PASSERELLE

N4 2013

Concepts de numérisation

Et de lexicométrie

En Réseau avec le projet PNR
Analyse du discours et des Objets Signifiants

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

CRSTDLA

UNIVERSITE D'ORAN

Projet National de Recherche

Analyse du discours et des objets signifiants

RAPPORT D'ACTIVITE

Chef de projet: Pr CHIALI Fatima Zohra

Chercheurs impliqués dans le projet **أعضاء المشروع و المؤسسة المستخدمة**

Nom et prénom الاسم و اللقب	Grade الرتبة	Etablissement employeur المؤسسة المستخدمة	Observation
1. TABET AOUL ZOULIKHA	MCB	USTO	En activité
2. HARIG FATIMA ZOHRA	MA	Univ Oran	En activité
3. BENABDELLAH IMENE	MCB	Univ Oran	En activité
4. BOULKIFANE MALIKA	MA	Univ Oran	En activité

SOMMAIRE

I. Présentation

Lorsqu'on réalise l'importance et la place que l'informatique a prises dans la vie du citoyen depuis bientôt une décennie, on ne peut que mettre en avant et louer l'initiative du projet de la BnTA, Base nationale des Textes Algériens, éditée par le projet PNR *Analyse du Discours et Des Objets Signifiants*. L'on reconnaîtra dès lors les deux objectifs primordiaux de ce projet :

-La maîtrise et la formation des chercheurs dans les domaines et les méthodologies en analyse du/des discours et en lexicométrie.
-Mise en place des fondements d'une base de données numériques BnTA qui référence et étiquette un ensemble vaste des textes algériens.

Ainsi, le nouveau contexte organisationnel au sein duquel l'outil numérique s'impose déplace les notions spatio-temporelles, les relations électroniques engendrent de nouvelles approches des pratiques de lecture, de réception du texte mais aussi d'apprentissage. Signalons que les nouvelles technologies ne se résument pas à l'utilisation d'ordinateurs, de logiciels ou de bases de données mais bien à une transformation des méthodes d'investigation, l'analyse débouchant sur une réflexion *actualisante*, amenant le lecteur à se situer dans le monde d'aujourd'hui.
L'objectif de la BnTA, constitution d'une base de données, est atteint de par le corpus numérisé mais désormais la BnTA accède au statut de plateforme de référence au même titre Gallica, Frantext ou Ibn Rushd. Si le fait d'appréhender, d'observer, de mesurer, de comparer l'apparition de différents types d'événements au fil des textes n'est pas nouveau, c'est la simultanéité de ces opérations sur de larges corpus qui en fait la nouveauté et l'intérêt.

Les étapes dans notre projet se sont déroulées sur deux parcours :

-Une formation théorique et pratique des membres de l'équipe et d'étudiants doctorants, magisterants et masterisants à travers un panorama général des outils lexicométriques exploitables par l'analyse du discours et de la sémiotique appliquée, à savoir :
-L'acquisition et l'application des techniques relatives à la numérisation et à l'océrisation.
-La collecte fastidieuse des corpus et leur traitement numérique. Catégorisation et étiquetage permettant l'étape de l'indexation préliminaire des corpus constitutifs de la plateforme de la BnTA.

-Une application par les traitements quantitatifs et qualitatifs des corpus. Ce qui pousse la recherche numérique à plus de numérisation et viser l'exhaustivité.
En adoptant un format de stockage des documents, nous donnons un aperçu de la méthodologie utilisée par des extractions-partitionnements de parties de corpus. La forme éditoriale d'origine (pagination, saut de ligne,...) et la macro-structure d'origine des textes (chapitre, section, paragraphe, ...) sont maintenues dans l'intention de présenter un maximum d'informations.

La base de données BnTA devient produit semi-fini pouvant déjà être exploité par des chercheurs avec les logiciels de traitement de texte, tels Hyperbase, Tropes. La partie technique relevant de la mise en ligne de la BnTA restera à poursuivre pour lui donner le statut des plateformes internationales telles Frantext.
Nous avons prévu à cet effet un partenariat avec l'équipe du PNR…. formée principalement d'informaticiens capables de matérialiser les données numérisées en

plateforme avec des principes d'accès Web. Ce qui garantira une diffusion plus large de ses enjeux. Car la concrétisation de la BnTA, motivant un nombre plus conséquent de chercheurs, devient une œuvre transmédia à caractère aussi bien national qu'international. Outil de travail adapté à l'ère numérique, la BnTA ambitionne de devenir « la Bibliothèque/Archive numérique digitale » de par l'immense espace de stockage des livres scannés et numérisés. En cela, la feuille de route numérique a démarré avec la BnTA et des terabytes en capital.

La dynamique de ce projet PNR a permis également :
-Sur le plan de la formation et de la formation continue la soutenance de deux thèses de magistère et l'inscription de trois thèses de doctorat, un magistère dans le domaine souscrit ainsi que la validation d'une formation en Master–Recherche à partir de Septembre 2013

-Sur le plan des manifestations scientifiques, nous enregistrons l'organisation de cinq Journées d'études et d'un colloque international ''Contexte et Discours'', en Novembre 2012. Ce qui a permis d'ouvrir le débat sur l'intérêt des nouvelles technologies et comment elles amènent à reconsidérer la création en dehors des frontières d'un territoire.

-Sur le plan de la production scientifique, l'équipe a publié des ouvrages en réseau avec le travail initié. Ce même rapport d'activité fait l'objet de publications.

II. Travaux de recherche réalisés

Cette rubrique regroupe les différentes contributions des chercheurs au cours des journées d'étude et des séminaires régulièrement assurés par le Laboratoire LOAPL et qui permettaient aux chercheurs de confronter leurs résultats et de débattre des nouvelles problématiques. Le but étant d'accéder à une cognition maîtrisée en Analyse de Discours, pratiques lexicométriques et numériques. Ces efforts ont permis d'initier, de former et de perfectionner les membres de l'équipe ainsi que les étudiants qui s'y rattachent.

La notion de corpus, point focal de nos recherches, a été définie à la lumière de l'éclairage de D. Maingueneau comme l'ensemble des productions liées à une société dans un contexte particulier. Des concepts comme la scénographie, la paratopie, l'image de l'auteur et instance auctoriale, etc. ont été convoqués pour aborder l'analyse de discours sous l'angle de la sémiotique différentielle, de la variation et de la mise en scène énonciative. La problématique qui entoure la définition même d'un auteur algérien (axe littéraire) trouve dans ces notions une mesure qui permet d'esquisser des hypothèses et d'en proposer des lectures. L'analyse s'est recentrée sur le statut du producteur, du produit en tant qu'événement vécu, de l'objet concret dans les divers contextes d'émergence (les politiques éditoriales). Ce qui amène l'analyse à identifier les représentations auctoriales investies au cœur même de l'identité de la production algérienne. Le discours constituant devient une notion opératoire permettant de découvrir les modes d'expression de la production algérienne et/ou liée à l'Algérie.

A long terme, La BnTA deviendra une référence.

Les articles qui se sont penchés sur l'utilisation des logiciels lexicométriques ont montré la priorité fondamentale de continuer le travail entrepris sous forme des diverses catégorisations et combinatoires. Ce qui va donner une accessibilité et une lisibilité plus performantes en termes de pertinence et de puissance pour la base de données. Il ne s'agit pas uniquement de corpus sous-spécialisés comme la banque de données de textes littéraires et journalistiques, mais aussi de former des autoroutes de données pouvant être croisées pour des fouilles textuelles par exemple des propositions d'archivage et de conservation des supports électroniques pour bibliothèques.

La lemmatisation et l'AFC ont été initiées par les chercheurs en micro-analyses afin de simuler des situations qui se prêteraient aux jeux des interprétations lexicales, discursives et sémiotiques.

L'hétérogénéité constitutive du discours rend plus problématique la pratique de la lemmatisation, son aspect peut réduire, « appauvrir » le sens d'un texte. Cependant l'apport de la lemmatisation devient fécond et bénéfique s'agissant du discours médiatique et des idéologies sous-jacentes. Elle s'avère aussi intéressante dans la mise en évidence d'un génome stylistique. L'utilisation du logiciel Hyperbase permet de réaliser des bases hypertextuelles avec les textes qu'on lui fournit (en mode ASCII, ou en convertissant des textes présentés en XML ou HTML). Le programme d'exploitation répond aux besoins classiques du traitement automatique des textes : index sélectifs ou systématiques, dictionnaires des fréquences, concordances, sélection de contextes élargis, cooccurrences, recherche des parties ou groupes de mots.

Quant aux travaux du Colloque Contexte et Discours, ils sont publiés dans la revue internationale Passerelle 5 et 6.