Modèle pour une analyse du phénomène de linéarité de catégories sémantiques dans les énoncés en français

Bernard Decobert
La Cavaille 24380 Veyrines de Vergt
Email : bernard.decobert@wanadoo.fr

1 Introduction

Des alignements répétitifs d'une vingtaine de catégories sémantiques (CS) indépendants de toutes limites de phrases ou de paragraphes, apparaissent dans des articles de presse en français (type dépêche - 300 à 1200 mots). L'analyse du phénomène, qui s'appuie notamment sur des étapes statistiques, conclut à l'existence probable dans le langage d'une structure linéaire sous-jacente non aléatoire plus précisément, d'une structure conceptuelle hiérarchisée se présentant sous la forme d'une séquence récurrente type. Les catégories sémantiques sont repérées grâce à des révélateurs lexicaux (Fig.1). L'étude prend le texte pour objet et s'inscrit dans le cadre général de l'élaboration d'une méthodologie et d'une instrumentation pour l'analyse des phénomènes interprétatifs (sémantique, analyse du discours, syntaxe).

La notion de catégorie sémantique rejoint pour partie celle des concepts primitifs (semantics primes) d'Anna Wierzbicka en Natural Semantic Metalanguage (NSM). Les CS tout comme ces derniers sont réductibles à des mots clés. Toutefois, malgré des ressemblances évidentes, les CS se différencient des concepts primitifs en particulier parce qu'elles tendent à s'articuler entre elles comme des étapes d'un seul et même processus.

L'étude s'est attachée à repérer et modéliser ces catégories sémantiques qui relèvent *a priori* d'un système organisé. Elle cherche à cerner la nature, la portée et les mécanismes inductifs et/ou logiques susceptibles d'incrémenter ces catégories sémantiques. Elle pose la question de la quantification du caractère contraint d'une telle structure conceptuelle hiérarchisée par rapport au hasard, aborde les prémices d'une grammaire permettant de décrire formellement ces contraintes, s'interroge sur le bien-fondé et la précision définitionnelle du catalogue de CS retenu et engage une réflexion sur un traitement algorithmique des données observées. En dehors des analyses statistiques, quatre éléments concourent à interpréter la présence d'un alignement répétitif de catégories sémantiques sous la forme d'une séquence type comme excluant le choix conscient des auteurs :

1. La répétition et l'étendue de la séquence type c'est-à-dire le nombre de catégories sémantiques qui la compose (plus d'une vingtaine sont identifiées à ce jour),
2. Le caractère transphrastique des séquences repérées dans les articles de presse,
3. la régularité des intervalles entre les occurrences représentatives de CS,
4. l'absence de causalité apparente entre ces occurrences.

L'approche méthodologique permet aujourd'hui de classer environ 12 000 unités lexicales (flexions comprises) dans une vingtaine de catégories sémantiques. Bien que cette étude se situe dans la lignée des travaux de recherche sur la construction automatique et approximative du sens d'un texte par les techniques de « clustering » révélant des thèmes sémantiques, elle s'en différencie principalement par l'originalité de la base lexicale.