Génération automatique de mots-valises : approche distributionnelle de la classification sémantique de construits

Lucas LAMBREY

Université de Lorraine
lucas.lambrey3@etu.univ-lorraine.fr

La morphologie constructionnelle étudie les procédés morphologiques permettant la création de nouvelles unités lexicales, dont les lexèmes. Pour ce faire, ces procédés font correspondre les trois dimensions qui caractérisent chaque lexème : la dimension phonologique, qui a trait au contenu phonique de l'unité lexicale ; la dimension sémantique, liée à son contenu référentiel ; la dimension syntactique, en lien avec les contraintes qui régissent son comportement au sein de la phrase — notamment, sa partie du discours.

Dans ce mémoire, nous nous intéressons à un procédé morphologique s'apparentant à la composition, puisqu'il fait correspondre trois lexèmes (deux bases et un dérivé), mais qui s'en distingue par son traitement des dimensions phonologiques et sémantiques : il s'agit du **mot-valisage**, parfois appelé *amalgamation lexicale* (Léturgie, 2011).

Le mot-valisage est phonologiquement particulier parce que la forme des construits est déterminée en grande partie par la position de séquences de segments communes aux deux lexèmes-bases, appelées séquences homophones (*poti**ron*** + *mar**ron*** > *potimar**ron*** ; *c**ou**rri**el*** + *p**ou**bel**le*** > *p**ou**rri**el***[1]).

Le mot-valisage est sémantiquement particulier de par l'imprédictibilité du sens des construits : les lexèmes-bases sélectionnés sont sémantiquement variés — toutes les parties du discours sont possibles — ; le mot-valisage est souvent employé de manière ludique, et le construit ainsi créé peut n'être interprété que dans le contexte d'énonciation. Cela pousse à classer ce procédé parmi les procédés marginaux de la grammaire (Fradin et al., 2009).

Le but de ce mémoire est de concevoir un générateur automatique de mots-valises. Ayant déjà mis au point un programme implémentant la Théorie de l'Optimalité pour générer la forme du construit, nous nous intéressons ici à la dimension sémantique de la génération.

Nous passons en revue quatre manières d'aborder le sens d'une unité lexicale, et évaluons leur potentiel prédictif. En premier lieu, nous remarquons que les tests habituels permettant d'établir certaines propriétés — la massivité et la comptabilité ; l'aspect évènementiel (Dugas et al., 2021) ; les rôles sémantiques liés à un évènement (Fradin, 2021) — sont difficilement implémentables, car il s'agit pour la plupart de jugements de grammaticalité. En deuxième lieu, l'approche paradigmatique de la morphologie dérivationnelle (Fradin, 2021) admet une certaine capacité de prédiction, qui, basée sur la régularité des procédés morphologiques, ne sied guère au mot-valisage. Ensuite, l'approche empirique de la sémantique distributionnelle (Huyghe & Wauquier, 2020, 2021) semble prometteuse de par la possibilité de manipuler mathématiquement des représentations vectorielles de la sémantique de tokens, mais nécessite de fait de définir une classification. Enfin, la base de données lexicale WordNet (Miller et al., 1990), qui illustre l'approche ontologique, offre justement un système de classification sémantique grâce à ses 25 ***Unique Beginners*** (UB) — racines de 25 sous-hiérarchies de WordNet (Miller, 1990). Nous préférons employer la classification proposée par (Barque et al., 2020) pour le français, car elle actualise les UB en leur attribuant des définitions partielles, et disjointes.

Ensuite, nous résumons les différents schèmes sémantiques du mot-valisage décrits par (Fradin, 2000) et indiquons comment nous pouvons les intégrer dans notre approche distributionnelle de la classification sémantique automatique de mots-valises.

Enfin, nous évaluons notre générateur en menant une tâche de génération automatique. Pour cela, nous constituons un corpus de mot-valises attestés (Cartier, 2019; Sajous & Hathout, 2015) — et identifions leurs bases —, puis annotons manuellement chacune des unités lexicales selon la classification sémantique que nous employons à l'aide du guide d'annotation établi[2]. Nous utilisons un modèle Word2Vec (Mikolov et al., 2013) — entraîné par Fauconnier (2015) sur un corpus de 600 000 tokens issu d'un dépôt de Wikipédia en français — pour calculer la représentation vectorielle des mots-valises à partir des représentations de leurs bases. Nous classifions les mots-valises en fonction de la proximité de leur vecteur avec les barycentres des classes sémantiques calculées en moyennant les représentations des unités lexicales les composant d'après le corpus FrSemCor (Barque et al., 2020).

1 En français québécois, terme désignant les courriels indésirables.

2 https://github.com/FrSemCor/FrSemCor/blob/master/guideAnno-FR-SemCor.pdf

BIBLIOGRAPHIE

Barque, L., Haas, P., Huyghe, R., Tribout, D., Candito, M., Crabbé, B., & Segonne, V. (2020, mai). FrSemCor : Annotating a French corpus with supersenses. *LREC-2020*. https://hal.archives-ouvertes.fr/hal-02511929

Cartier, E. (2019). Neoveille, plateforme de repérage et de suivi des néologismes en corpus dynamique. *Neologica*, *2019*(13), 23-54.

Dugas, E., Haas, P., & Marin, R. (2021). Héritage sémantique des noms aux verbes : Étude des verbes dénominaux en français. *Verbum*, *38*, 30p.

Fauconnier, J.-P. (2015). *French Word Embeddings*. http://fauconnier.github.io

Fradin, B. (2000). Combining forms, blends and related phenomena. *Extragrammatical and Marginal Morphology. München: Lincolm Europa*, 11-59.

Fradin, B. (2021). De la variété des paradigmes dérivationnels. *Verbum*, *38*, 26p.

Fradin, B., Montermini, F., & Plénat, M. (2009). Morphologie grammaticale et extragrammaticale. In *Aperçus de morphologie du français* (p. 21-45).

Huyghe, R., & Wauquier, M. (2020). What's in an agent? *Morphology*, *30*(3), 185-218. https://doi.org/10.1007/s11525-020-09366-2

Huyghe, R., & Wauquier, M. (2021). Une étude distributionnelle des noms d'agent en -ant, -eur, -ien, -ier, et -iste. *Verbum*, *38*, 28p.

Léturgie, A. (2011). À propos de l'amalgamation lexicale en français. *Langages*, *183*(3), 75. https://doi.org/10.3917/lang.183.0075

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1990). Nouns in WordNet : A Lexical Inheritance System. *International Journal of Lexicography*, *3*(4), 245-264. https://doi.org/10.1093/ijl/3.4.245

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet : An On-line Lexical Database*. *International Journal of Lexicography*, *3*(4), 235-244. https://doi.org/10.1093/ijl/3.4.235

Sajous, F., & Hathout, N. (2015). *GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary*. 22.