Analyse textuelle de manuscrits mayas et égyptiens : apports d'un codage par n-grammes, et de représentations multidimensionnelles graduées

Bruno Delprat[1], Martine Cadot[2], Alain Lelu[3]

[1]Université Grenoble Alpes (UFR de Langues étrangères) – brunodelprat@club-internet.fr

[2]Laboratoire LORIA, Nancy – martine.cadot@loria.fr

[3]Retraité, Université de Franche-Comté – alelu@orange.fr

Abstract

For ancient logosyllabic scripts, without separators between lexical units, we propose to explore methods without prior tokenization, adapted to small corpora. We present here a comparative analysis of literary and religious texts, Egyptian tale of the *Shipwrecked Sailor*, and the only three available Mayan manuscripts, using their representation in n-grams of elementary signs, visualized with mayaTeX, and their processing by Correspondence Analysis and graded unsupervised classification (Axial K-Means and Non-negative Matrix Factorization). We identify intra- and inter-text features of the narrative structures in these literary corpora, such as parallelism and *mise en abyme*. The groupings identified on nuanced axes and their correspondences within original texts make it possible to clarify the meaning of certain poorly understood passages, by situating them in contexts easier to interpret.

Keywords: logosyllabic scripts, Maya, Egyptian, n-grams, correspondence analysis, CA, axial k-means, non-negative matrix factorization, NMF, intrinsic dimension, Monte-Carlo simulations, Tournebool algorithm.

Résumé

Pour des écritures logosyllabiques anciennes, ignorant les séparateurs entre unités lexicales, on se propose d'explorer des méthodes sans tokenisation préalable, adaptées à de petits corpus. Nous présentons ici une analyse comparative de textes littéraires et religieux, d'une part égyptien du *Conte du naufragé*, et d'autre part mayas des trois seuls manuscrits mayas disponibles, utilisant leur représentation en n-grammes de signes élémentaires, visualisés avec mayaTeX, et leur traitement par Analyse Factorielle des Correspondances et classification non supervisée graduée (K-Moyennes Axiales et Non-negative Matrix Factorization). Nous en dégageons des manifestations, intra- et inter-textes, des structures narratives dans ces corpus littéraires, comme le parallélisme et la mise en abîme. Les regroupements dégagés sur des axes nuancés et leur report dans le texte original permettent d'éclairer la signification de certains passages peu compris, en les resituant dans des contextes interprétables.

Mots clés : écritures logosyllabiques, Maya, Egyptien, n-grammes, analyse factorielle des correspondances, AFC, k-moyennes axiales, KMA, dimension intrinsèque, simulations de Monte-Carlo, algorithme TourneBool.