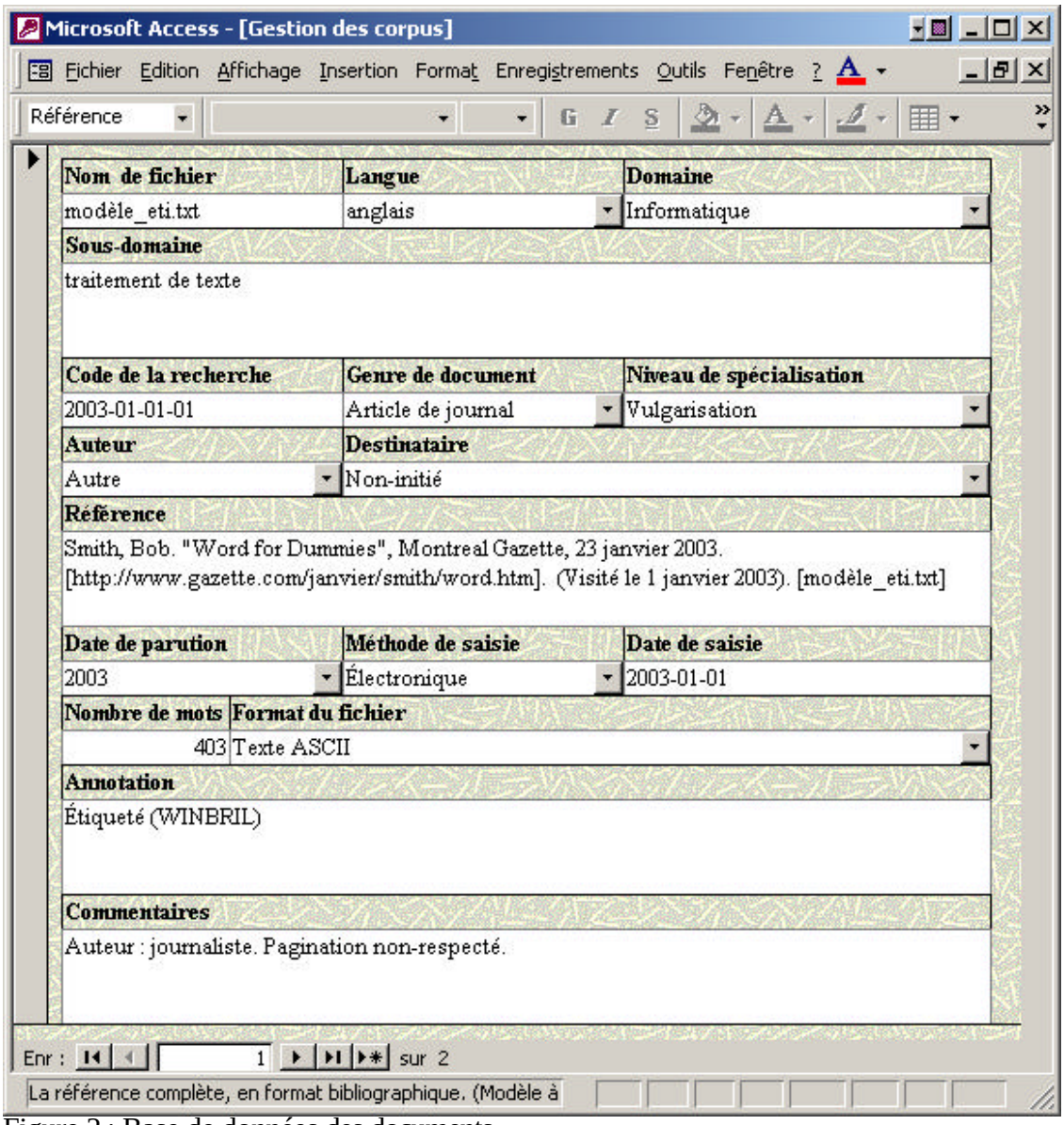

Figure 2 : Base de données des documents

3.2.1 Nom de fichier

Le nom de fichier servira d'identificateur *unique* pour chaque document. Il est donc important de bien noter ce nom sur le formulaire Access, dans le champ « Nom de fichier ».

Ce nom de fichier indique le plus souvent le sujet du texte. Il est souhaitable de limiter le nom de fichier à 8 caractères.

3.2.2 Langue

Évidemment, dans tout travail linguistique, ce sera le critère le plus important pour le tri des textes.

Il est conseillé de prendre des textes originaux et d'éviter des traductions (qui risquent d'être des re-rédactions par des non-experts), et lorsque possible de prendre des textes écrits dans la langue maternelle de l'auteur (ou au moins dans une langue que ce dernier maîtrise très bien). Il convient également de se méfier de tout texte qui semble contenir beaucoup d'erreurs d'ordre linguistique

La langue est choisie à partir de la liste dans le formulaire Access associé à la table « Gestion de corpus ».

3.2.3 Domaine et sous-domaines

Le domaine sur lequel porte les recherches devrait être bien défini et délimité avant de commencer la construction du corpus. Par la suite, il faut déterminer avec précision si le texte traite du domaine en question.

Par exemple, si on ajoute un texte au corpus juridique, on pourrait ajouter un texte journalistique sur une décision importante d'une Cour, mais non pas un éditorial qui évalue *l'impact social* de cette décision. Il est aussi important d'identifier le sous-domaine, ce qui permet non seulement de construire des sous-corpus de textes, mais aussi de revoir la gamme de sous-domaines représentés, et donc l'équilibre du corpus.

Le domaine général est choisi à partir de la liste dans le formulaire Access. Les domaines disponibles pour l'instant sont l'informatique, le droit, la mécanique, et la médecine. Le sous-domaine est tapé par l'utilisateur dans le champ approprié. Si aucun sous-domaine n'est identifié, ce champ reste vide.

3.2.4 Code de recherche

Le code de la recherche dont provient le texte est reproduit tel qu'il apparaît dans la table « Recherches » dans la base de données « Gestion de corpus » (voir la section 3.1).

3.2.5 Genre de document

Le genre du texte indique souvent son niveau de spécialisation et les qualifications d'un auteur. Par exemple, un texte d'une revue spécialisée est généralement plus approprié pour un corpus spécialisé qu'un article journalistique, qui risque d'être écrit par un non-expert pour un public qui a peu de connaissances du domaine.

Le genre du texte est choisi à partir de la liste dans le formulaire. Les options disponibles incluent : article de journal ; article de revue générale ; article de revue spécialisée ; article de journal académique ; article de vulgarisation ; manuel technique ; manuel de cours ; notes de cours ; autre ouvrage.

3.2.6 Niveau de spécialisation

Pour des corpus spécialisés, on cherche le plus souvent des textes écrits par des personnes qui ont certaines connaissances dans le domaine, pour un public qui en partage une certaine proportion de ces connaissances. Donc, pour évaluer le niveau de spécialisation, on considère le genre du texte, l'auteur et le public visé.

La plupart du temps, on a tendance à éviter des textes très informels ou personnels et ceux qui sont écrits par des membres du grand public, qui risquent de contenir du vocabulaire et de la phraséologie inexacts.

Le niveau de spécialisation est sélectionné à partir de la liste dans le formulaire. Les options disponibles incluent : très technique, technique, spécialisé, vulgarisation, formation, général.

3.2.7 Qualifications/expertise de l'auteur

Les qualifications de l'auteur sont très importantes pour des travaux sur corpus spécialisés. L'auteur du texte devrait être un expert dans le domaine, ou au moins un semi-expert. Il ou elle devrait être reconnu par ses pairs ou par une institution (par exemple, il ou elle publie des ouvrages dans le domaine, est associé avec une université ou un centre de recherche, le texte apparaît sur un site officiel).

En choisissant une des options de la liste sur le formulaire, on indique les qualifications de l'auteur. Les options disponibles incluent : expert, semi-expert, enseignant.

Si on n'est pas capable d'identifier ces informations, on peut choisir l'option « Inconnu ». Dans ce cas, cependant, il sera d'autant plus important de vérifier les autres critères pour être certain de la qualité du texte.

3.2.8 Destinataire

Ce critère est étroitement lié avec le genre et le niveau de spécialisation du texte. C'est à partir de ces critères qu'on peut deviner le public visé par le texte.

On choisit de la liste sur le formulaire l'option qui représente mieux ce destinataire. Les options disponibles incluent : expert, initié, non-initié, étudiant, autre.

Par exemple, pour un cours de biochimie de premier cycle universitaire, on pourrait classifier le destinataire comme « étudiant ». Pour une monographie pharmaceutique destiné aux médecins, on classifierait le destinataire comme « initié ».

Si on n'est pas capable d'identifier le niveau de formation du public visé par un document, on peut aussi choisir l'option « Inconnu », à condition de bien vérifier les autres critères pour être certain que le texte est admissible au corpus.

3.2.9 Référence

Dans ce champ, on tape la référence bibliographique *complète* du document, y compris l'*URL* et la *date de consultation* du site pour les documents puisés sur Internet. Cette référence prendra la forme suivant :

3.2.9.1 Pour un ouvrage publié

AUTEUR, Le (2003). Titre de l'ouvrage, collection, volume, Maison d'édition, Ville. [nom de fichier]

AUTEUR, Le (2003). « Titre du chapitre », Titre de l'ouvrage, Nom du directeur (dir), Maison d'édition, Ville, pages. [nom de fichier]

3.2.9.2 Pour un article publié

AUTEUR, Le (2003). « Titre de l'article », Titre de la périodique, volume(numéro), pages. [nom de fichier]

AUTEUR, Le (2003). « Titre de l'article », Titre du journal, jour mois année, sectionpage. [nom de fichier]

3.2.9.3 Pour un site Web

AUTEUR, Le (2003). « Titre de la page ». [http://www.url.com]. (Visité le jour mois année). [nom de fichier]

AUTEUR, Le (2003). « Titre de la page », Organisation, Titre de publication en ligne, volume(numéro). [http://www.url.com]. (Visité le jour mois année). [nom de fichier]

Cette référence est aussi copiée et collée au début du texte lui-même, pour servir de repère pour assurer qu'elle ne sera pas perdue.

3.2.10 Date de parution

Pour les besoins de notre recherche, l'équipe veut inclure des textes de diverses années dans les corpus. Cependant, et surtout pour les domaines en pleine évolution, tels que l'informatique, le plus important est d'inclure une majorité de textes récents. Généralement, on s'intéresse à des textes écrits depuis 1990.

Dans ce champ on entre la date de parution de l'ouvrage ou de l'ajout au site Web. Si, dans ce dernier cas, on ne peut trouver cette information, mais le caractère du texte ne fait pas de doute, on peut laisser ce champ vide. Si on connaît la date de mise à jour du site, on note cette information dans le champ « Commentaires » du formulaire.

3.2.11 Méthode de saisie

On indique ici si le texte a été numérisé ou si c'était en format électronique à l'origine, en choisissant soit *électronique*, soit *numérisé* à partir de la liste.

3.2.12 Date de saisie

On indique dans ce champ la date de saisie du texte (c'est à dire, la date de consultation d'une page Web, ou la date de reconnaissance optique de caractères pour des textes numérisés). Cette information prend le format aaaa-mm-jj.

3.2.13 Nombre de mots

Le décompte de mots se fait dans Word et le résultat est tapé dans le champ approprié du formulaire.

3.2.14 Format de fichier

Les textes sont généralement stockés en format .txt pour faciliter leur exploitation avec divers outils (concordanciers, étiqueteurs, etc.). Le plus simple est de copier le texte et de faire un collage spécial comme texte non formaté dans un document Word, puis d'enregistrer ce document en format Texte seulement (.txt).

D'autres formats sont cependant disponibles dans le formulaire, pour des cas où le document ne serait pas enregistré en format .txt. Ces formats incluent SGML, HTML, XML, .rtf, .doc, etc.

3.2.15 L'annotation

On gardera une copie de tout document original. Les textes étiquetés pour les fins des recherches seront enregistrés sous le même nom de fichier avec l'ajout de « _eti » à la fin, et seront stockés dans un répertoire séparé.

Le plus simple est de copier tous les textes à étiqueter dans ce répertoire et puis de les étiqueter, afin de ne pas écraser l'original accidentellement.

S'il y a lieu, la mention « étiqueté » et le logiciel utilisé pour étiqueter le texte sera noté sur la fiche du document, dans le champ « Annotation ».

3.2.16 Commentaires

Ce champ sert à consigner toute information supplémentaire qui pourrait être utile. Ceci inclurait, par exemple, des commentaires sur le respect du formatage du texte, l'existence d'une traduction du texte, et des approfondissements de l'information inclus dans un autre champ du formulaire.

3.2.17 Autres questions

3.2.17.1 Corpus parallèle

Si on voit qu'il y a une traduction facilement accessible d'un document stocké, il serait intéressant d'enregistrer ce document aussi. Ceci permettrait éventuellement de construire un corpus parallèle ou un bitexte.

La traduction sera identifiée par le même nom de fichier que l'original, avec l'ajout de « _tra » à la fin, et sera enregistrée dans un répertoire séparé. Son existence pourra être notée sur la fiche du document original, dans le champ « Commentaires » du formulaire Access.

3.3 Suppression de documents

Si on supprime des fichiers du corpus, on coupe l'enregistrement associé au texte de la table « Gestion des corpus » et on le colle directement dans la table « Textes supprimés » en utilisant l'option Coller par ajout. À l'information déjà indiquée, on ajoute la date de suppression et la raison de la suppression dans les champs prévus à ces fins. Les principales raisons pour la suppression seraient les suivantes : *texte en double*, *texte vieilli*, etc.

4. Bibliographie, ouvrages cités et ressources supplémentaires

Aarts, J. et W. Meijs, dirs. (1986). *Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora.* Série : Costeurs, vol. 47. Amsterdam : Éditions Rodopi.

Aarts, J. et W. Meijs, dirs. (1990). *Theory and Practice in Corpus Linguistics.* Série : Language and Computers: Studies in Practical Linguistics, no. 4. Amsterdam : Éditions Rodopi.

Ahmad, K. (1995). « Machine Translation and the Lexicon ». Dans *Third International EAMT Workshop.* Steffens : Springer Verlag. pp. 51-76.

Ahmad, K. (1993). « Terminology and Knowledge Acquisition: A Text-Based Approach ». Dans Klaus-Dirk Schmitz (dir.) *TKE'93: Terminology and Knowledge Engineering, Proceedings of the Third International Congress on Terminology and Knowledge Engineering,* 25-27 août 1993, Cologne, Allemagne. Frankfurt: Indeks Verlag. pp. 56-70.

Association européenne pour les ressources linguistiques (ELRA). [http://www.icp.inpg.fr/ELRA/fr/]. (Visité en janvier 2003).

Atkins, B.T.S. (1992). « Tools for C-A Corpus Lexicography ». *Papers in Computational Terminology. Proceedings of Complex '92.* pp. 1-60.

Atkins, S., J. Clear et N. Ostler. (1992). « Corpus Design Criteria ». *Literary and Linguistic Computing* 7(1). pp. 1-16.

Biber, D. (1993). « Representativeness in Corpus Design ». *Literary and Linguistic Computing* 8(4). pp. 243-257.

Biber, D. et E. Finegan. (1986). « An Initial Typology of English Text Types ». Dans *Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora.* Dirs. J. Aarts et W. Meijs. Série : Costeurs, vol. 47. Amsterdam : Éditions Rodopi. pp. 19-46.

Bowker, L. (1996). « Towards a corpus-based approach to terminography ». *Terminology* 3(1):27-52.

Brekke, M. (1999). « Popular vs. Professional Aspects of Economics Texts in English ». Hermes, Journal of Linguistics, 23. pp. 24-40.

Brekke, M., J. Myking et K. Ahmad. (1996). « Terminology Management and Lesser-Used Living Languages: A critique of the corpus-based approach ». *TKE'96: Terminology and Knowledge Engineering, Proceedings of the 4th International Congress on Terminology and Knowledge Engineering,* Vienne, 26-28 août 1996. Frankfort: Indeks-Verlag. pp.179-189.

Collins, P. et P. Peters. (1987). « The Australian Corpus Project ». Dans *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora.* Dirs. M. Kytö, O. Ihalainen et M. Rissanen. Série : Language and Computers: Studies in Practical Linguistics, no. 2. Dirs. J. Aarts et W. Meijs. Amsterdam : Éditions Rodopi. pp. 103-120.

Équipe de recherche en combinatoire lexicale, terminologie et informatique (ÉCLECTIK). (2003). « ÉCLECTIK : Accueil ». [http://www.fas.umontreal.ca/Ling/lhomme/eclectik.htm]. (Visité en janvier 2003).

Filmore, C.J. (1992). « 'Corpus Linguistics' or 'Computer-aided armchair linguistics' ». Dans *Directions in Corpus Linguistics: Proceedings of Novel Symposium 82.* Dir. J. Svartik. Stockholm, Suède, 4-8 août 1991. Série : Trends in Linguistics Studies and Monographs, no. 65. Berlin/New York : Mouton de Gruyter. pp. 35-60.

Francis, W.N. (1992). « Language Corpora B.C .» Dans *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82.* Dir. J. Svartik. Stockholm, Suède, 4-8 août 1991. Série : Trends in Linguistics Studies and Monographs, no. 65. Berlin/New York : Mouton de Gruyter. pp. 17-32.

Francis, W.M. (1982). « Problems of assembling and computerizing large corpora ». Dans *Computer Corpora in English Language Research.* Dir. S. Johanson. Bergen : Norwegian Computing Centre for the Humanities. pp. 7-24.

Glaser, R. (1982). « The Problem of Style Classification in LSP(ESP) ». In *Pragmatics and LSP, Proceedings of the 3rd European Symposium on LSP.* pp. 69-81.

Habert, B., A. Nazarenko et A. Salem. (1997). *Les Linguistiques de corpus.* Paris : Armand Colin.

Ide, N., G. Priest-Dorman et Jean Véronis. (1996). *Corpus encoding standard*. [http://www.cs.vassar.edu/CES/]. Cité dans J. Gamper, « Construction of a Parallel Text Corpus Encoding Primary Data ». *Academia* 18, mars 1999. (Visité en janvier 2003).

Kaye, G. (1987). « The Design of the Database for the Survey of English Usage ». Dans *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora.* Dirs. M. Kytö, O. Ihalainen et M. Rissanen. Série : Language and Computers: Studies in Practical Linguistics, no. 2. Dirs. J. Aarts et W. Meijs. Amsterdam : Éditions Rodopi. pp. 145-168.

Kaye, G. (1990). « A Corpus Builder and Real-Time Concordance Browser for an IBM PC ». Dans *Theory and Practice in Corpus Linguistics.* Dirs. J. Aarts et W. Meijs. Série : Language and Computers: Studies in Practical Linguistics, no. 4. Amsterdam : Éditions Rodopi. pp. 137-161.

Kytö, M. et M. Rissanen. (1987). « The Helsinki Corpus of English Texts: Classifying and Coding the Diachronic Part ». Dans *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora.* Dirs. M. Kytö, O. Ihalainen et M. Rissanen. Série : Language and Computers: Studies in Practical Linguistics, no. 2. Dirs. J. Aarts et W. Meijs. Amsterdam : Éditions Rodopi. pp. 169-179.

Kytö, M., O. Ihalainen et M. Rissanen, dirs. (1987). *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language*

Research on Computerized Corpora. Série : Language and Computers: Studies in Practical Linguistics, no. 2. Amsterdam : Éditions Rodopi.

McEnery, T. et A. Wilson. (2000). « ICT4LT Module 3.4 : Corpus Linguistics ». Information and Communications Technology for Language Teachers. [http://www.ict4lt.org/en/en_mod3-4.htm]. (Visité en décembre 2002).

Meyer, I. (2000). Cours TRA6905 : Terminologie et lexicographie. Université d'Ottawa, hiver 2000.

Meyer, I. et K. Mackintosh. (1996). « The Corpus from a Terminographer's Viewpoint ». *International Journal of Corpus Linguistics.* 1(2): 257-285.

Pearson, Jennifer. (1998). *Terms in Context.* Amsterdam/Philadelphia: John Benjamins.

Petek Kurtböke, N. (1998). « A Corpus-driven study of Turkish-English Language Contact in Australia ». Thèse doctorale, Monash University, Australie. [http://home.vicnet.net.au/~petek/thesis/]. (Visité en décembre 2002).

Renouf, A. (1987). « Coding Metalanguage: Issues Raised in the Creation and Processing of Specialized Corpora ». Dans *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora.* Dirs. M. Kytö, O. Ihalainen et M. Rissanen. Série : Language and Computers: Studies in Practical Linguistics, no. 2. Dir. J. Aarts et W. Meijs. Amsterdam : Éditions Rodopi. pp. 197-206.

Renouf, A. (1984). « Corpus Development at Birmingham University ». Dans *Corpus Linguistics.* Dirs. J. Aarts et W. Meijs. Amsterdam : Rodopi.

Renouf, A. (1987). « Corpus Development ». Dans *Looking Up: An Account of the COBUILD Project in Lexical Computing.* Dir. J. Sinclair. London/Glasgow : Collins.

Sinclair, J. (1991). « Chapter 1: Corpus Creation ». Dans *Corpus, Concordance, Collocation.* Série : Describing English Language. Dirs. J. Sinclair et R. Carter. Oxford : Oxford University Press.

Svartik, J. (dir). (1992). *Directions in Corpus Linguistics: Proceedings of Novel Symposium 82.* Stockholm, Suède, 4-8 août 1991. Série : Trends in Linguistics Studies and Monographs, no. 65. Berlin/New York : Mouton de Gruyter.

TEI (1994). « The Text Encoding Initiative ». [http://www.uic.edu/orgs/tei/]. (Visité en janvier 2003).