Maison méditerranéenne des sciences de l'homme

30-31 Janvier 2015 **Journées d'étude**

Programme de Recherche TALA

Traitement par Automates de la Langue Arabe

Invariance et calculabilité en langue arabe et en sémitique

MMSH - 5, rue du château de l'horloge, Aix-en-Provence

Intervenants

Claude Audebert (AMU, MMSH)
Serge Rosmorduc (Paris VIII, CNAM)
François Barthélemy (CNAM)
Joseph Dichy (ICAR, Lyon 2)
Christian Gaubert (IFAO, MMSH)
André Jaccarini (MMSH)
Abbdel Majid Arrif (MMSH)
Samir Zardan (MMSH)

Programme

Journée du 30/01/2015

10h- 11h : **Modélisation d'une hypothèse cognitivo-didactique sur l'optimisation de la recherche dans le dictionnaire arabe**

Claude Audebert, André Jaccarini, Christian Gaubert

Abstract
L'expérience de pensée qui est à l'origine de ce projet consiste à supposer qu'il existe un processeur abstrait et optimal de décodage d'un texte arabe non vocalisé qui minimise le recours au lexique. On suppose résolu le problème de l'analyse morphologique de la détermination de la racine que nécessite l'accès au dictionnaire et l'on imagine à l'aide d'un programme fictif l'optimisation de l'ensemble des processus sous-jacents à cette analyse (méthodologie *top-down*), en se libérant de la contrainte de la

progression mot à mot. L'idée principale est que la stratégie consistant à avancer dans le texte, en recourant le moins possible au dictionnaire, est la plus efficace à condition toutefois de savoir repérer les éléments les plus contraignants (segmentation entropique) et mettre en œuvre en en tirant judicieusement parti les *attentes* morphosyntaxiques qu'ils déclenchent. L'expression de ces contraintes sous forme d'automates s'imbriquant les uns dans les autres découle donc de l'hypothèse, que suggère l'expérience pédagogique, de l'existence de ce processeur idéalisé. Or celui-ci, s'il existe, ne peut être appréhendé que par approximations successives et la définition de méthodes de construction des grammaires par agrégation de fragments en ayant recours à la théorie algébrique des automates. Par ailleurs la mise au point progressive des fragments de grammaire ne peut se faire qu'en procédant par rétroaction continue, d'où la nécessité de la mise au point d'un outil permettant d'assurer le *feedback* entre corpus et grammaires lequel permettra ainsi de boucler rétroactivement jusqu'à obtenir un niveau jugé satisfaisant.

11h - 12h : Claude Audebert - **Quelques remarques sur les tokens arabes en tant qu'opérateurs syntaxiques**

Abstract
Description des attentes syntaxiques déclenchées par certains marqueurs morpho-graphiques facilement repérables.

14h - 15h : Serge Rosmorduc - **Spécificités de la langue égyptienne et de son système de représentation hiéroglyphique**

15h - 16h : François Barthélemy et Serge Rosmorduc -**Modèle de calcul en vue du traitement par automates des langues sémitiques ou apparentées**

Abstract
Translitération des hiéroglyphes égyptiens en alphabet latin pour comparer deux formalismes de machines à états finis: le premier est une cascade de transducteurs binaires; le second repose sur des transducteurs multi-rubans exprimant des contraintes simultanées et produit des machines de taille plus importantes mais permet en revanche des descriptions plus abstraites.

16h- 17h : Joseph Dichy- **Les limites calculables de la calculabilité: la théorie des spécificateurs et son application à l'arabe**

17h-18h **:** Moulay Ismail Elamrani - **Vers la base de données lexico-sémantique trilingue LASMAR (Lexique analytique et sémantique multilingue de l'arabe)**

Matinée du 31 janvier

- 9h – 13h : **Séances de travail**

  Discussion générale:
- De l'intérêt de représenter les automates morphosyntaxiques arabes issus de l'application Kawakib-pro dans le système KNG (automates multi-rubans pondérés) en vue d'études comparatives en vue d'établir des critères de comparaison des méthodes
  tant au niveau de la modélisation linguistique qu'au niveau des modèles de calcul afin de fixer des méthodes d'évaluation et de comparaison sur des cas précis d' analyses mises en œuvre dans les deux systèmes Kawakib et KNG (Karamel New Generation); esquisse d'un plan de travail en commun.

- **Recherche de principes généraux en vue**
  - d'établir une méthode pour déterminer les compromis optimaux entre grammaire et lexique;
  - De rechercher des critères précis de pondération des règles;

- **ATALA** : Organisation de la journée internationale que nous devons organiser à l'Association pour le Traitement Automatique des Langues (ATALA) dans le courant de l'année 2015.