Diffusion des innovations lexicales sur Twitter : description et prédiction de l'influence de la position des locuteurs dans le réseau

Louise Tarrade, Jean-Pierre Chevrot, Jean-Philippe Magué

Diffusion des innovations lexicales sur Twitter : description et prédiction de l'influence de la position des locuteurs dans le réseau

Louise Tarrade [1], Jean-Pierre Chevrot [2], Jean-Philippe Magué [1]
[1] Laboratoire ICAR (UMR 5191), École Normale Supérieure de Lyon
[2] Laboratoire LIDILEM (EA 609), Université Grenoble Alpes
louise.tarrade@ens-lyon.fr, jean-pierre.chevrot@univ-grenoble-alpes.fr, jean-philippe.magueé@ens-lyon.fr

Introduction[1]

L'étude de la variation et du changement linguistique est au cœur de la sociolinguistique variationniste. Des théories majeures sur ce sujet ont émergé de ce domaine : parmi celles-ci, les propositions de Milroy & Milroy (1985) sur l'importance des liens faibles dans l'introduction des innovations, inspirées par le sociologue des réseaux Granovetter (1973) qui a mis en évidence la fonctionnalité de ce type de connexions. Ainsi Milroy & Milroy (1985) ont défini les innovateurs, les personnes apportant l'innovation dans leur communauté, comme ayant des positions périphériques à leur communauté avec de nombreux liens faibles à l'intérieur et à l'extérieur de celle-ci. De même, ils ont mis en évidence que pour qu'une variante s'établisse au sein d'une communauté, une condition nécessaire est qu'elle ait été préalablement adoptée par des personnes à la fois centrales et bien ancrées dans celle-ci. De son côté, Labov (2001) décrit les leaders du changement linguistique comme des personnes à la fois très centrales à leur communauté mais disposant également de nombreux liens à l'extérieur de celle-ci. Cependant, ces études portaient sur des variantes phonétiques, étaient menées à l'échelle d'une centaine d'individus, et souvent en synchronie.

La sociolinguistique computationnelle (Nguyen et al., 2016) permet d'échapper aux limites imposées par l'enquête de terrain, et d'étudier diachroniquement une variété de langue à mi-chemin entre l'écrit et l'oral, très propice à la variation et à l'innovation. À l'instar de la sociolinguistique variationniste traditionnelle, elle s'est donc en partie attelée à étudier le changement linguistique et son processus de diffusion. Par exemple, des simulations multi-agents ont permis de mettre en évidence l'importance des membres périphériques dans l'apport d'innovations et celle d'individus au centre de sous-groupes denses dans l'apparition de normes (Fagyal et al., 2010). Au niveau des médias sociaux, une attention particulière a été portée à l'importance des liens faibles dans l'apport des innovations et à l'influence des liens forts, sur des corpus issus de Facebook (Bakshy et al., 2012), Twitter (Goel et al., 2016), ou Reddit (Del Tredici & Fernández, 2018), ou encore à l'influence de la structure du réseau sur la diffusion des innovations (Zhu & Jurgens, 2021 ; Würschinger, 2021). Cependant, ce changement d'échelle questionne les notions traditionnelles de réseau, de lien et de communauté. De plus, le schéma global de l'influence de la structure du réseau aux

[1] Les auteurs remercient le LABEX ASLAN (ANR-10-LABX-0081) de l'Université de Lyon pour son soutien financier dans le cadre du programme français "Investissements d'Avenir" géré par l'Agence Nationale de la Recherche (ANR).

différentes étapes de diffusion des innovations ainsi que les paramètres qui influent sur le succès ou l'échec de ces dernières restent encore flous.

En nous basant sur la trajectoire en forme de S des innovations réussies (Blythe & Croft, 2012 ; Fagyal et al., 2010 ; Rogers, 2003), nous chercherons à caractériser les sous-populations d'utilisateurs qui s'emparent des innovations aux différentes étapes de leur diffusion. Plus précisément, il s'agira d'essayer de dégager un schéma général de la façon dont le profil des utilisateurs évolue au fur et à mesure de la diffusion d'une innovation lexicale. En comparaison avec des innovations dont la trajectoire aboutit à l'échec de leur établissement dans la communauté linguistique, nous nous demanderons qui sont les acteurs du changement et si leur position dans le réseau aux différentes phases entrave ou participe à la diffusion de ces innovations.

Corpus et méthodologie

Corpus

Nous nous appuyons pour ce travail sur un corpus d'environ 650 millions de tweets en français rédigés de 2007 à début 2019, et provenant d'environ 2,5 millions d'utilisateurs. Une collecte initiale de 170 millions de tweets produits entre 2014 et 2017 dans les fuseaux horaires GMT et GMT+1 constitue le socle de ce corpus (Abitbol et al., 2018). Celui-ci a été complété dans un second temps par une récupération des derniers tweets des utilisateurs à l'aide de l'API Twitter, puis filtré en fonction de la langue et du client utilisé afin de ne garder que les tweets en français et d'éliminer autant que possible les tweets provenant de bots.

En parallèle a été récupéré pour chaque utilisateur l'ensemble de ses followees, c'est-à-dire des personnes qu'il suit. À partir de ces informations, nous avons reconstitué le réseau des utilisateurs de notre corpus, dont les nœuds sont les utilisateurs, qui peuvent avoir des liens entrants (followers) et sortants (followees) internes, c'est-à-dire que nous considérons seulement les liens entre les utilisateurs de notre corpus. Nous obtenons finalement un réseau dirigé et statique de plus de 2,5 millions de nœuds et 300 millions de liens.

Méthodologie

À partir du réseau modélisé des utilisateurs du corpus, nous avons distingué les communautés en nous appuyant sur le principe de l'algorithme de Louvain (Blondel et al., 2008), puis nous avons caractérisé chaque utilisateur[2] en fonction des mesures suivantes :

- le coefficient de clustering local, qui rend compte pour un utilisateur du degré d'ouverture de son réseau ;
- le score de PageRank, qui est un indicateur du prestige d'un individu et va dépendre à la fois du nombre de ses liens entrants mais également de si ces liens entrants ont eux-mêmes un score de PageRank élevé ;
- la centralité d'intermédiarité, qui mesure à quel point l'utilisateur est central à sa communauté et fait office de "pont" au sein de celle-ci ;

[2] Au regard de la taille conséquente du réseau, la modélisation du graphe ainsi que les calculs des différentes variables de réseau - à l'exception de la proximité avec l'extérieur de la communauté - ont été effectués à l'aide de la librairie Networkit (Staudt et al., 2014).

- la proximité avec les autres communautés, qui correspond au nombre de « pas » moyen nécessaire pour un utilisateur avant de pouvoir atteindre une autre communauté.

Dans un précédent travail (Tarrade et al., 2022) nous avons détecté les innovations lexicales apparues dans le corpus de tweets entre mars 2012 et février 2014. Plus précisément, nous avons sélectionné tous les mots[3] apparus pour la première fois au cours de cette période et avons conservé les mots dont la trajectoire d'utilisation sur 5 ans s'ajustait à une courbe logistique ou à une courbe gaussienne. Nous les avons ainsi catégorisés en tant que changement ou buzz selon si leur taux d'utilisation réussissait à se stabiliser au fil du temps ou non (figure . 1). Nous avons ensuite délimité de façon automatique leurs trois phases de diffusion, à savoir innovation, propagation, puis fixation pour les changements ou déclin pour les buzz. Nous avons ainsi réussi à identifier 141 changements et 251 buzz.

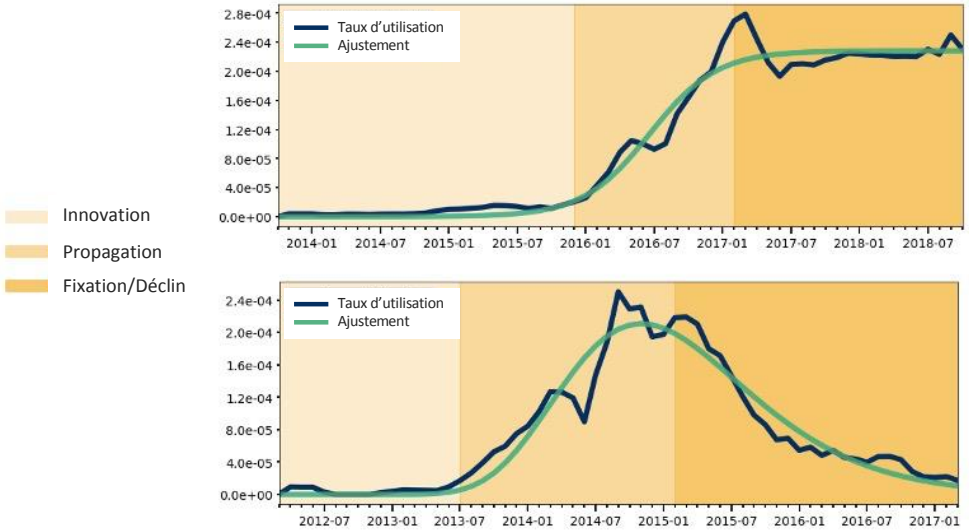

figure . 1 Deux exemples d'ajustement : « malaisante » (changement, en haut) et « sweg » (buzz, en bas)

Pour le travail présenté ici, nous avons ajouté une condition contrôle composée de 200 mots utilisés entre février 2013 et janvier 2018, dont le taux d'utilisation est stable tout au long de cette période, et dont la distribution en termes de nombre d'utilisateurs est similaire à celle des innovations lexicales que nous avons préalablement identifiées.

Afin de dégager un schéma global de diffusion de ces innovations lexicales, nous avons caractérisé chaque innovation, à chacune de ses phases de diffusion, en fonction des caractéristiques de réseau des utilisateurs l'ayant employée pour la première fois durant la phase observée. Chaque innovation et chaque mot contrôle se voit ainsi attribuer une valeur pour les quatre variables de réseau décrites précédemment. Nous comparons ensuite les distributions de ces valeurs pour les différentes catégories de mot, afin de déterminer si et comment elles diffèrent à chacune de ces phases.

Pour confirmer nos observations, nous tentons ensuite de prédire le destin des innovations lexicales avant que leur trajectoire ne se stabilise ou ne décline, soit dès la phase d'innovation ou de propagation. Pour se faire, nous entraînons un modèle de régression logistique[4] sur les innovations lexicales de notre jeu de données pour faire de la classification binaire : la

[3] C'est à dire toute suite de caractères alphanumériques pouvant contenir une apostrophe ou un tiret.
[4] Nous utilisons pour cela l'algorithme de classification par régression logistique de la librairie Scikit-learn.

variable à prédire est le type d'innovation lexicale (changement ou buzz), et les variables explicatives sont les valeurs médianes de l'ensemble des utilisateurs de chaque mot pour chacune des variables de réseau.

Résultats

Dans un premier temps, nous avons pu établir que les innovations lexicales sont utilisées pour la première fois par des utilisateurs aux caractéristiques de réseau relativement similaires, que ces innovations connaissent plus tard le succès ou l'échec. Contrairement à ce que nous pouvions attendre, elles ne sont pas utilisées lors de la phase d'innovation par des individus dont le réseau personnel est plus ouvert ou plus fermé que la moyenne. Ceux-ci ont la possibilité d'être plus vite en contact avec d'autres communautés, sans être ni centraux dans leur propre communauté, ni prestigieux au sein du réseau global. Un profil des premiers adoptants des innovations lexicales (les innovateurs) se dessine donc dès la première phase de diffusion, mais sans que celui-ci ne se distingue réellement d'un type d'innovation à l'autre - buzz ou changement.

Mais le destin des innovations lexicales se joue en phase de propagation. Lors de cette phase, les changements sont caractérisés par des adoptants plus prestigieux que ceux des buzz, à la fois centraux à leur communauté, mais également situés plus à proximité des autres communautés. À l'inverse, les buzz ne réussissent pas à atteindre des utilisateurs centraux ou ayant une proximité avec des utilisateurs extérieurs à leur communauté, ce qui entrave à priori leur diffusion dans le réseau global.

Cette configuration de paramètres distinguant les buzz des changements est confirmée par les résultats de la prédiction par régression logistique obtenus en phase de propagation, avec un taux de précision de plus de 80%. Ces résultats esquissent un schéma général de diffusion du changement linguistique en fonction des positions des locuteurs dans le réseau.

En conclusion, nous discuterons de leur adéquation avec les conclusions des précédentes études menées en sociolinguistique variationniste, notamment en ce qui concerne le profil des innovateurs décrit par Milroy & Milroy (1985) et celui des meneurs du changement par Labov (2001).

Références bibliographiques

Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P., & Fleury, E. (2018). Socioeconomic Dependencies of Linguistic Patterns in Twitter : A Multivariate Analysis. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 1125-1134. https://doi.org/10.1145/3178876.3186011

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. *Proceedings of the 21st International Conference on World Wide Web*, 519–528.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Blythe, R. A., & Croft, W. (2012). S-CURVES AND THE MECHANISMS OF PROPAGATION IN LANGUAGE CHANGE. *Language*, *88*(2), 269–304. JSTOR.

Del Tredici, M., & Fernández, R. (2018). The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities. *ArXiv:1806.05838 [Cs]*. http://arxiv.org/abs/1806.05838

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L., & Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua*, *120*(8), 2061–2079. https://doi.org/10.1016/j.lingua.2010.02.001

Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F., & Eisenstein, J. (2016). The social dynamics of language change in online networks. *International Conference on Social Informatics*, 41–57.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469

Labov, W. (2001). *Principles of linguistic change*. Blackwell.

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, *21*(2), 339–384.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational Sociolinguistics: A Survey. *ArXiv:1508.07544 [Cs]*. http://arxiv.org/abs/1508.07544

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed). New York: Free Press.

Staudt, C. L., Sazonovs, A., & Meyerhenke, H. (2014*). NetworKit: A Tool Suite for Large-scale Complex Network Analysis*. https://doi.org/10.48550/ARXIV.1403.3005

Tarrade, L., Magué, J.-P., & Chevrot, J.-P. (2022). Detecting and categorising lexical innovations in a corpus of tweets. *Psychology of Language and Communication*, *26*(1), 313-329. https://doi.org/10.2478/plc-2022-15

Würschinger, Q. (2021). Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter. *Frontiers in Artificial Intelligence*, *4*, 648583. https://doi.org/10.3389/frai.2021.648583

Zhu, J., & Jurgens, D. (2021). The structure of online social networks modulates the rate of lexical change. *ArXiv Preprint ArXiv:2104.05010*.