

Detecting semantic or structural similarities for theater play comparison

Pierre Bourhis¹, Aaron Boussidan², Céline Fournial³, Philippe Gambette⁴

¹Université de Lille, CNRS, Inria, Centrale Lille – pierre.bourhis@inria.fr

²Université Gustave Eiffel – aaron.boussidan@univ-eiffel.fr

³Université Clermont Auvergne – celine.fournial@uca.fr

⁴Université Gustave Eiffel – philippe.gambette@univ-eiffel.fr

Abstract

From the very beginnings of modern theater in France, dramatic creation has been using several practices of imitation and adaptation of earlier texts, from biblical, antique or historical sources. Here, we focus on plays inspired by other plays, and on ways of automatically detecting similarities between the texts of such plays. We present approaches based on structural similarities and others based on semantic similarities between play texts, using word embedding methods. Finally, we look at the advantages of developing hybrid approaches that combine these two criteria, using on the one hand the corpus of sources on dramatic creation provided on the *Hyperpièces* website for comedies, tragi-comedies and tragedies published in France from 1550 to 1650, and on the other hand the *Dracor* corpus.

Keywords: theater plays, text comparison, textual alignment, word embeddings, parameterized words.

Résumé

Dès les débuts du théâtre moderne en France, la création dramatique s'est nourrie de plusieurs pratiques d'imitation ou d'adaptation de textes antérieurs, de sources bibliques, antiques ou historiques. Nous nous intéressons ici aux pièces inspirées par d'autres pièces de théâtre, et aux moyens de détecter automatiquement les similarités entre les textes des pièces dans ce cas. Nous présenterons des approches fondées sur des similarités de structure et d'autres fondées sur des similarités sémantiques entre les textes des pièces, en utilisant des méthodes de plongements de mots. Enfin, nous étudierons l'intérêt de développer des approches hybrides qui mêlent ces deux critères, en mobilisant d'une part le corpus des sources de la création dramatique fournies sur le site *Hyperpièces* à propos des comédies, tragi-comédies et tragédies publiées en France de 1550 à 1650, et d'autre part le corpus *Dracor*.

Mots clés : théâtre, comparaison de textes, alignement de textes, plongements de mots, mots paramétrés.