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

Texte préparé par Elizabeth Marshman, janvier 2003

1. 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.

1.1 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.

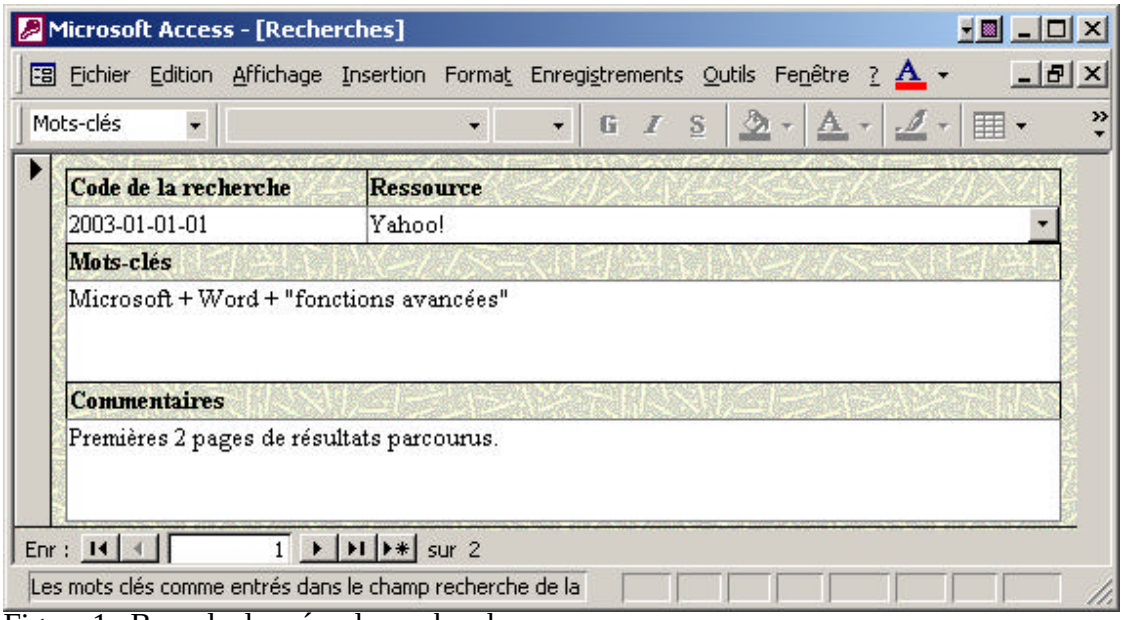

Figure 1 : Base de données des recherches

1.2 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.

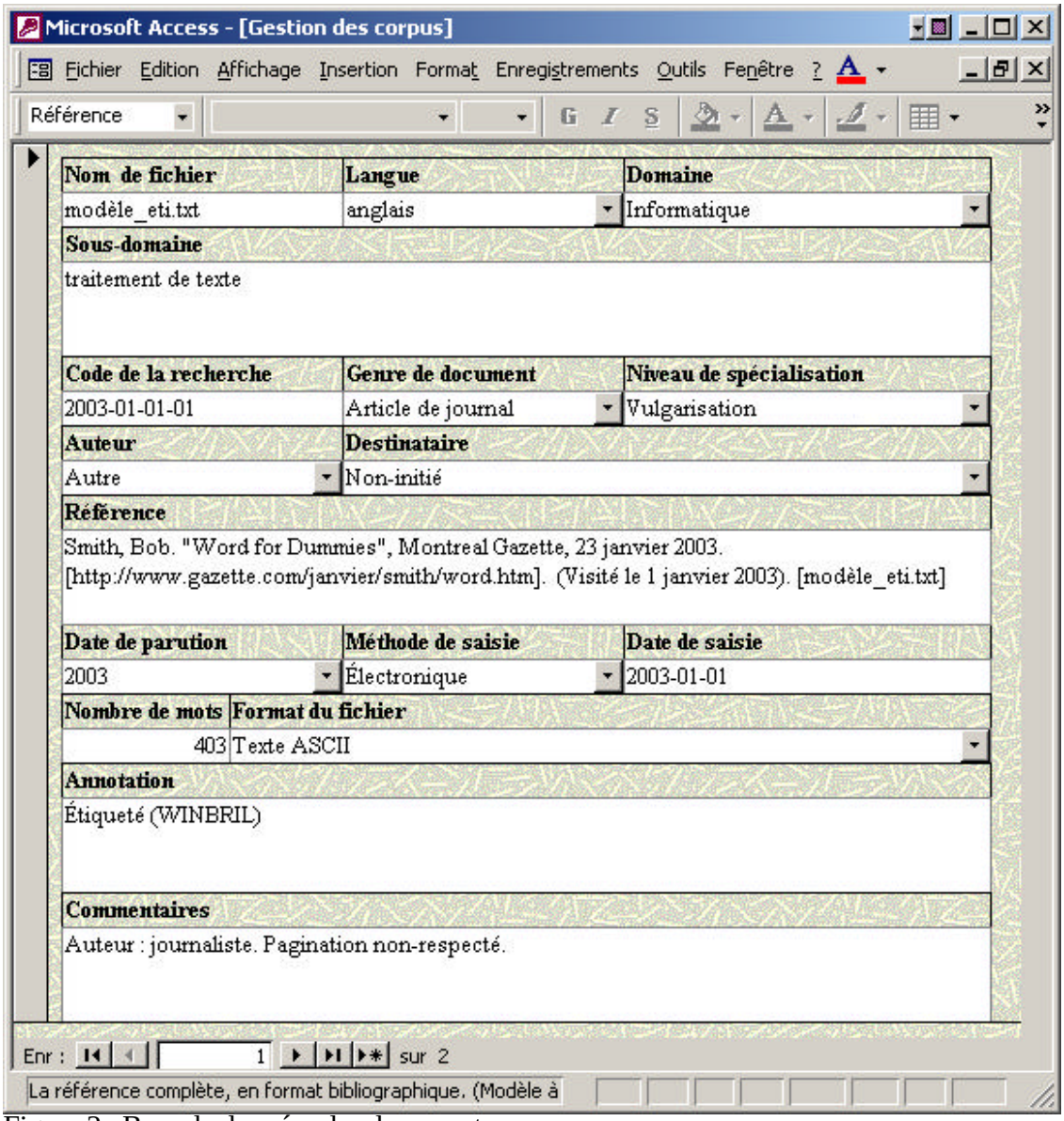

Figure 2 : Base de données des documents

1.2.1 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.

1.2.2 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 ».

1.2.3 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.

1.2.4 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).

1.2.5 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.

1.2.6 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.

1.2.7 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.

1.2.8 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.

1.2.9 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 :

1.2.9.1 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]

1.2.9.2 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]

1.2.9.3 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.

1.2.10 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.

1.2.11 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.

1.2.12 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.

1.2.13 Nombre de mots

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

1.2.14 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.

1.2.15 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 ».

1.2.16 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.

1.2.17 Autres questions

1.2.17.1 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.

1.3 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.