MEDITE : logiciel d'alignement de textes pour l'étude de la génétique textuelle

Zied Sellami[1] Jean-Gabriel Ganascia[1] Mohamed-Amine Boukhaled[1]
(1) Laboratoire d'Informatique de Paris 6 (LIP6) 4 Place Jussieu, 75005 Paris
{zied.sellami, jean-gabriel.ganascia, mohamed.boukhaled}@lip6.fr

Résumé. MEDITE est un logiciel d'alignement de textes permettant l'identification de transformations entre une version et une autre d'un même texte. Dans ce papier nous présentons les aspects théoriques et techniques de MEDITE.

Abstract.

MEDITE: text alignment software for the study of textual genetics
MEDITE is an alignment software able to identifying transformations between two versions of a same text. In this paper we show the theoretical and technical aspects of this tool.

Mots-clés : Alignement de textes, Génétique textuelle, Détection d'homologies dans les séquences textuelles
Keywords: Text alignment, Textual genetics, Homology detection in text sequences

1 MEDITE : aspects théoriques

MEDITE est un logiciel d'alignement de textes issu d'une collaboration entre l'ITEM (Institut des Textes et Manuscrits Modernes) et l'équipe ACASA du LIP6.

Initialement MEDITE était prévu pour aligner des transcriptions linéarisées d'avant-textes afin de mettre en évidence les différences et les invariances. Il s'est révélé utile dans de nombreuses autres applications, comme pour établir l'appareil critique d'éditions savantes en comparant les différentes versions publiées d'une œuvre, pour l'étude des variations de textes collectifs, ou encore pour la comparaison de bi-textes afin d'améliorer les outils de traduction statistique (Ganascia, Bourdaillet, 2006). MEDITE est construit sur un algorithme d'alignement de textes par fragments qui recourt à une détection des homologies par la méthode des arbres de suffixes. Il met en évidence les suppressions, les insertions, les remplacements et les déplacements. La première étape de l'algorithme identifie les blocs homologues maximaux. Il s'agit ensuite de distinguer, parmi ces blocs, des pivots et des blocs dits déplacés. Le processus est itéré de façon récursive afin d'éviter les phénomènes de masquage. Enfin, les insertions, les suppressions et les remplacements se déduisent de l'alignement des blocs non répétés (Fenoglio, Ganascia, 2008). L'algorithme de MEDITE étant fondé sur des principes d'algorithmique des séquences, il est indépendant de la langue et peut donc traiter n'importe quel texte, sans ressources spécifiques. En outre, il peut repérer des réutilisations de parties de mots, ce qui s'avère très utile, en particulier pour les langues flexionnelles (Bourdaillet, Ganascia, 2007).

2 MEDITE : aspects pratiques

Une version en ligne de MEDITE est disponible dans http://obvil.paris-sorbonne.fr/developpements/medite. La page d'accueil de MEDITE permet à l'utilisateur d'introduire deux textes à comparer et de paramétrer l'outil (voir Figure 1). Les paramètres de MEDITE sont aux nombres de 8 :

- **Sensible à la casse** : l'alignement est sensible ou pas aux caractères majuscules/minuscules ;
- **Sensible aux signes diacritiques** : l'alignement est sensible ou pas aux caractères accentués ;
- **Sensible aux séparateurs** : l'alignement est sensible ou pas à la ponctuation ;
- **Algorithme mots (coché) ou caractères (non coché)** : l'algorithme de comparaison effectue un découpage par mots ou par caractères des segments à comparer ;
- **Colorer uniquement les blocs en commun** : il s'agit d'une option d'affichage des résultats. En cochant cette option, uniquement les blocs commun entre les deux textes seront colorés ;

– **Longueur minimale des chaînes pivots** : paramètre lié aux arbres de suffixes et fixé à une valeur de 5 ;
– **Ratio minimal des chaînes remplacés** : Il s'agit d'indiquer en pourcentage le taux de textes différents entre deux blocs comparés pour considérer cela en tant qu'un remplacement ;
– **Seuil de longueur pour validation lissage** : paramètre lié aux arbres de suffixes et fixé à une valeur de 50 %.

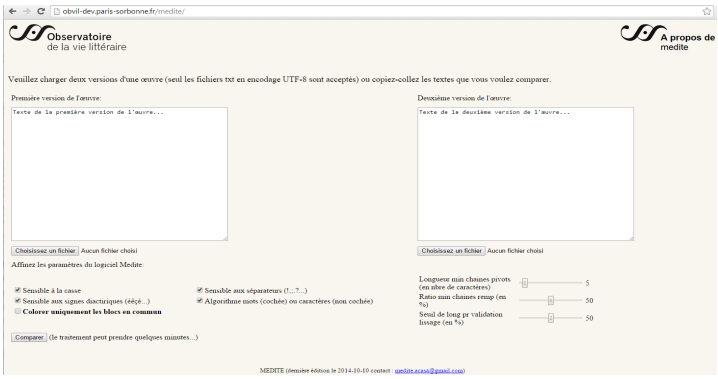

FIGURE 1: Capture d'écran de la page d'accueil de l'outil MEDITE

Les résultats sont affichés après que l'utilisateur clique sur le bouton comparer. Les résultats des alignements sont affichés dans une page de résultats (voir Figure 2). Les deux textes sont présentés côte à côte sur une interface qui met en évidence, au moyen de différentes couleurs, les blocs insérés, supprimés, remplacés et déplacés. Les blocs en commun entre les deux textes sont reliés entre eux par un simple clic de souris. Les résultats peuvent être sauvegardés en cliquant sur l'icône disquette de la page des résultats.

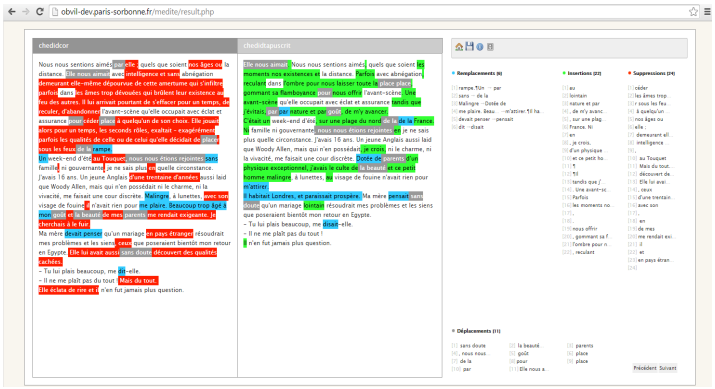

FIGURE 2: Capture d'écran de la page des résultats de MEDITE

Remerciements

Ce travail a bénéficié d'une aide d'État gérée par l'Agence Nationale de la Recherche dans le cadre des Investissements d'Avenir portant la référence ANR-11-IDEX-0004-02

Références

FENOGLIO I., GANASCIA J.-G. (2008). "Le logiciel MEDITE: approche comparative de documents de genèse", in *L'édition du manuscrit - De l'archive de création au scriptorium électronique, Aurèle Crasson, Academia A|B Bruylant, col. Au coeur des textes*, n°10, 209-228.

BOURDAILLET J., GANASCIA J.-G. (2007) Alignement of Noisy Unstructured Text Data, *IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*, Hyderabad, India.

GANASCIA, J.-G., BOURDAILLET, J. (2006). Alignements unilingues avec MEDITE. *Actes des Huitièmes Journées Internationales d'Analyse Statistique des Données Textuelles*.