

Extraction et chaînage supervisés de connaissances d'un corpus d'entretiens en histoire des sciences

Benjamin Hervy**, Matthieu Quantin*,***, Pierre Teissier***

*École Centrale de Nantes, IRCCyN UMR CNRS 6597
prenom.nom@irccyn.ec-nantes.fr

**Université d'Angers, CERHIO UMR CNRS 6258
prenom.nom@univ-angers.fr

***Université de Nantes, Centre François Viète EA 1161
prenom.nom@univ-nantes.fr

1 Introduction

Les données des sciences de l'homme forment souvent des corpus de textes, qui sont hétérogènes par leurs forme et contenus ; spécifiques par leurs terminologie et signification. À partir d'un corpus d'entretiens en histoire des sciences, nous présentons une méthode supervisée générant un réseau de documents liés par leurs proximités de contenus. Il s'agit d'un graphe multiple flou, basé sur l'extraction de *n-grams* à taille variable.

1.1 Présentation du corpus

Le corpus est formé par la retranscription de 37 entretiens de chercheurs racontant leur carrière parfois depuis les années 1940. Le corpus global contient 293k mots et a été étudié manuellement (thèse, articles, livres) par Teissier (2007). Il entrecroise ainsi des questions techniques (recherche), des énoncés relationnels et affectifs (interpersonnels) et des positionnements identitaires (discipline, génération, genre, etc.).

1.2 Objectifs et hypothèses

L'approche proposée vise à mettre en évidence de nouveaux éléments de réflexion pour l'historien via la création d'un graphe de co-occurrences d'expressions entre documents.

Nous faisons l'hypothèse qu'il est préférable que l'automatisation génère des inférences bas-niveau s'articulant avec des inférences (manuelles) qualitatives haut-niveau de l'historien. L'objectif de la méthode proposée est de favoriser cette articulation pour saisir les structures et les dynamiques de communautés scientifiques.

L'hyper-spécialisation du corpus étudié participe de la richesse de l'analyse historique. La méthode numérique devra préserver cette richesse terminologique. Les connaissances extraites sont issues du corpus ou des choix de l'historien et non d'un modèle défini *a priori*. Dans cette optique, nous favorisons le *rappel* à la *précision* en produisant des indicateurs pour la supervision de l'historien.