ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

Recherche doctorale débutée
en septembre 2019 (Nikolay Chepurnykh)

Nom du projet en français	*Traitement de l'accès à l'information lexicale par une approche duelle fondée sur le parcours des grands réseaux lexicaux et l'extraction d'information en corpus*
Nom du projet en anglais	*Processing the Access to Lexical Information Through a Dual Approach Based on the Exploration of Large Lexical Networks and Information Extraction from Corpora*
Directeur	Alain POLGUÈRE – Université de Lorraine, CNRS, ATILF (Nancy) `https://perso.atilf.fr/apolguere/`
Codirecteur	Leo WANNER – ICREA & Université Pompeu Fabra (Barcelone) `https://www.icrea.cat/Web/ScientificStaff/leo-wanner-324`
École doctorale	SLTC (Sociétés, Langages, Temps, Connaissances)
Mots-clés	lexicologie ; grands réseaux lexicaux; extraction automatique d'information linguistique ; traitement de corpus ; accès lexical ; phraséologie

Domaines scientifiques de la thèse

- Lexicologie formelle et informatisée
- Lexicographie des grands réseaux lexicaux
- Traitement Automatique des Langues (TAL)
- Extraction d'information en corpus

Contexte et visées de la recherche

L'*accès à l'information lexicale* pour la lexicalisation et l'organisation syntaxique de la phrase est une opération récurrente effectuée de façon quasi instantanée par le Locuteur dans le contexte de la production de tout énoncé. Il s'agit d'un processus invisible quand il s'effectue de façon normale. Il est toutefois facilement révélé par les accidents d'énonciation et les différentes formes d'erreurs lexicales, qui peuvent avoir de multiples causes : fatigue, émotion, maîtrise non native de la langue, trouble du langage d'origine physiologique ou psychique, etc. Il serait particulièrement intéressant de parvenir à développer des outils qui permettent d'aider le Locuteur à effectuer et exploiter l'accès à l'information lexicale. De plus, ces outils pourraient aussi être utilisés dans des contextes aussi variés que l'étude linguistique, le travail lexicographique et terminographique, l'enseignement de la langue.

> La thèse explore une nouvelle approche de la construction et de l'organisation de l'information lexicale qui pourrait être exploitée par de tels outils.

La structure et le contenu des ressources linguistiques nécessaires à la modélisation informatique de l'accès à l'information lexicale doivent permettre de reproduire automatiquement les processus linguistiques impliqués dans la lexicalisation ou, à l'inverse, l'interprétation des éléments lexicaux dont sont constitués les énoncés (Miller, 1986, 1999; Zock et Schwab, 2011). Deux types de ressources présentent tout particulièrement de l'intérêt pour le développement de ce type d'outils : les *grands réseaux lexicaux* et les *corpus textuels*.

La présente recherche repose sur l'hypothèse que l'on peut mettre en place des stratégies qui utilisent conjointement ces deux types de ressources pour modéliser l'accès à l'information lexicale et guider le Locuteur vers : **(i) l'identification d'un contenu à exprimer et (ii) la recherche du « mot juste » pour exprimer ce contenu**. La thèse explore cette hypothèse à travers l'implémentation d'un système expérimental d'accès automatique à l'information lexicale à partir de :

1. déclencheurs conceptuels → trouver les mots qui lexicalisent une idée à exprimer ;
2. déclencheurs linguistiques → trouver les mots et les structures syntaxiques dont le choix est contingent à des choix linguistiques déjà opérés (*cf.* phraséologie).

Méthodologie et résultats attendus

L'exploitation conjointe de ressources lexicales et du traitement de corpus n'est pas nécessairement productive et certaines tâches peuvent être effectuées de façon équivalente ou, même, avec de meilleures performances par le seul recours aux données de corpus (Wanner et al., 2017). L'hypothèse d'une approche duelle sur laquelle se fonde la recherche semble cependant prometteuse pour au moins deux raisons :

1. L'opération qu'il s'agit d'implémenter – l'accès à l'information lexicale – est particulièrement compatible avec l'exploitation de modèles lexicaux en graphe de type *petits mondes* (Watts et Strogatz, 1998), où le lexique est structuré comme un « réseau social » de mots (Gaume et al., 2008; Polguère, 2016).
2. Les modèles lexicaux qui sont exploités sont éminemment relationnels – ce ne sont pas des modèles ontologiques fondés sur une classification des mots (Polguère, 2014) – et l'on peut s'attendre à ce que les données qu'ils contiennent permettent des interactions productives avec les données extraites des corpus sous forme notamment de représentations vectorielles de type *word embedding*.

Le projet de thèse élaborera des algorithmes d'analyse des grands réseaux lexicaux développés à l'ATILF [1] et exploitera les méthodes statistiques existantes d'extraction des données lexicales de corpus, dont le codirecteur de thèse (Leo Wanner) est un spécialiste. À travers la construction d'un système prototype, la recherche permettra de mieux évaluer la part respective que peuvent assumer les modèles lexicologiques et les ressources textuelles dans la modélisation des processus langagiers.

Les résultats de la recherche peuvent avoir de multiples débouchés ; par exemple :

- la description des erreurs typiques de différents groupes d'utilisateurs et l'identification des causes de ces erreurs en regard de l'organisation de l'information lexicale ;
- la réalisation d'aides à l'expression linguistique – que ce soit dans un contexte d'apprentissage de la langue ou de troubles langagiers ;
- l'enrichissement semi-automatique des ressources lexicales (Sajous et al., 2011).

Bibliographie

Gaume, Bruno, Duvignau, Karine, Prévot, Laurent et Desalle, Yann. Toward a cognitive organization for electronic dictionaries, the case for semantic proxemy. Dans *Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, pages 86–93, Manchester, 2008.

Miller, George A. Dictionaries in the mind. *Language and Cognitive Processes*, 1(3):171–185, 1986.

Miller, George A. On Knowing a Word. *Annual Review of Psychology*, 50:1–19, 1999.

Polguère, Alain. From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 27(4):396–418, 2014. Special Issue: Dictionaries and the Digital Revolution: A Focus on Users and Lexical Databases.

Polguère, Alain. La question de la géométrie du lexique. *SHS Web of Conferences*, 27:01002, 2016. Actes du *5e Congrès Mondial de Linguistique Française (CMLF 2016)*.

Sajous, Franck, Navarro, Emmanuel et Gaume, Bruno. Enrichissement de lexiques sémantiques approvisionnés par les foules : le système WISIGOTH appliqué à Wiktionary. *Traitement Automatique des Langues (T.A.L.)*, 52(1):11–35, 2011.

Wanner, Leo, Ferraro, Gabriela et Moreno, Pol. Towards Distributional Semantics-Based Classification of Collocations for Collocation Dictionaries. *International Journal of Lexicography*, 30(2):167–186, 2017.

Watts, Duncan J. et Strogatz, Steven H. Collective Dynamics of 'Small-World' Networks. *Nature*, 393:440–442, 1998.

Zock, Michael et Schwab, Didier. Storage does not Guarantee Access: The Problem of Organizing and Accessing Words in a Speaker's Lexicon. *Journal of Cognitive Science*, 12:233–259, 2011.

1. *Réseau Lexical du Français* (RL-fr), *de l'Anglais* (RL-en) et, plus marginalement, *du Russe* (RL-ru).