Extraction automatique et modélisation lexicale des néologismes chinois (2015–2025) : approches hybrides et sémantique vectorielle

CHEN Lian 陈恋
LLL université d'Orléans & CRLAO - CNRS-INALO-EHESS

Cette communication propose une approche hybride pour l'extraction automatique et la modélisation lexicale des néologismes chinois à partir de corpus médiatiques couvrant la période 2015–2025. Elle s'inscrit dans le cadre des projets ENEOLI COST Action (COST Action CA22126, 2024–2027) et NeoLex (Chen, Dao, Nouvel, Delaporte, https://github.com/Computational-Lexico/NeoLex), et vise à articuler méthodes quantitatives et analyse linguistique dans une perspective diachronique.

La problématique centrale concerne la détection des néologismes (Lü, 1984; Wang, 1992 ; Neveu, 2011 ; Sablayrolles, 2019 ; Tao, 2017 ; citecabre :2004 ; Boulanger, 2010 ; Fan, 2015 ; Cartier, 2016 ; Bernal et al., 2020 ; Bernal et al., 2020; Boulanger, 2010) dans le chinois à partir d'un corpus médiatique diachronique couvrant la période 2015–2025, constitué notamment d'articles issus de *People s Daily, The Papers* et de *Xinhua*, où les innovations lexicales peuvent être formelles (créations morphologiques) ou sémantiques (émergence de nouveaux sens). Dans ce cadre, un néologisme est défini comme une unité lexicale absente des ressources de référence avant 2015 et attestée dans les usages contemporains (Sagot et al., 2013a ; Sagot & Nouvel, 2013b ; Chen et al. 2025).

La méthodologie repose sur une chaîne de traitement intégrée. L'extraction initiale des candidats s'appuie sur une approche par séquences de caractères (fenêtre glissante de 1 à 5 caractères), complétée par des mesures statistiques (fréquence, PMI, entropie). Afin de dépasser les limites des approches purement formelles, une modélisation sémantique est ensuite mise en œuvre à partir de représentations vectorielles (embeddings).

Plusieurs niveaux d'analyse sont combinés : (i) une validation sémantique fondée sur la cohérence des contextes, (ii) une modélisation de la structure sémantique (stabilité, polysémie, dispersion) à l'aide de techniques de clustering, et (iii) une analyse diachronique permettant de mesurer la dérive sémantique (semantic drift) et de détecter l'émergence de nouveaux usages. Cette approche permet de caractériser les unités lexicales selon différents profils (stables, polysémiques, en évolution) (Reimers & Gurevych, 2019 ; Mikolov et al., 2013 ; Reimers & Gurevych, 2019 ; Hamilton et al., 2016 ; Kutuzov et al., 2018 ; Tahmasebi et al., 2019).

Les résultats montrent que la combinaison de méthodes constitue une stratégie plus robuste que les approches isolées, notamment pour identifier les néologismes sémantiques et suivre leur évolution dans les discours médiatiques. Dans une perspective linguistique, ce travail permet de mieux caractériser les dynamiques de variation et de stabilisation sémantiques. Il ouvre également des perspectives en lexicographie computationnelle, en particulier pour la structuration de ressources lexicales telles que des dictionnaires de néologismes fondés sur des données, ainsi que leur intégration dans des cadres comme OntoLex ou des plateformes collaboratives telles que ENEOLI Wikibase (https://eneoli.wikibase.cloud/wiki/Main_Page). Par ailleurs, ce travail pourra être étendu à des corpus plus larges et à d'autres langues, afin de tester la reproductibilité de la méthode. Enfin, il s'inscrit dans une dynamique plus large d'activités scientifiques autour de la néologie computationnelle et de l'analyse des corpus.

Références

Bernal, E., Freixa, J., & Torner, S. (2020). Néologicité et dictionnarisabilité : Deux conditions inverses ? *Neologica, 14*, 47–60.

Boulanger, J.-C. (2010). Sur l'existence des concepts de « néologie » et de « néologisme » : propos sur un paradoxe lexical et historique. In T. Cabré, O. Domènech, R. Estopà, J. Freixa, & M. Lorente (Éds.), *Actes del I Congrés internacional de neologia de les llengües romàniques* (pp. 31–73). Université Pompeu Fabra.

Cartier, E. (2016). Neoveille, système de repérage et de suivi des néologismes en sept langues. *Neologica, 10*, 101–131.

Chen, L., Dao, H. L., Nouvel, D., & Delaporte, A. (2025). *Extraction automatique et modélisation lexicale des néologismes en chinois et en vietnamien (2015–2025) : le projet NeoLex*. NLP & TAL – Traitement Automatique de Langues, INALCO, 1–75.

Fan, H. (2015). The interaction of Chinese translations of Western books and early modern Chinese and Japanese neologisms. *Japanese Language Learning Research, 181*(6), 27–33.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489–1501). Association for Computational Linguistics.
Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384–1397). Association for Computational Linguistics.
Lü, S. (1984). 汉语语法论文集 *[Recueil d'articles sur la grammaire du chinois]*. Maison d'Édition du Commerce.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781
Neveu, F. (2011). *Dictionnaire des sciences du langage* (2e éd.). Armand Colin.
Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
Sablayrolles, J.-F. (2019). *Comprendre la néologie : Conceptions, analyses, emplois*. Lambert-Lucas.
Sagot, B., & Nouvel, D. (2013). Edylex : Enrichissement dynamique de ressources lexicales multilingues. In *Les Rencontres du numérique 2013*.
Sagot, B., Nouvel, D., Mouilleron, V., & Baranes, M. (2013). Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel. In *TALN-RÉCITAL 2013*.
Tahmasebi, N., Borin, L., & Jatowt, A. (2019). Survey of computational approaches to lexical semantic change. *arXiv*. https://doi.org/10.48550/arXiv.1811.06278
Tao, Y. (2017). An investigation into Chinese internet neologisms. *Canadian Social Science, 13*(12), 65–70.
Wang, T. (1992). The criteria of neologisms and principles of dictionary compilation. *Language Application*, (4), 14–21.