Stage M2R CEDRIC
Requêtes complexes dans un corpus de hiéroglyphes

Serge Rosmorduc (ILJ), Nicolas Travers (Vertigo)

Mots-clés : Hiéroglyphes, Indexation, similarité,

Contexte

La gestion documentaire des corpus de hiéroglyphes est riche et complexe. Ces corpus intègrent aussi bien les encodages des graphies, que le son (phonème) et des métadonnées diverses. Retrouver des informations pertinentes demande une connaissance précise du contenu du corpus et de la manière dont les informations sont codées dans celui-ci.

Ces informations reposent principalement sur la graphie et les phonèmes utilisés sur les hiéroglyphes. L'expression de la langue égyptienne et son encodage dans un corpus se complexifie d'autant plus lorsque l'on intègre les problèmes de flexions (préfixe, suffixe), le contexte de l'époque où a été écrit le texte (4500 ans d'évolution de l'écriture), les habitudes des égyptologues (200 ans d'égyptologie), et le fait que des signes peuvent être corrélés (étiquette, imbrication).

La base de données *Ramsès* [1] est la seule à intégrer les graphies originales des mots, ouvrant ainsi la porte à des études morphologiques et des interrogations complexes. Elle propose à l'aide d'automates [2] de composer des motifs de phrase en combinant différentes graphies/phonème pour retrouver les documents correspondant à la structure définie.

Toutefois, certains documents ne répondent pas toujours à ce type d'interrogation. Principalement dû à des structures un peu trop fixes, à des variations dans l'encodage, ou à des fusions de corpus demandant une interprétation différente.

Sujet de stage :

Ce stage a pour but de proposer une approche complémentaire aux automates pour intégrer à la base de données un moteur de recherche par similarités. Le but serait de proposer un module de recherche textuel (e.g. *Lucène* [3], SolR [4]) associé à Ramsès

1. Un modèle de représentation des hiéroglyphes, intégrant les différentes facettes possibles : les mots, leurs flexions, les imbrications, les étiquettes, les phonèmes et le contexte.
2. Définir un langage de recherche à la fois simple et riche pour les égyptologues. Pour ce faire, nous analyserons les logs de recherche effectuées sur Ramsès

pour en comprendre les structures indispensables.

3. Proposer un modèle de calcul de distance entre les documents en donnant à chaque facette un poids, et une fonction de similarité inspirée des méthodes de recherche d'information [5] (i.e., cosinus).

Pour valider le modèle, un jeu de documents sera sélectionné pour donner une valeur qualitative au modèle de calcul.

Références :

[1] S. Polis, S. Rosmorduc, J. Winand. "Ramses goes online. An annotated corpus of Late Egyptian texts in interaction with the Egyptological community", International Congress of Egyptologists, August 2015, Vol. XI, pp.n/a, Florence, Italie,

[2] S. Rosmorduc, "Automated Transliteration of Egyptian Hieroglyphs" in N. Strudwick ed., *Information Technology and Egyptology in 2008,* p. 167-183.

[3] N. Travers. "Putting into Practice: Full-Text Indexing with LUCENE", Titre du livre: "*Web Data Management*", February 2012, Cambridge University Press, pp. 355--363, (isbn: 9781107012431)

[4] R. Fournier, P. Rigaux, N. Travers. « SolR, un moteur de recherche » http://b3d.bdpedia.fr/solr.html

[5] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008