Dr. Miri- Benabdallah Imene,
MCB Université d'Oran.

5. Constitution d'un corpus linguistique pour une analyse textuelle du discours : Modalités et enjeux

Plan :

1- Introduction
2- Méthodologie et constitution d'un corpus de la presse
3- Le mode de numérisation
4- Conclusion

1. Introduction

Dans le cadre du projet PNR intitulé « Analyse des Discours et des Objets Signifiants », et dans un contexte de multiplication des sources pour la constitution des corpus linguistiques.
Nous nous sommes intéressés en tant que chercheurs du laboratoire LOAPL et en tant que membres de ce projet à la numérisation d'articles de presse.
Cet intérêt émane d'une part du fait que ce genre de corpus devient l'une des sources les plus sollicitées par les chercheurs en sciences humaines et sociales. Et d'autre part du fait que les chercheurs en Sciences du Langage rencontrent des difficultés pour l'établissement scientifique des données textuelles. Les questions que nous nous posons sont les suivantes :
quels sont les enjeux de la constitution du corpus numérisé ?
Et si des possibilités de documents numérisés sont ouvertes aux chercheurs, les utilisateurs que nous sommes sont- ils prêts à exploiter ces textes ?

La constitution du corpus d'articles de presse numérisé porte sur les discours de différents journaux, notamment des quotidiens d'expression française et arabe. Dans un premier temps, notre travail se focalise sur des chroniques et éditoriaux de quotidiens d'expression française.

Notre objectif consiste à faire une numérisation diachronique et au recensement d'archives disponibles afin d'en numériser des parties essentielles et / ou représentatives en mode texte, c'est-à-dire accessible à des recherches linguistiques.
Nous prévoyons d'appliquer à ce corpus en constitution les outils conceptuels et logiciels de l'analyse textuelle des discours, afin de l'analyser sur les plans lexico-thématique, énonciatif et pragmatique.

C'est une étude qui s'inscrit dans le champ des connaissances empiriques des médias et qui vise aussi à intéresser celui des sciences sociales en général, d'un point de vue interdisciplinaire.

2. Méthodologie et constitution d'un corpus de la presse

La constitution de ce corpus commence par la numérisation. Cette dernière se fait souvent en deux grandes étapes : la numérisation en mode image et l'océrisation en mode plein-texte.
D'une part, le mode image permet de préserver l'intégralité des propriétés d'un texte inscrit dans son *document*. D'autre part, le mode plein-texte ouvre la voie à tout un champ d'études sur les textes.
Au niveau méthodologique, la numérisation assure au texte une fidélité à son document d'origine.

Au niveau théorique, la numérisation d'un document constitue l'ensemble des opérations de restitution du texte.

Ainsi, la numérisation est actuellement le mode le plus favorable de conservation et de diffusion des documents pour les nombreux avantages qu'elle apporte aux documents.
Premièrement : elle rend un document plus maniable.
Deuxièmement : grâce à elle, nous évitons une manipulation fréquente des documents fragiles.
Enfin : elle facilite un accès plus rapide au document et permet un accès aux outils informatiques tels que les outils d'analyse statistique textuelle permettant ainsi une exploration fine du texte.

Parmi ces outils, se trouve le logiciel lexico 3 destiné aux explorations textométriques, réalisé par l'équipe universitaire *A* de l'université Paris 3 Sorbonne nouvelle, sous la direction d'André Salem.
Dans le cas de notre travail, le logiciel Lexico 3 sera utilisé pour mettre le corpus en base textuelle exploitable afin de pouvoir procéder à la segmentation, aux concordances, aux décomptes portant sur les formes graphiques, aux spécificités et analyses factorielles portant sur les formes et les segments répétés.

3. Le mode de numérisation

La numérisation en mode texte constitue une opération double,
il s'agit de scanner des documents papier dans un premier temps et de convertir cette première forme de numérisation du mode image au mode plein texte.
Le passage du mode image au mode plein-texte s'effectue obligatoirement par l'océrisation.
Techniquement, il s'agit du traitement d'une image de texte sur laquelle intervient un logiciel de reconnaissance de caractères : le logiciel déchiffre les formes et les traduit en lettres.
Une étape d'apprentissage est parfois nécessaire, c'est-à-dire qu'à chaque caractère non reconnu, il faut lui indiquer quelle est la lettre en question. Mais, malgré cet apprentissage, il existe toujours un taux d'erreur dans la reconnaissance de caractères, lié à la qualité du document initial, aux polices employées, aux notes et à la forme du texte...
Dans le travail du corpus d'articles de presse, l'océrisation a été réalisée à partir des images numérisées à l'aide du logiciel *ABBYY FineReader* qui permet de convertir différents types de documents tels que les documents papiers scannés, les fichiers numériques vers des formats modifiables et exploitables.
Il fonctionne sur le principe de reconnaissance des caractères permettant une transformation d'un fichier image en fichier texte.

Conclusion

Pour conclure, constituer un corpus linguistique à partir de documents numérisés constitue un défi, car chaque document présente des difficultés différentes. Donc chaque corpus a son propre mode de constitution. Ainsi, chaque constitution de ce type de corpus restera un nouveau défi.