Psychology of Language and Communication 2022, Vol. 26, No. 1

sciendo

DOI: 10.2478/plc-2022-15

Louise Tarrade[1], Jean-Philippe Magué[1], Jean-Pierre Chevrot[2]
[1] ICAR laboratory (UMR 5191), École Normale Supérieure de Lyon, France
[2] LIDILEM laboratory (EA 609), Université Grenoble Alpes, France

Detecting and categorising lexical innovations in a corpus of tweets

In this paper, we present the methodology we have developed for the detection of lexical innovations, implemented here on a corpus of 650 million of French tweets covering a period from 2012 to 2019. Once detected, innovations are categorized as change or buzz according to whether their use has stabilized or dropped over time, and three phases of their dynamics are automatically identified. In order to validate our approach, we further analyse these dynamics by modelling the user network and characterising the speakers using these innovations via network variables. This allows us to propose preliminary observations on the role of individuals in the diffusion process of linguistic innovations which are in line with Milroy & Milroy's (1997) theories and encourage further investigations.

Key words: computational sociolinguistics, linguistic innovations, S-curve, Twitter, big data, network, diffusion of innovations

Address for correspondence: Louise Tarrade
ICAR laboratory (UMR 5191), École Normale Supérieure de Lyon, 15 parvis René Descartes, 69007, Lyon, France.
E-mail: louise.tarrade@ens-lyon.fr

The diffusion process of linguistic innovations has long been a topic of interest in sociolinguistics (Weinreich et al., 1968). Many studies have highlighted the influence of social structures on this process (Labov, 2010; Milroy & Milroy, 1997). The recent access to massive social network data and the advent of computational sociolinguistics (Nguyen et al., 2016) allow an approach to this phenomenon that combines a large amount of data and a fine-grained temporality. It is from this perspective that we present an approach for detecting and categorising linguistic innovations, more specifically lexical innovations. We relied on the idealized S-shaped trajectory of successful innovations (Blythe & Croft, 2012; Feltgen et al., 2017; Rogers, 2003) to identify lexical innovations in a corpus of French tweets and categorize them according to whether their use has stabilised (change) or not (buzz) over time. We then automatically detected the three phases of diffusion (Chambers, 2013; Fagyal et al., 2010) of these new forms: innovation, propagation, fixation for a change, or decline for a buzz.

In order to validate our approach for detecting lexical innovations, categorizing according to their fate, and identifying the successive phases of their dynamics, we tested the hypothesis that changes and buzzes spread differently across the network of users.

The first section presents relevant previous works which will allowed us to formulate our hypothesis more precisely. The second section presents in detail the data, and describes the method we propose to detect, categorizes, and delimit the diffusion phases of lexical innovations as well as our validation method. The third section presents the results obtained, which are discussed in the last section. All the codes used and the results obtained are available on our GitHub repository.[1] However, the corpus of tweets cannot be made available in order to respect the privacy of users.

Previous Work

A topic of interest in sociolinguistics for many years, linguistic change corresponds to the outcome of a process in several steps described by Weinreich et al. (1968). First, a speaker introduces a new form in their use of language, then this form is taken up and used by other speakers, and finally, the use of this form stabilizes in a community. We can consider this to be a change. These phases in the establishment of a linguistic innovation as a change are variously named *innovation*, *propagation*, and *fixation* by Fagyal et al. (2010) or initial stasis, rapid rise, and tailing off by Chambers (2013). The idealised S-shaped trajectory observed by Rogers (2003), confirmed at the linguistic level by Blythe and Croft (2012), and later validated on a large scale by Feltgen et al. (2017), accounts for these phases in the case of successful innovations. The role of social structures and individuals in this process of diffusion has also been discussed, and

1 https://github.com/LTarrade/lexical_innovation_detection

sociolinguistics has tried to identify the innovators in the process of diffusion of a linguistic innovation, as well as how their position in the network influences this process. Labov (2010) describes leaders of linguistic change as generally female, middle-class individuals who are both very central to their local community and have a large number of connections outside of it. Milroy and Milroy (1997) observed phonological variation in three different neighbourhoods of the city of Belfast and noted that people who were very central to their neighbourhoods were also very conservative with respect to vernacular norms, but also, that innovations were mainly introduced by young women who worked in shops where the different communities met, and who were therefore in regular contact with them, but had no strong ties to them. For Milroy and Milroy (1997), the innovators are therefore people with weak ties, on the periphery of the communities. They are the ones who bring the different variants into the community, but for that variant to be adopted by a community, it is necessary that it is first adopted by people with strong ties and who are very central to the community. Earlier, Granovetter (1973) had also highlighted the importance of weak ties in the transmission of innovations. However, these studies were most often conducted on phonetic changes, on populations of hundreds of individuals at most.

Over the past decade, access to massive digital data has allowed the emergence of computational sociolinguistics, which approaches the issues of sociolinguistics by combining the methodologies of natural language processing and data science (Nguyen et al., 2016). In addition to the methodological renewal it has brought about, the interest of computational sociolinguistics also lies in the nature of the data it uses. Often derived from social media, these data document language varieties that are not very standardized, showing a high variability and a high rate of innovation. Computational sociolinguistics has thus been able to address the issue of language change and some of its theories have been partly transposed into social media studies. Thus, Del Tredici and Fernández (2018) highlighted in the observation of the diffusion of linguistic innovations within Reddit communities that the innovators seem to correspond to the hubs of the community, that is, individuals with many but weak connections. They also demonstrated that the adoption of an innovation by the community seems to be conditioned by the fact that it is adopted by members with strong ties. Moreover, Laitinen et al. (2020), using a corpus of tweets, underlined the importance of the size of the network and showed that, above a certain size, the fact that networks are mostly composed of weak (more propitious to innovation) or strong ties (more conservative and resistant to change) no longer seems to constitute a significant distinction in the resistance to change. Fagyal et al. (2010) used multi-agent simulations to study the role of individuals in the diffusion of a linguistic innovation and its adoption as a norm. Using a number of tests carried out by varying the parameters of their network, they highlighted that, in general, leaders (hubs) push forward the change in progress and are indispensable for establishing it as a norm, and loners are the repositories of old or new variants

and their absence from the network results in a lack of innovation. One of the most promising avenues of research in the field of linguistic variation and change on social media is the consideration of psychological and social factors through the relationship between personality and social network. The structure and size of the communities of contacts established by Facebook users depend on their degree of extraversion (Friggeri et al., 2012), a trait that is also associated with a greater general tendency to innovate (Ali, 2019), which could also manifest itself in the field of language. While these studies partly addressed the way in which social structure influences the process of linguistic diffusion, they often remain confined to one aspect and do not provide a comprehensive view of this diffusion process in action at different population levels.

However, before being able to study this mechanism, one must first identify the linguistic innovations to be studied, and the very noisy nature of social media corpora does not facilitate this task. The methods used for the detection of linguistic innovations often focus on the detection of semantic changes, in particular because of the possibilities offered by the first word embedding techniques allowing to represent words in vector spaces with, for example, the word2vec algorithm proposed by Mikolov et al. (2013) and in particular since the appearance of language models based on deep learning such as BERT (Devlin et al., 2019) and its successors. For our part, we were more interested in lexical innovations, that is, the appearance of new words. The majority of studies on lexical changes in computer mediated corpora rely on frequency analysis to detect them. Among them, Eisenstein et al. (2014), in their work on linguistic diffusion networks on Twitter between different metropolitan areas in the USA, selected the words used in their analysis based on the 100,000 most frequent terms. After imposing a minimum frequency of use, they calculated the variance of the algorithmic probability of each of them to select only the words with a variance above a certain threshold. From the 5,000 words obtained, they manually eliminated both named entities and non-English words, and determined for each remaining word whether its usage is similar to that of an English word based on examples of word usage in context. Meanwhile, Costin-Gabriel and Rebedea (2014) used word evolution images provided by the Google Books N-gram Viewer and principal component analysis techniques to find the general patterns of three types of words: common words, neologisms, and archaisms, which they listed manually. They calculated the proximity of the evolution of the word to be classified to each of the three trends to determine which type of word it was. Tjong Kim Sang (2016) tested two methods of detecting neologisms and archaisms in a corpus of magazine texts as well as a corpus of tweets, one calculating a score from the ratio of the starting frequency to the ending frequency, and another, less effective on tweets, whose score depended on the correlation coefficients between word frequencies and time. At the same time, Kershaw et al. (2016) relied on two methods originally intended for lexicographers to measure the acceptance of linguistic innovations. They implemented three statistical

tests based on the variations of the frequency, meaning, and morphology of the forms, which they applied to a corpus of tweets and a corpus of Reddit posts. To measure the importance of tie strength in the adoption of an innovation within communities, Del Tredici and Fernández (2018) used an already existing online lexicon of slang terms from the internet to identify these innovations. They then categorized innovations as either successful or unsuccessful based on the slope of diffusion of each term. Stewart and Eisenstein (2018) highlighted the importance of linguistic rather than social diffusion in the maintenance or decline of nonstandard words, and showed that one of the factors determining their stabilisation is their membership to a wider variety of lexical contexts. To do so using a Reddit corpus, they used the frequency of words over time to identify the words with increasing frequency using the Spearman coefficient and the words with decreasing frequency by fitting the frequency series of words with a two-phase piecewise linear regression and with a logistic distribution for the more discrete growth and decay trajectories. On the other hand, Kerremans and Prokić (2018), chose a semi-automatic detection of neologisms on the web using correspondence dictionaries.

The approaches for detecting lexical innovations mentioned above almost systematically using English corpora, only partially respond to our needs. Often, they either require manual steps that are quite time-consuming, they require the use of dictionaries, or they seem to focus more on innovations in the growth phase rather than on innovations that have stabilised, and do not seem to seek to identify the three phases of diffusion of innovations. Furthermore, as with the detection of semantic changes (Schlechtweg et al., 2019), it is difficult to assess the performance of these methods given the diversity of the datasets and the lack of evaluation of this task.

Methodology

Data Presentation

Our data was collected in two steps. The first corpus of tweets was first collected as described in (Abitbol et al., 2018), spanning from June 2014 to March 2018. Afterwards, the second collection of tweets was carried out, which consisted in updating the first corpus by taking each of its users and retrieving their last tweets using the Twitter API. The tweets thus obtained were filtered according to the language detected by Twitter (French) and the client used (in order to filter out robots). In the end, the cleaned corpus data included about 650 million tweets in French from just over 2.5 million users, covering a period from 2007 to February 2019, with 98% of the corpus concentrated between 2012 and 2019. For each tweet, we gathered a set of metadata such as the creation date or the identifier as well as information about the user who produced the tweet such as the identifier, the number

of followers or the number of followees,[2] that is, the other accounts they follow.

However, for about 20,000 accounts, this information could not be retrieved for various reasons, as in the case when the Twitter account has been deleted in the meantime. This collection allowed us to reconstruct the network of the corpus users by modelling it as a directed graph composed of nodes (users) that can include incoming ties (followers) and outgoing ones (followees). This network is thus a directed, static, and closed graph.

Detection and Categorization of Lexical Innovations

From the tokenized tweets, we retrieved all of the new tokens in the corpus as follows. Since we were interested in the dynamics of innovation after their appearance, we selected the tokens that appeared between March 2012 and February 2014 (which were totally absent from the corpus for the whole of the previous year) in order to have at least a five-year period to observe the evolution of their use. We filtered these forms by keeping only those that have been used by at least 200 different users, and using regular expressions, we excluded hashtags, emojis, or punctuation marks in order to focus the rest of our analysis exclusively on words, which were then defined as any sequence of alphanumeric characters that may contain an apostrophe or a dash.

For each of the new forms thus recovered, we retrieved their usage rate each month over a period of five years. The usage rate of a form in a given month corresponds to the ratio between the number of people who used this form that month and the number of people who tweeted in the month. Therefore, for each emerging form, and for each linguistic innovation, we obtained the trajectory of its use among the users of our corpus during its first five years of use.

In order to categorise each innovation as either a change (an innovation whose usage rate stabilised after experiencing exponential growth) or a buzz (an innovation that also experienced a phase of exponential growth, but whose usage eventually declined to a very low rate), we used a curve fitting method. To reduce the influence of accidental peaks in the trajectory of the usage rate of each observed form, we considered the rolling average of this rate with a three-month window. Then, using the LMFIT[3] library for Python, we fit the usage trajectories of each form to two reference functions:

- The logistic function (the S-shaped curve followed by the changes), defined as $f(x, A, \mu, \sigma) = A[1 - \frac{1}{1+e^{\alpha}}]$, where $\alpha = (x - \mu)/\sigma$ and where A is the amplitude, μ is the center, and σ is the sigma parameter (which influences the steepness of the inclination of the curve slope).
- The lognormal function (a skewed bell-shaped curve followed by the buzzes), which is defined by $f(x; A, \mu, \sigma) = \frac{A}{\sigma\sqrt{2\pi}} \frac{e^{-(ln(x)-\mu)^2/2\sigma^2}}{x}$, where A is the amplitude, μ is the center, and σ is the sigma, that is, the characteristic width of the peak.

2 https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

3 https://lmfit.github.io/lmfit-py/index.html

Forms for which the reduced chi-square (a measure of quality of fit) was above a certain threshold (0.00005) for both functions were ruled out. Otherwise, the best of the two fits determined whether the form was a change or a buzz. In order to observe the trajectory of innovations in their entirety, we also sorted out the values of some output parameters of the fit.[4]

As the latter influenced each other, the choice of the limit attributed to their value was decided following observations of the impact of each of these parameters on the others. Following this selection, we ended up with just over 500 words whose use seems to follow a logistic or lognormal trajectory very closely and whose trajectory is almost entirely observable over the five-year period.

As mentioned in the Introduction section, a successful language innovation follows three phases of diffusion. During the first phase, the innovation phase, the form spreads only very slowly through the population. In the second phase, propagation, the form spreads among users exponentially until it reaches a threshold at which the use of the form stops growing but remains high, which is the fixation phase. When its use stabilizes, a change can be considered to have occurred. If an innovation fails to stabilise and instead experiences a third phase of decline, we consider it to be a buzz.

To delimit the three phases of diffusion of innovations, we searched for the maximums of the third derivative, that is, the moments when the acceleration in the spread of the form varied the most. The first maximum marks the boundary between the innovation and the propagation phases, the second marks the boundary between the propagation and the fixation or decline phases.

Validation of the Method for Detecting and Classifying Linguistic Innovations

In order to validate our approach of detecting lexical innovations and our classification as buzzes or changes, we hypothesized that they are distinguished from each other by the position in the network of the users who adopted them at different phases of diffusion. Based on what is reported in the literature, in particular the theory on the process of diffusion of an innovation within a language community proposed by Milroy and Milroy (1997), lexical innovations should be introduced by people who are rather on the periphery and in contact with several different social groups. Similarly, they should only start to stabilize after being adopted by people who are very central to the community, who, transposed on our corpus of tweets, could be identified as people with many incoming ties and, therefore, with a certain prestige. To confirm this, we characterized each user of our corpus by network variables and looked at the distribution of these variables for each of the phases defined above.

4 Sorting on the output parameters of the logistic curve fit: (((center>=16) & (center<=31) & (sigma<=8)) | ((center>31) & (center<=46) & (sigma<=7))) & (redchi<0.00005) & (amplitude>0.02) & (center_err<5); Sorting on the output parameters of the lognormal curve fit: (fwhm>=4) & (fwhm<=40) & (redchi<=0.00005) & (amplitude<=1.1) & (maxPoint>=21) & (maxPoint<=46) & (((center<=3.6) & (sigma<=0.65)) | ((center>3.6) & (center<=3.8) & (sigma<=0.35)) | ((center>3.8) & (sigma<=0.15)))

To extract network variables, we relied on the user network, modelled as described in the Data Presentation section. For each user, we calculated a PageRank score (Brin & Page, 1998) using the SNAP network analysis library (Leskovec & Sosič, 2016). The PageRank score corresponds to a measure of user prestige. More concretely, it is calculated from the frequency of visits of each node (user) by a random walk. Therefore, the score will be influenced both by the number of incoming ties (followers) of each node, but also by the respective prestige of the incoming nodes. Thus, the higher the score, the more prestige a node has. In the same way, we characterized each node by its local clustering coefficient. The clustering coefficient of a node is calculated by considering the graph as an undirected graph (all links between users are considered, regardless of whether they are incoming or outgoing) and by calculating the number of effective triads out of the number of possible triads for each node, thus looking at the proportion in which the neighbors of a node are connected to each other. Therefore, this measure is an indicator of the openness of each user network. The higher the clustering coefficient of a user, the more closed their own network is, so the more their friends are also friends with each other.

Our goal was to understand how and when these variables affect the process of acceptance or nonacceptance of innovations. For each form and for each user using that form, we recorded the phase when they used the form for the first time. Thus, for each type of innovation (change or buzz), we recovered all the users who adopted a form of this type at each phase (innovation, propagation, and fixation/decline). We then compared the distribution of the variables across the type of innovation and the phases. These six distributions were compared to the distribution in the whole population as well. A user may appear twice since they may have used one form in one phase and another form during another phase. This nonindependence constrained the statistical tests used to compare the distributions. However, only 18% of the users have used at least one buzz and one change. The great majority of users used either only buzzes or only changes, and on average, they used two different forms, the median being one form.

Since the distributions of the variables were not normal, we based our comparisons on the median and the first and third quartiles rather than on the means. We also used the nonparametric Kruskal-Wallis and Wilcoxon-Mann-Whitney statistical tests to ensure the significance of our observations. Finally, as the number of individuals in the observed samples varied considerably, we also ensured that this parameter did not influence our results by using a bootstrapping technique. More precisely, for each observed sample (e.g., the distribution of the number of incoming ties of users who used a buzz for the first time during the period of innovation of the form), we carried out 1,000 random samples without replacement, of the same number of individuals as in the observed sample, and then ensured that the median of this sample did not lie within the 95% confidence interval of the medians observed on the 1,000 random samples.

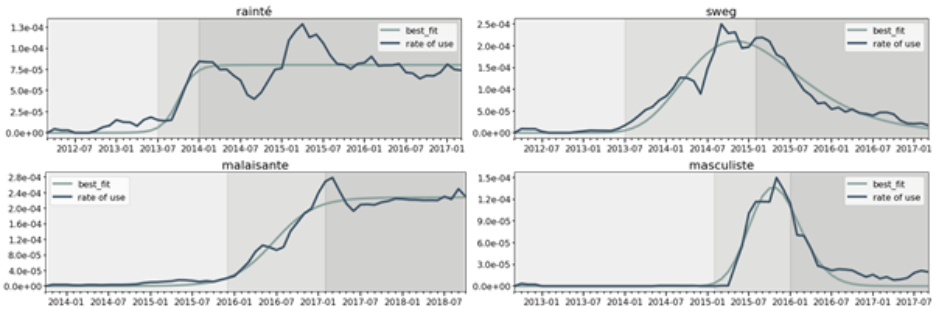
Results

Identifying Changes and Buzzes

The method of detection and categorization according to the trajectory of use of new linguistic forms over the months described in the previous section allowed us to identify a set of lexical innovations, and categorize them as changes (C) or buzzes (B). Figure 1 shows examples of linguistic forms identified as changes (left) or as buzzes (right) and for which the monthly rate of use over five years follows either a logistic function or a lognormal function, respectively. However, note that while the method correctly recognized the different types of curves we were looking to identify, the usage trajectory of the words was never ideally fitted with the reference curve due to the noisy nature of the data.

Focusing on the changes and buzzes identified by our method, we can observe a large number of neologisms, most of them linked to new realities (fullstack (C), émoji (C), mutuals (C), snapé (C), streameur (C)) or practices (twerké (C), dabé (C), binge-watching (C)), as well as to social phenomena (islamogauchiste (C), féminazis (C), masculiste (B), agenre (C)). We can also find archaic reactualized linguistic forms (malaisante (C)), new linguistic forms (enlové (B), sweg (B), baé (C), ggwp (C), oklm (C)) with a large number of morphological derivations (cuissance (B), mignonance (B), coulance (C), génance (C)), but also borrowings (mskn (C), mutuals (C), kehba (B), sadlife (B)) or slang words (bresom (C), rainté (C), kecho (B), peufra (B)). Many variations of new linguistic forms were also present, especially in the buzzes (miskinou (B), oklmus (B), tchuips (B)), but also phonological variations (chumor (C), aoé (C), aeq (B), caley (B)), agglutinations (heinquoi (C), balecouilles (B)), abbreviations (batrd (B), qtv (C)), or simple spelling variations (parasyte (C), embiancer (B)). The buzzes also included a very large number of lengthenings (oklmm (B), mddddddrr (B),

Figure 1. *The Usage Rate per Month of Two Changes (Left) and Two Buzzes (Right) Represented by a Rolling Average with a Three-Month Window (Blue), as well as the Result of the Curve Fitting (Green). The Three Diffusion Phases are Represented by the Grey Shading in the Background.*



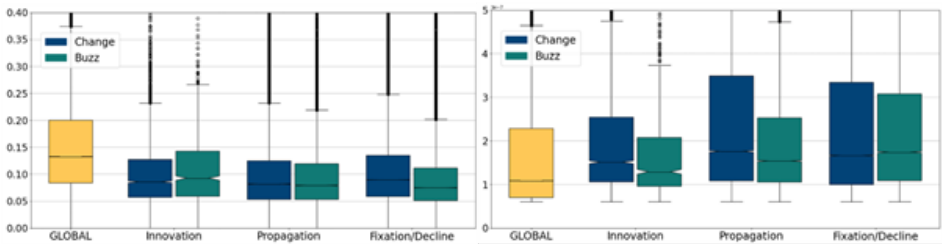
flmmm (C)). Finally, despite an initial automatic filtering of proper nouns, there remained a large number of named entities (just over 200 in the two categories of innovations combined), which we set aside after manual filtering. To conclude, we have identified 141 changes and 251 buzzes.

Analysis of Network Variables

A first comparison of the distribution of the two network variables of the buzz and change users at the different phases with the distribution in the global population allowed us to perceive some interesting phenomena. First, the users who used buzzes or changes were both clearly different from the overall population, even if not to the same degree and not always following the same dynamics. Second, the comparison of the network variables was in line with our hypothesis, especially for the changes.

The left panel of Figure 2 shows the distributions of the clustering coefficient of the overall population and of the users of the buzzes and changes during the different diffusion phases. As a reminder, the clustering coefficient of a user is an indicator between 0 and 1 of how much their friends are also friends with each other. The higher this coefficient, the more the user is part of a closed network, the less new information has the possibility to reach this subnetwork. The clustering coefficient distributions of the users of changes and buzzes in the innovation phase had much lower values than those of the overall population (Mann-Whitney = 3.9e+10; $p < .01$ (C); Mann-Whitney = 1e+10, $p < .01$ (B)), which would suggest that the innovations appear in both cases in more open networks than normal, in which new information can more easily enter. In the propagation phase, if buzzes and changes follow the same dynamics with a clustering coefficient that continues to decrease, buzzes experience a much stronger decrease (median at -1.4e-02 compared to the innovation phase; Mann-Whitney = 1.6e+08, $p < .01$) than changes (median at -3.9e-03 compared to the innovation phase; Mann-Whitney = 1e+09, $p < .01$). Finally, users who adopted changes during their last diffusion

Figure 2. *Distributions of the Clustering Coefficients (Left) and PageRank Scores (Right) for the Whole Corpus in Yellow and for the Users of the Changes (Blue) and Buzzes (Green) at the Three Different Phases of Diffusion (from Left to Right: Innovation, Propagation, Fixation/Decline).*



phase (fixation) had a clustering coefficient that rose and tended to get closer to that of the global population, with a median at 8.9e-02 and, above all, a greater dispersion of the data towards higher clustering coefficient values (thus, a more closed network), contrary to those who adopted buzzes during the corresponding phase, with a median at 7.5e-02 (Mann-Whitney = 4.6e+09, $p < .01$).

The right panel of Figure 2 shows the distributions of the PageRank score, an indicator of user prestige, influenced both by the number of incoming degrees of a user and by the prestige of these incoming degrees. The higher the PageRank score, the more prestige the user has. To illustrate this, the Twitter account in our corpus with the highest PageRank score (6.55e-04) was the official Twitter account of the French newspaper Le Monde. Here too, the dynamics between the different phases of diffusion of changes and buzzes differed. While users of the two categories of innovations (changes and buzzes) who adopted the innovation during its first phase of diffusion had a much higher distribution of their PageRank score than that of the overall population (Mann-Whitney = 4.6e+09, $p < .01$ (C); Mann-Whitney = 1.2e+09, $p < .01$ (B)), that of users of changes was even higher (with a median of 1.5e-07, i.e., 1.17 times the median of buzzes, Mann-Whitney = 2.4e+06, $p < .01$, in the same phase and 1.39 times that of the overall population). In the propagation phase, the distribution of PageRank scores for changes continued to expand strongly towards higher values (median 1.62 times that of the overall population and third quartile at 3.48e-07, i.e., 1.53 times higher than that of the overall population; Mann-Whitney = 4.2e+10, $p < .01$) and that of buzzes followed the same dynamic, but with less expansion (median 1.42 times that of the overall population but with a third quartile at 2.5e-07, i.e., only 1.11 times higher than that of the overall population; Mann-Whitney = 3.1e+10, $p < .01$). Users who used a change for the first time during the fixation phase had lower median PageRank scores than during the previous phases (Mann-Whitney = 4.1e+09, $p < .01$), which would suggest the beginning of a decline in the distribution, while those who used a buzz for the first time during its decline period had higher median PageRank scores (Mann-Whitney = 9e+08, $p < .01$), almost reaching the values of that of the change propagation phase (1.73e-07). These observations suggest that if innovations are initially employed by more prestigious users than the average in both cases, those who adopt the changes in the propagation phase are even more prestigious, and one might suppose that they thus favor their diffusion. Also interesting was the decrease in the PageRank score during the last phase of diffusion of changes, with the opposite phenomenon in the phase of decline of buzzes. This could be interpreted as an indicator of the introduction of the innovation to a wider part of the population, with users with a lower prestige value who would in turn adopt the innovation, unlike the buzz.

The overall observation of the distributions of the two observed network variables, that is, the clustering coefficient, a measure of network openness, and the PageRank score, a measure of prestige, is in line with previous sociolinguistics studies. Indeed, the decrease of the clustering coefficient during the first two

phases of diffusion and its slight increase during the last phase suggests that the individuals described by Milroy and Milroy (1997) and Granovetter (1973) with weaker ties who are at the intersection of several communities (and therefore with less closed networks) are also at work here, before the innovation is appropriated by people with tighter networks. Similarly, the inverse dynamic observed during the diffusion of change phases with the PageRank score is in line with Milroy and Milroy (1997), for whom an innovation must first be adopted by the central (i.e., more prestigious) members of their community before being passed on to the rest of the community.

Discussion and Future Perspectives

We have presented a method for automatically detecting lexical innovations based on the trajectory of their rate of use in a corpus of Tweets over a period of five years, categorized them as changes or buzzes according to this same criterion, and then determined for each of them their three phases of diffusion: innovation, propagation, and diffusion for changes or decline for buzzes. This method is semiautomatic, in the sense that it required two manual interventions: one to restrict the values of the parameters of the curves fitted during the automatic categorization of the innovations, the other for a final filtering of the words obtained in order to remove the named entities. Although these two steps need to be improved to avoid any manual recourse afterwards, it should be noted that they are not extremely costly in terms of time.

Using two network variables reflecting, respectively, the degree of network openness of each user and their level of prestige, we validated the relevance of this detection method by showing that the distributions of these variables for the users of the two classes of lexical innovations both differed clearly from that of the overall population of the corpus, and that the evolution of the distribution of these variables, in particular for the changes, was in line with what had been observed in earlier sociolinguistic work. Nevertheless, from a more distant perspective, the rather similar dynamics observed in particular at the level of the first two phases of buzz and change diffusion suggests that what determines the adoption or not of a lexical innovation could also be located at other levels, which should be explored in parallel, in particular at the level of user communities. We can legitimately question the possibility whether the diffusion of innovations (whether buzzes or changes) that we observed is not really the diffusion of innovations in the overall population of our corpus of tweets, but rather the diffusion of these innovations within communities. In order to know whether these innovations are accepted within the communities in which they were created or whether have spread outside of them, we believe it is essential to model the different communities in our user network. Therefore, the next immediate step of our work will be to detect these communities in two ways: on the one hand, with community detection algorithms such as the Louvain algorithm (Blondel et al., 2008), and on the other, by finding communities of interest in a more traditional way, for example, by exploiting hashtags.

In order to continue to explore the importance of social structure in the diffusion and acceptance of innovations, it would be interesting to broaden the field of variables to be observed in order to characterize users more completely. Thus, it would be interesting to complete the network variables characterizing each user, in particular by calculating centrality scores such as the betweenness centrality, which gives an indication of the extent to which the observed node is a passage point for the other nodes of the graph, the centrality of proximity, or the Katz centrality, which would allow us to obtain an exposure index for each user (the higher the index, the more the user is exposed to the information). Similarly, we can consider including linguistic variables such as the number of tweets, measures of lexical diversity, and so forth. Another short-term objective would be to complete this characterisation with social variables, such as age or gender, which could be inferred using machine learning algorithms as has already been done, for example, by Wang et al. (2019) for gender, age, and organization status, Bamman et al. (2014) for gender, or Flekova et al. (2016) for age and income. Thus, we can also question the strength of the influence of other variables in the acceptance of a linguistic innovation. For example, does it depend more strongly on the variety of linguistic contexts in which it is used, as suggested by Stewart and Eisenstein (2018), or is the duration of the innovation or propagation phases decisive in the acceptance process?

Finally, we also plan in the near future to complete our study by including the detection and classification of semantic innovations, but also by addressing the diffusion of lexical change through a more qualitative approach, extracting the different contexts of use of a number of innovations at different periods of their diffusion to analyze the evolution of their usage and lexical contexts.

Conclusion

From a corpus of a hundred million tweets in French, produced by more than 2.5 million users and spanning several years, we automatically identified the lexical innovations present in the corpus and categorized them in the same way as changes (innovations whose use stabilised in the corpus over time) or buzzes (innovations whose use, after a period of growth, declined). We used the speed of diffusion of the rate of use of each form to determine the three characteristic phases of the diffusion of these linguistic innovations. We also modeled the network of users in the corpus and characterized each of these speakers with network variables indicating the level of openness of their own network and their level of prestige. The first observations of the distribution of these user variables at the different phases of buzz and change diffusion validated the efficiency of our method by bringing out, in particular for the changes, a diffusion dynamic described in the literature (Milroy and Milroy, 1997). These results encourage us to continue in this direction and to explore new directions to study the influence of social structure on the process of diffusion of linguistic innovation.

Conflict of Interest Disclosure

The authors do not have any conflicts of interest to report.

Funding

Funding source: LabEx ASLAN (ANR-10-LABX-0081) of the Université de Lyon.

Research Ethics Statement

The whole project including design, data collection and processing, data handling, storing and sharing, privacy protection were screened and approved by the ethics committee of INRIA (National Institute for Research in Digital Science and Technology) (favorable opinion, reference 2017-005, IRB00013144).

Authorship Details

Louise Tarrade: research concept and design, collection and/or assembly of data, data analysis and interpretation, writing the article, critical revision of the article, final approval of the article.

Jean-Philippe Magué: research concept and design, collection and/or assembly of data, data analysis and interpretation, critical revision of the article, final approval of the article.

Jean-Pierre Chevrot: research concept and design, collection and/or assembly of data, data analysis and interpretation, critical revision of the article, final approval of the article.

References

Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P., & Fleury, E. (2018). Socioeconomic dependencies of linguistic patterns in Twitter: A multivariate analysis. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (pp. 1125–1134). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186011

Ali, I. (2019). Personality traits, individual innovativeness and satisfaction with life. *Journal of Innovation & Knowledge, 4*(1), 38–46. https://doi.org/10.1016/j.jik.2017.11.002

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135–160. https://doi.org/10.1111/josl.12080

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.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Chambers, J. K. (2013). Patterns of variation including change. In J. K. Chambers & N. Schilling (Eds.), *The handbook of language variation and change* (pp. 129–297). Wiley Blackwell.

Costin-Gabriel, C., & Rebedea, T. E. (2014). Archaisms and neologisms identification in texts. In: *2014 RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference* (pp. 1–6). IEEE. https://doi.org/10.1109/RoEduNet-RENAM.2014.6955312

Del Tredici, M., & Fernández, R. (2018). The road to success: Assessing the fate of linguistic innovations in online communities. *ArXiv:1806.05838* [cs.CL]. https://doi.org/10.48550/arXiv.1806.05838

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv:1810.04805* [cs.CL]. https://doi.org/10.48550/arXiv.1810.04805

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS One, 9*(11), e113114. https://doi.org/10.1371/journal.pone.0113114

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

Feltgen, Q., Fagard, B., & Nadal, J.-P. (2017). Frequency patterns of semantic change: Corpus-based evidence of a near-critical dynamics in language

change. *Royal Society Open Science, 4*(11), 170830. https://doi.org/10.1098/rsos.170830

Flekova, L., Preoţiuc-Pietro, D., & Ungar, L. (2016). Exploring stylistic variation with age and income on Twitter. In: K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 313–319). Association for Computational Linguistics

Friggeri, A., Lambiotte, R., Kosinski, M., & Fleury, E. (2012). Psychological aspects of social communities. In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (pp. 195–202). IEEE. https://doi.org/10.1109/SocialCom-PASSAT.2012.104

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

Kerremans, D., & Prokić, J. (2018). Mining the web for new words: Semi-automatic neologism identification with the NeoCrawler. *Anglia, 136*(2), 239–268. https://doi.org/10.1515/ang-2018-0032

Kershaw, D., Rowe, M., & Stacey, P. (2016). Towards modelling language innovation acceptance in online social networks. In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 553–562). Association for Computing Machinery.

Labov, W. (2010). *Principles of linguistic change. 2: Social factors*. Wiley-Blackwell.

Laitinen, M., Fatemi, M., & Lundberg, J. (2020). Size matters: Digital social networks and language change. *Frontiers in Artificial Intelligence, 3*, 46. https://doi.org/10.3389/frai.2020.00046

Leskovec, J., & Sosič, R. (2016). SNAP: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology, 8*(1), 1–20. https://doi.org/10.1145/2898361

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3111–3119). Curran Associates, Inc.

Milroy, J., & Milroy, L. (1997). Network structure and linguistic change. In: N. Coupland & A. Jaworski (Eds.), *Sociolinguistics* (pp. 199–211). Springer.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics, 42*(3), 537-593. https://doi.org/10.1162/COLI_a_00258

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

Schlechtweg, D., Hätty, A., Del Tredici, M., & im Walde, S. (2019). A wind of change: Detecting and evaluating lexical semantic change across times and domains. ArXiv, 1906.02979 [cs.CL]. https://doi.org/10.48550/arXiv.1906.02979

Stewart, I., & Eisenstein, J. (2018). Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. *ArXiv:1709.00345* [cs.CL]. https://doi.org/10.48550/arXiv.1709.00345

Tjong Kim Sang, E. (2016). Finding rising and falling words. In: E. Hinrichs, M. Hinrichs, & T. Trippel (Eds.), *Proceedings of the workshop on language technology resources and tools for digital humanities* (LT4DH) (pp. 2–9). The COLING 2016 Organizing Committee

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). Demographic inference and representative population estimates from multilingual social media data. In: L. Loiu & R. White (Eds.), *WWW' 19: The World Wide Web Conference* (pp. 2056–2067). https://doi.org/10.1145/3308558.3313684

Weinreich, U., Labov, W., & Herzog, M. (1968). *Empirical foundations for a theory of language change (Vol. 58)*. University of Texas Press Austin.