

# Développement automatique de lexiques pour les concepts émergents : une exploration méthodologique

Revekka Kyriakoglou<sup>1</sup>, Anna Pappa<sup>1</sup>, Jilin He<sup>1</sup>, Antoine Schoen<sup>2</sup>, Patricia Laurens<sup>2</sup>, Markarit Vartampetian<sup>3</sup>, Philippe Laredo<sup>2</sup>, Tita Kyriacopoulou<sup>3</sup>

<sup>1</sup>LIASD - Université Paris 8 – [kyriakoglou@up8.edu](mailto:kyriakoglou@up8.edu), [ap@up8.edu](mailto:ap@up8.edu),  
[patrick.he1024@gmail.com](mailto:patrick.he1024@gmail.com)

<sup>2</sup>LISIS - Université Gustave Eiffel – [antoine.schoen@esiee.fr](mailto:antoine.schoen@esiee.fr), [patricia.laurens@esiee.fr](mailto:patricia.laurens@esiee.fr),  
[philippe.laredo@univ-eiffel.fr](mailto:philippe.laredo@univ-eiffel.fr)

<sup>3</sup>LIGM - Université Gustave Eiffel – [markaritvar@gmail.com](mailto:markaritvar@gmail.com), [tita@univ-mlv.fr](mailto:tita@univ-mlv.fr)

## Abstract

This paper presents the development of a lexicon centered on emerging concepts, focusing on *non-technological innovation*<sup>1</sup>. It introduces a four-step methodology that combines human expertise, statistical analysis, and machine learning techniques to establish a model that can be generalized across multiple domains. This process includes the creation of a thematic corpus, the development of a *Gold Standard Lexicon*, annotation and preparation of a training corpus, and finally, the implementation of learning models to identify new terms. The results demonstrate the robustness and relevance of our approach, highlighting its adaptability to various contexts and its contribution to lexical research. The developed methodology promises applicability in conceptual fields.

**Keywords:** semantic lexicon development, conceptual domain modeling, emerging concepts identification, knowledge extraction, machine learning annotation, corpus creation.

## Résumé

Cet article présente le développement d'un lexique centré sur les concepts émergents, en se focalisant sur *l'innovation non technologique*. Il décrit une méthodologie en quatre étapes, combinant expertise humaine, statistiques et techniques d'apprentissage automatique, pour établir un modèle généralisable à plusieurs domaines. Cette procédure comprend la création d'un corpus thématique, la constitution d'un lexique de référence, l'annotation et la préparation d'un corpus d'entraînement, et enfin, l'implémentation de modèles d'apprentissage pour identifier de nouveaux termes. Les résultats montrent la robustesse et la pertinence de notre approche, mettant en évidence sa capacité à être adaptée à plusieurs contextes et sa contribution à la recherche lexicale. La méthodologie développée promet une applicabilité dans des domaines conceptuels.

**Mots clés :** développement de lexique sémantique, modélisation de domaine conceptuel, identification de concepts émergents, extraction de connaissances, annotation par apprentissage automatique, création de corpus.

---

<sup>1</sup> *L'innovation non technologique* englobe les changements dans les domaines de l'organisation, du management, des pratiques de travail, de la stratégie d'entreprise, du marketing, ou encore des modèles d'affaires apportant une valeur ajoutée sans nécessairement s'appuyer sur de nouvelles technologies (Tidd et Bessant, 2020).