Phœbus : un Logiciel d'Extraction de Réutilisations dans des Textes Littéraires

Mohamed-Amine Boukhaled, Zied Sellami, Jean-Gabriel Ganascia

LIP6 (Laboratoire d'Informatique de Paris 6), Université Pierre et Marie Curie and CNRS (UMR7606),
ACASA Team, 4, place Jussieu,
75252-PARIS Cedex 05 (France),
{mohamed.boukhaled, zied.sellami, jean-gabriel.ganascia}@lip6.fr

Résumé. Phœbus est un logiciel d'extraction de réutilisations dans des textes littéraires. Il a été développé comme un outil d'analyse littéraire assistée par ordinateur. Dans ce contexte, ce logiciel détecte automatiquement et explore des réseaux de réutilisation textuelle dans la littérature classique.

Abstract.

Phoebus: a Reuse Extraction Software for Literary Text
Phoebus is a reuse extraction software for literary text. It was developed as a computer-assisted literary analysis tool. In this context, the software automatically detects and explores textual reuse networks in classical literature.

Mots-clés : Extraction de réutilisations, empreintes digitales textuelles, analyse littéraire assistée par ordinateur
Keywords: Reuse Extraction, textual fingerprintng, computer-assisted literary analysis.

1 Introduction et Motivation

Les études littéraires et le néo-structuralisme des années soixante et soixante-dix mirent en évidence l'importance des influences mutuelles dans la production littéraire et intellectuelle. Celles-ci se traduisent par des réemplois plus ou moins littéraux et par l'usage d'un vocabulaire similaire. Des notions comme l'intertextualité (Kristeva, 1974) et l'hyper-textualité (Genette, 1982), ont été introduites dès les années soixante-dix pour approcher et formaliser ces phénomènes à partir de l'étude des paraphrases, les réécritures et les citations. L'extraction des réutilisations dans les textes littéraires permet aux chercheurs en littérature de tracer les réemplois d'écrit antérieurs, à regarder l'origine des citations partagées et la façon dont elles sont introduites, d'énumérer les utilisations de proverbes dans les romans et d'étudier d'une manière plus contrôlée la notion d'influence littéraire en termes d'idées et de styles.
L'augmentation de la puissance de calcul des machines et la numérisation des corpus de très grande taille, y compris les corpus de textes littéraires, ont participé au développement des possibilités d'extraction automatique de la réutilisation dans de tel corpus. Ainsi plusieurs méthodes basées sur différentes approches ont été proposé pour effectuer de telles tâches. L'une des approche les plus réussites est l'approche basée sur les empreintes digitales textuelles (Broder, 1997). Cette approche consiste tout d'abord en l'indexation des textes avec des séquences récurrentes de mots. Puis comparer les séquences appartenant à deux textes différents pour extraire les segments textuels communs.

Dans ce contexte, Phœbus a été conçu comme un outil d'analyse littéraire assistée par ordinateur permettant l'extraction de réutilisations dans des textes littéraires. Ce logiciel détecte automatiquement et explore des réseaux de réutilisation textuelle dans la littérature classique.

2 L'approche utilisée

L'approche d'extraction des réutilisations dans Phœbus est basée sur une mise en place et une adaptation de la méthode des empreintes digitales textuelles (Ganascia et al., 2014). Plus précisément, elle se compose de quatre étapes principales :

1. Préparation du texte en utilisant des techniques de traitement automatique du langage naturel ;

2. Extraction de séquences récurrentes élémentaires de mots ;
3. Regroupement des séquences récurrentes élémentaires qui se chevauchent en séquences récurrentes de plus grande taille ;
4. Filtrage des séquences résultantes et élimination des redondances.

3 Fonctionnement de Phœbus

Phœbus a été développé comme une application web avec une architecture client/serveur. L'utilisateur peut y accéder en utilisant un navigateur web via l'adresse suivante : `http://obvil-dev.paris-sorbonne.fr/phoebus` L'interface utilisateur de Phœbus est assez intuitive, elle se compose de deux champs textuels et de trois contrôleurs de paramètres : Les deux champs textuels servent à copier ou à charger les textes à comparer (textes dont on voudrait extraire des parties similaires présentant une réutilisation). Les contrôleurs permettent de choisir les trois paramètres du programme qui définissent la granularité et la finesse d'analyses à travers les trois propriétés suivantes :

- Le nombre de mots réutilisés à considérer dans les empreintes digitales textuelles, ce qui correspond au nombre de mots en commun entre une réutilisation et son texte original.
- Le nombre de trous autorisés dans ces empreintes pour leurs donner une plus grande capacité de généralisation à des réutilisations non littérales (réutilisations avec changement de quelques mots)
- Le fait de respecter ou pas l'ordre des mots réutilisés entre la réutilisation et le texte original.

Une fois les deux textes choisis et les paramètres définis, l'utilisateur lance l'extraction et les résultats seront affichés et synthétisés (voir Figure.1) d'une façon très ergonomiques en incluant entre outre :

- Un alignement automatiques des réutilisations entre texte source et texte cible
- La possibilité de navigation dans les différentes réutilisations.
- Une valeur d'importance donnée à chaque réutilisation par critère de taille (Plus la couleur verte est foncé, plus la réutilisation est importante).
- Le couplage avec le logiciel MEDITE qui compare en finesse les transformations textuelles élémentaires (suppressions, insertions, remplacements et déplacements) qui font passer d'un bloc à son semblable.

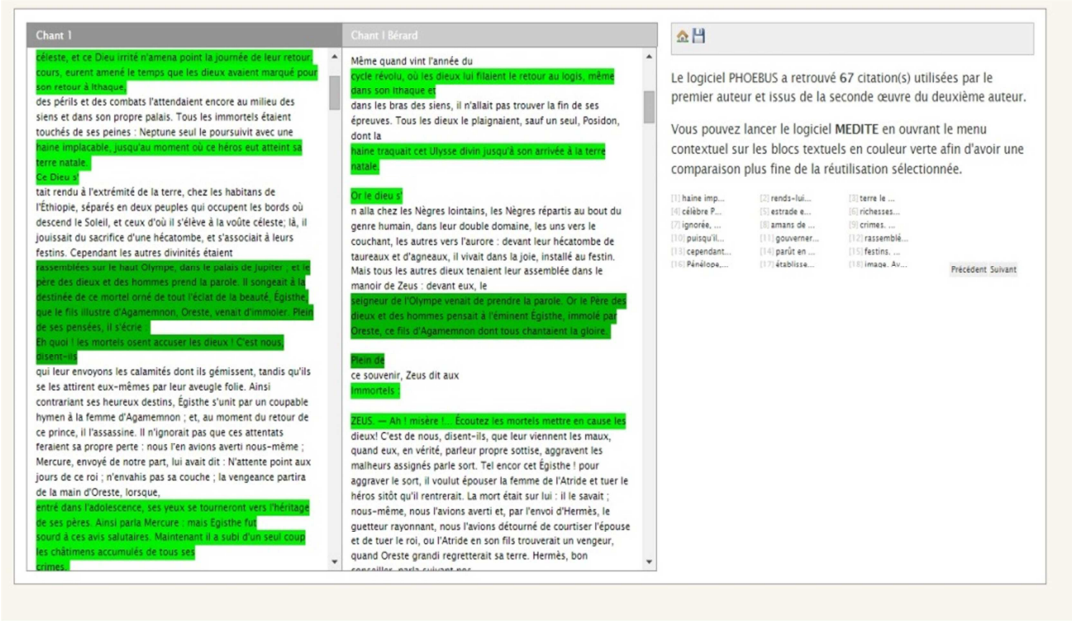

Figure 1. Exemple d'une page de résultats produite par Phœbus

Remerciement

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

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings* (pp. 21–29)

Ganascia, J.-G., Glaudes, P., & Del Lungo, A. (2014). Automatic detection of reuses and citations in literary texts. Literary and Linguistic Computing, 29(3), 412–421.

Genette, G. (1982), Palimpsestes : La Littérature au second degré, Seuil, coll. « Essais », Paris.

Kristeva, J. (1974). La Révolution du langage poétique, col. "Tel Quel", Editions du Seuil.