

Sens phraséologique ou distributionnel : démêler l'écheveau sémantique par l'algorithmie

Prof. Jean-Pierre Colson
Université catholique de Louvain (Louvain-la-Neuve)

Le linguiste anglais John Rupert Firth (1890-1960) est célèbre pour son affirmation « You shall know a word by the company it keeps » (Firth 1957 : 11). De nombreux linguistes adhèrent à ce principe général, peut-être en raison précisément de son ambiguïté, ainsi que le montrent ses traductions très diverses. Il est souvent invoqué dans les recherches sur les collocations et la phraséologie en général, mais aussi en sémantique distributionnelle, dont Firth est même considéré comme l'un des précurseurs, malgré lui sans doute.

Notre hypothèse de départ est que le sens peut difficilement être à la fois phraséologique et distributionnel. Par exemple, les associations phraséologiques (mesurées par des scores statistiques) vont associer *espoir* et *vie* (en raison de l'expression *l'espoir fait vivre*) ou même *fond* et *caisse* (en vertu de l'expression populaire *à fond la caisse*). Il est clair que des associations purement sémantiques (notamment la synonymie : *vieux* et *ancien*) ou socio-culturelles (*art* et *culture*) constituent un réseau bien différent d'associations phraséologiques.

Même si de nombreux points de rencontre peuvent exister entre les associations phraséologiques et sémantiques, notre hypothèse de travail est que leur nature différente brouille les pistes des recherches en TAL (traitement automatique du langage). Afin de tester notre hypothèse, nous partirons tout d'abord des recherches les plus avancées en phraséologie informatique : l'extraction des expressions multi-mots (EMM) à partir d'un texte continu. La constitution d'un standard de référence multilingue par le projet *Parseme* (<https://typo.uni-konstanz.de/parseme/>) a permis ces dernières années de faire progresser nettement la réflexion dans ce domaine. Les résultats récents des tâches d'extraction automatisée laissent toutefois apparaître les limites des modèles les plus performants, issus de l'apprentissage profond (*DL*, *Deep Learning*) en particulier des réseaux neuronaux (*neural networks*). Si cette méthodologie permet effectivement de reproduire par apprentissage profond des décisions humaines lors de la reconnaissance des EMM, le problème de l'adéquation forcée (*overfitting*) se pose. En outre, la reproduction de telles décisions nous renseigne imparfaitement sur le fonctionnement sémantique du système. En particulier, les filtres grammaticaux (étiquetage, analyse ou *parsing*) influencent le fonctionnement des algorithmes et l'interprétation des données. Un modèle plus simple, basé sur la grammaire de construction et les motifs récurrents, sera proposé (Colson 2017, 2018).

A partir d'un modèle théorique purement distributionnel, les résultats fournis par les outils neuronaux les plus récents, tels que *word2vec* (<https://fr.wikipedia.org/wiki/Word2vec>) offrent des résultats mitigés en matière de synonymie et d'analyse de l'idiomaticité. Selon une brillante étude, pionnière en la matière (Cordeiro et al. 2019), les modèles neuronaux permettraient de détecter la *non-compositionnalité* des phrasèmes nominaux, mais nous montrerons par d'autres expériences les limites d'une telle approche. Enfin, parmi les pistes prometteuses, nous mentionnerons une combinaison possible du sens phraséologique et du sens distributionnel dans la pondération d'un modèle alternatif.

Références

Jean-Pierre Colson. 2017. “The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions”. In: Ruslan Mitkov (ed.), *Computational and Corpus-based phraseology, Lecture Notes in Artificial Intelligence 10596*. Cham, Springer International Publishing: 16-28.

Jean-Pierre Colson. 2018. “From Chinese Word Segmentation to Extraction of Constructions: Two Sides of the Same Algorithmic Coin”. In: Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan and Miriam R. L. Petruck (eds.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, Coling 2018, Santa Fe NM, USA, Association for Computational Linguistics: 41-50.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart and Carlos Ramisch. 2019. “Unsupervised Compositionality Prediction of Nominal Compounds”. *Computational Linguistics* 45: 1-57.

John R. Firth. 1957. *Papers in Linguistics 1934–1951*. London, Oxford University Press.