How position in the network determines the fate of lexical innovations on Twitter

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

PLOS COMPLEX SYSTEMS

RESEARCH ARTICLE

How position in the network determines the fate of lexical innovations on Twitter

Louise Tarrade[1]*, **Jean-Pierre Chevrot**[1,2]*, **Jean-Philippe Magué**[1,3]

1 ICAR laboratory (UMR 5191), École Normale Supérieure de Lyon, France, **2** LIDILEM laboratory (EA 609), Université Grenoble Alpes, France, **3** IXXI, Complex Systems Institute, Lyon, France

* louise.tarrade@ens-lyon.fr (LT); jean-pierre.chevrot@univ-grenoble-alpes.fr (J-PC)

Citation: Tarrade L, Chevrot J-P, Magué J-P (2024) How position in the network determines the fate of lexical innovations on Twitter. PLOS Complex Syst 1(1): e0000005. https://doi.org/10.1371/journal.pcsy.0000005

Editor: Jennifer Badham, Durham University, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: January 21, 2024

Accepted: July 9, 2024

Published: September 3, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcsy.0000005

Data Availability Statement: Data is available on the following repository Ortolang, that is a French government supported infrastructure for the

Abstract

This study analyzes the diffusion of lexical innovations on Twitter to understand how the social network position of adopters impacts their success. Looking at both successful and failed neologisms, we categorize them into "changes" which become established and "buzzes" which decline over time. Using a corpus of 650 million French tweets, we reconstruct user networks and characterize adopters of innovations during different diffusion phases based on prestige, centrality, clustering, and external ties. In the early innovation phase, change and buzz adopters have similar peripheral profiles. During propagation, changes spread to prestigious, central individuals while buzzes do not, which predicts their eventual success or failure. By the establishment phase, changes reach highly central users with closer external ties. The results align with sociolinguistic theories about weak ties for innovation and strong ties for establishment. Additionally, logistic regression models based on early adopter profiles can predict the fate of innovations. This work sheds light on the diffusion dynamics of online lexical innovations and the crucial role of user network factors.

Author summary

In everyday language, words are constantly being created, and these words either persist or disappear. Although this phenomenon has been the subject of much linguistic research, the factors which influence the fate of a new word remain largely unknown, partly because of the difficulty of recording spontaneous language use over time. Examining the varieties of language used on social media allows us to overcome these limitations. We collected over 650 million tweets written in French, covering several years of ordinary interactions between 2.5 million users. We also collected the network of social links between these users. We identified nearly 400 words that appeared in the corpus between 2012 and 2014, and tracked their diffusion over 5 years within the network of users. Some of these words lead to changes, while others generate only ephemeral buzz. By looking at the position in the network of users who adopt these innovations, we show that words adopted by users who are more central in their community and easily in contact with other communities become established in the language, and vice versa. Thus, the position in the network of speakers who adopt these words is enough to predict their fate.

language data. url: www.ortolang.fr/market/corpora/sosweet.

Funding: J.-P. M., J.-P. C. and L.T. are grateful to the ASLAN project (ANR-10-LABX-0081, https://aslan.universite-lyon.fr/) of the Université de Lyon for its financial support within the French program "Investments for the Future" operated by the National Research Agency (ANR). The data collection has been supported by the SoSweet ANR project (ANR-15-CE38-0011-03, https://anr.fr/) attributed to J.-P. M. and J.-P. C. The authors are also grateful to University of Grenoble Alpes and Ecole Normale Supérieure de Lyon for the support for publication. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Previous work

Since language evolves within a social context, its usage diversifies according to the heterogeneity and changes in society, and sociolinguistic variation is omnipresent. Different variants of the same form are constantly in competition at all levels of the linguistic structure. Every human being is able to vary his or her way of speaking or to opt for a particular variant depending on whom he or she is addressing, for what purpose and in what context, with a varying degree of consciousness. Variation is the phenomenon observed in synchrony, change is its outcome from a diachronic point of view: "all change (with the exception of certain lexical innovations) results from a situation of variation—but not all variation leads to change [our translation]" (p. 23) [1].

As theorised by Weinreich et al. [2], variationist sociolinguistics is mainly concerned with explaining the mechanisms of linguistic change and establishing the influence of linguistic, cognitive, cultural and social factors on change. While external pressure and the influence of ones social groups (e.g. class, race, gender) have been shown to be explanatory factors for variation, social ties between individuals are also an important parameter to take into account when looking at the dynamics of the circulation of change. Thus, in his survey in Philadelphia Labov [3,4] establishes a significant correlation, particularly for women, between the use of advanced forms of the sound changes in progress and the structure of the individual's network. Thus, the people leading the change are people with a certain local prestige, having both a high density of interaction in their local block, but also a large proportion of their friends living outside it. For their part, Milroy & Milroy [5,6] were particularly interested in the influence of network structures on the circulation of sociolinguistic variants. Significant results concerning the relation between linguistic change and network emerge from their study of Belfast. First, they confirm and complete Granovetter's contribution [7] on the importance of weak ties in the transmission of innovations by defining innovators as people with weak ties, peripheral to communities. The denser a network, and therefore the stronger its ties, the more conservative it is regarding the vernacular local norms and the more resistant it will be to change. In contrast, speakers with weaker and more peripheral ties will be less close to these norms and more exposed to external variants. The different variants thus pass from one linguistic community to another through peripheral individuals acting as bridges between the groups. However, according to Milroy & Milroy [5], the adoption of a variant by individuals who are both central and well-established in the community is essential for its establishment within the community. In addition, before central members adopt it, the variant must be transmitted through a large number of ties as it is less socially risky to accept an innovation that is already widely spread at the margins of the community.

While these studies have considerably highlighted the process of change circulation, they have also revealed a few limitations such as the limited number of speakers considered or the lack of continuous, homogeneous longitudinal data implying a synchronic approach to linguistic change—a process which, by nature, extends over time. Furthermore, sociolinguists historically favoured field surveys—inspired by the sociological approaches—often focusing on phonetic variables.

The diachronic study of linguistic change has thus long been left to the domain of historical linguistics which, by definition, is concerned with long-term changes, often spanning several centuries, and generally of a morphosyntactic nature. Moreover, the corpora on which it relies are written corpora often reflecting a language much more standardised than oral language. Emerging with the digital age, computational sociolinguistics [8], applied to social media, allows us to study less standardised varieties of language, which are highly propitious to variation and innovation, both synchronically and diachronically. The focus on media has

increased the amount of attention paid to the lexicon, and work on lexical variation and diffusion has flourished [9–18]. Observations of lexical changes are indeed more tractable on a shorter time scale, "the lexicon [being] the component where change is the quickest (new words are constantly being created), and grammar the most stable, change taking place over a long period of time [our translation]" [19]. Furthermore, one can assume that the acceleration and multiplicity of exchanges on social media induce a phenomenon pointed out by Lorenz-Spreen et al. [20], namely that the ever-faster dissemination and consumption of information leads to a decrease in the collective attention span given to it. Consequently, the ever-increasing mass of content can lead to an acceleration of the diffusion process of linguistic innovations, whose fate would also be sealed more quickly.

Computational sociolinguistics has leveraged on social interaction data to address the relationship between the diffusion of linguistic innovations and the network structure connecting individuals. Particular attention has been paid to the importance of weak ties in the introduction of innovation and strong ties in their establishment within the language community. For instance, the innovative nature of information transmitted via weak ties and the greater influence of strong ties has been confirmed by a large-scale study on the transmission of information on Facebook, involving 250 million users [21]. At the linguistic level, studies on a short time scale on Twitter and Reddit have shown that the innovators, the people who introduce new linguistic forms, are individuals who have many weak ties and who are more central to the network [14]. This is in line with both Milroy's definition of innovators [5] and Labov's definition of linguistic change leaders [3,4] in terms of their centrality. On the other hand, it has been shown that people with strong ties have more influence than others [12,14].

The belonging of individuals to an area of high density in their local network generally results in the maintenance of vernacular forms [22] and, in the same way, the more isolated a community is from others, the more its members converge linguistically [23]. On the other hand, it is likely, as Milroy & Milroy [5] suggest, that the adoption of an innovation by individuals strongly embedded in local groups facilitates the spread of the innovation through these more cohesive subgroups and the establishment of this innovation in the linguistic community more generally. Multi-agent simulations effectively showed that while the absence of solitary and very peripheral members in a network leads to a lack of innovation, the absence of people defined as leaders (highly connected agents) prevents variants from stabilizing as norms [24].

Other studies have examined the relationship between some structural properties of the network and the circulation of innovations. At the egocentric network level for instance, individuals with smaller networks are more linguistically malleable [25] and are therefore more likely to adopt a linguistic innovation. At the level of the network as a whole, the study of the diffusion of neologisms has showed that a larger network as well as dense connections within and between communities increase the number of new words as well as their chances of survival, in contrast to communities fragmented into many local clusters [26]. The diffusion of a neologism is also more likely to succeed if it is not limited to a few subgroups of speakers but rather spreads across different speaker communities [17].

As we have seen, variationist sociolinguistics has highlighted the fact that the position occupied by speakers in their community can play an important role in the diffusion of linguistic change. In brief, two main theories have emerged about individuals driving change in their local networks: one defining them as people with weak ties, peripheral to their community [5] and the other as people central to their community, but with many ties outside it [4]. As the starting point of linguistic change is complicated to identify, it is likely that these two descriptions simply refer to two different phases in the diffusion of linguistic change. Computational studies on this issue have relied mainly on social media corpora to examine the link between networks and the diffusion of change on a larger scale. In addition to the impact of certain

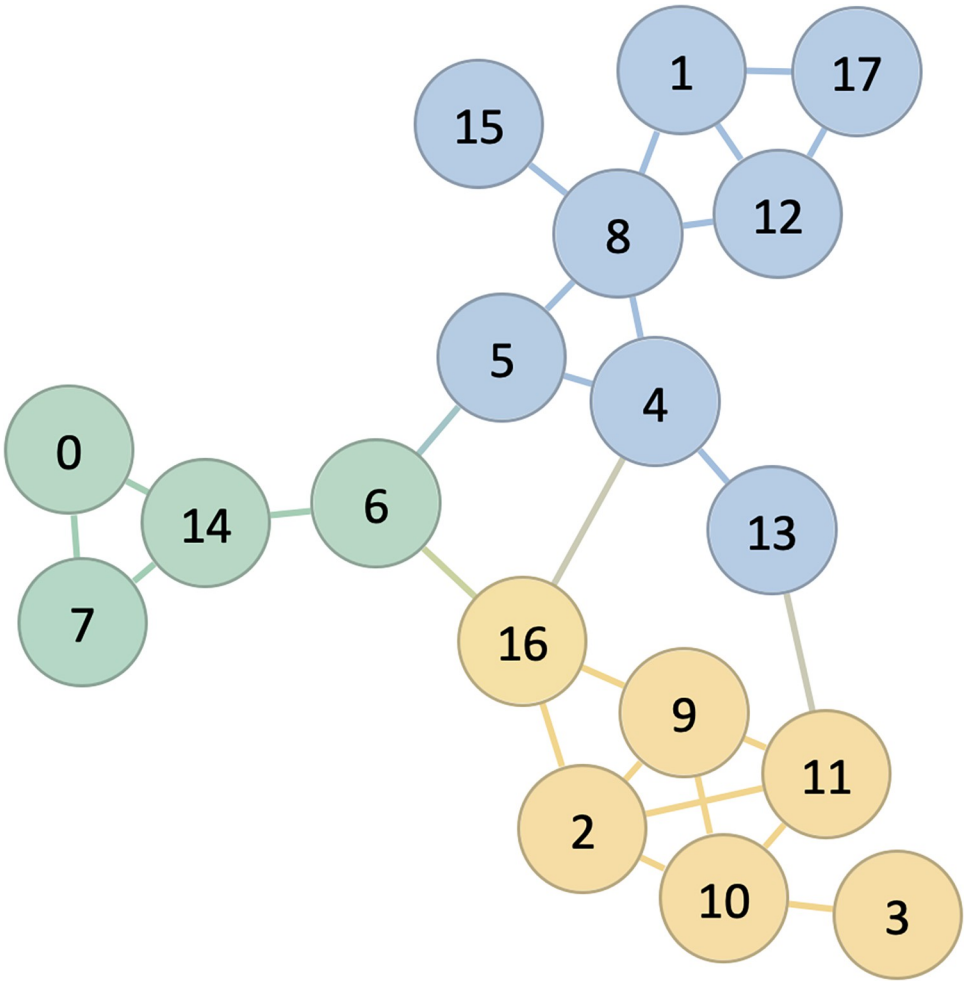

Fig 1. Social network formed by 18 individuals belonging to three different communities.

https://doi.org/10.1371/journal.pcsy.0000005.g001

structural properties of the network on the diffusion of linguistic innovations, they have mainly confirmed the role of the margin and weak ties in the introduction of innovations, as well as the influence of strong ties on their stabilization. They have also shown the conservative attitudes towards vernacular norms of more closely-knit groups. Fig 1 schematises a hypothetical toy social network formed by 18 individuals, each belonging to one of the three communities represented by the colours green, blue, and yellow. Speakers with a very closed network–such as those belonging to the triads 0-7-14 and 1-12-17, or the tetrad 2-9-10-11 –should therefore tend to be less innovative than others and intervene at a later stage of propagation. Conversely, individuals whose networks are smaller or who are located on the periphery of communities–such as nodes 6, 16 or 13—are more linguistically malleable, less conservative, and therefore more likely to take up innovations and, by extension, to facilitate their circulation. The role played by the centrality of innovators remains slightly unclear at this stage. The research carried out to date, which has focused almost exclusively on English, highlights the importance of links between individuals in the process of diffusion of linguistic innovations and sheds light on certain aspects in its own way, without however offering a complete overview of this phenomenon. Moreover, with a few exceptions, they have generally concentrated on successful innovations, leaving aside unsuccessful innovations.

Based on a corpus of tweets in French and a short diachronic observation of the diffusion of successful and unsuccessful lexical innovations from their appearance to their stabilization or decline, we will examine **a)** how the structural properties of their adopters within the social network evolve over time, and **b)** whether the position of the speakers who adopt them at the successive phases of their diffusion can predict the fate of the lexical innovations. Our contribution is to provide a global overview of the circulation of lexical innovations within a social network. Moreover, we work with data in French, a language that is rarely studied in this type of study, where English is over-represented.

Materials and methods

Corpus

For this work, we rely on a corpus of around 650 million tweets in French coming from about 2.5 million users, and spanning the period from 2007 to early 2019, the largest part of which is contained between March 2012 and January 2019. An initial collection of 170 million tweets produced between 2014 and 2017 was collected using the data providers Gnip and Datasift and constitutes the user base of this corpus [27]. The selection criteria for the tweets were that they should be written in French and come from the GMT and GMT+1 time zones. In a second phase the corpus was completed—directly via the Twitter API (using the Tweepy library)—by retrieving iteratively the latest tweets of the users having produced this initial corpus, excluding retweets. The corpus was filtered according to language and client used in order to keep only tweets in French and to eliminate as much as possible tweets from bots. For the language, we simply relied on the language of the tweet as automatically identified by twitter. For the bots, we relied on the Twitter clients. Since bots produce very stereotyped tweets, we have kept the clients exhibiting sufficient tweet lengths variability. The list of retained clients and the selection criteria are available at [28]. The corpus of tweets is available on the Ortolang platform [29].

Lexical innovations

As explained in [30], we first selected all the words (i.e. any sequence of alphanumeric characters that can contain an apostrophe or a hyphen) that appeared in the corpus for the first time between March 2012 and February 2014. For each of these words, we then reconstructed their usage trajectory over 5 years from their first appearance, by recovering their usage rate—i.e. the number of people who used this form out of the number of people who tweeted during the month.

For each of the trajectories obtained, we used a curve-fitting method using the LMFIT library for Python to fit them as closely as possible to two functions: the logistic function and the lognormal function. These functions correspond respectively to the ideal theoretical S-shaped trajectory of successful innovations [31–34] and the skewed bell-shaped trajectory of innovations whose use, after a growth phase, declines rather than stabilizes. We then used the adjustment output parameters to retain the words whose trajectory of use over 5 years most closely obeyed one or other of these laws.

A manual filtering stage was then necessary to remove the named entities from the almost 500 words retained. In the end, we have two types of lexical innovation:

1. The changes correspond to lexical innovations whose monthly trajectory of use follows a (logistic) S-shaped curve. It is possible to identify three distinct phases in the diffusion of this type of innovation: an initial phase—the innovation phase—during which the usage rate of the word remains at a very low level for a few months, followed by a more or less

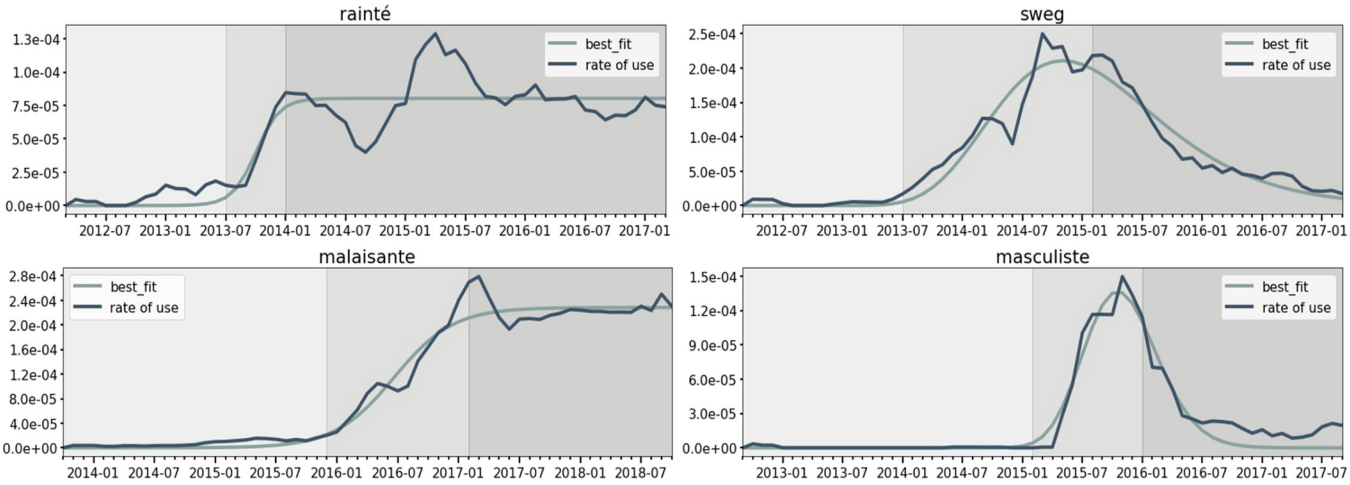

Fig 2. Two trajectories of lexical innovations. The usage rate per month of two changes (left) and two buzzes (right) represented by a rolling average with a three-month window (blue), as well as the result of the curve fitting (green). The three diffusion phases are represented by the grey shading in the background [30].

https://doi.org/10.1371/journal.pcsy.0000005.g002

long propagation phase during which its usage rate takes off exponentially, to finally stabilise in the fixation phase. We identified 141 changes.

2. The buzzes correspond to lexical innovations whose use trajectory per month follows a Gaussian curve. The first two phases of diffusion of innovations categorised as buzz are identical to those observed for changes. However, the last phase shows a significant decline in the rate of use of the word, until it returns to a very low rate; this is what we call the decline phase. The number of buzzes is 251.

To automatically delimit the three diffusion phases described above—innovation, propagation, then fixation for changes or decline for buzzes—we used the third derivative of the fitted distribution. More precisely, we looked for its maximums to identify the moments in the trajectory where the acceleration varies the most, delimiting the beginning and end of the propagation phase.

Fig 2 shows two changes ("rainté" and "malaisante") and two buzzes ("sweg" and "masculiste") identified with this method, their trajectory of use over 5 years, the adjustment to the reference function, and the three phases of diffusion.

The scripts used to detecting and categorizing the lexical innovations and the resulting data are available at [35].

Control words

In order to characterize the dynamics of lexical innovations in the network of users, we designed a third group of control words whose use is stable. The period and duration taken into account for the control words was matched with the lexical innovations, 5 years from February 2013 to January 2018.

We retrieve all the words of this period with at least 100 occurrences, as well as their number of users per month. Stable words are defined by a five-year usage rate whose standard deviation is below a certain threshold. In order to make this threshold comparable from one word to another, the monthly uses were normalized over the 5 years period. After manual observation of a large sample of words, this threshold was set at 0.007. In parallel, we check that each

Changes	selfie, mskn, putaclick, compwend, crimetime, bruuuuh, iencli, tfcn, padamalgam, cheatmeal, salpute, srapel, féminazis, unranked, qtv, malaisance, huuuuuurle, fullstack, obah, tsoulever
Buzzes	peufra, mddddddrr, sweg, lacelle, ololooo, fangirlage, ennuyance, oklmus, mvenere, batrd, caley, pttttttttdr, flippagram, miskinou, teamer, dhrs, ptddddddddddddddddddddr, partît, nptki, jgrail
Control words	bouges, tsk, boostez, mohhhh, été*, accélérateurs, vegetarien, soil, frnchmnt, cdes, rpas, mouffetard, jramène, blettes, 08:08, arthrite, limougeauds, bolas, aériennes, okaaaaay

Fig 3. Examples of changes, buzzes and control words.

https://doi.org/10.1371/journal.pcsy.0000005.g003

form has at least as many non-zero values as the linguistic innovation that has the least, in order to avoid words with to long periods with a zero use rate.

We obtain almost 40,000 words from which we randomly select 200 words whose number of users is matched to that of lexical innovations. Fig 3 shows a random sample of 20 forms belonging to each of the categories, change, buzz and control word.

User network

For each user of the corpus, we have retrieved the list of his followees, i.e. the people he follows. From this information, we reconstructed the static network restricted to the other users of our corpus. We did not rely on mentions to reconstruct the network of users in the corpus because this would have led to the exclusion of the vast majority of users who do not use mentioning. The resulting network counts 2.5 million users and 300 million ties.

From this network, we can then characterize each user according to the following network variables: local clustering coefficient, PageRank score, betweenness centrality and proximity to the outside of the community. The computations of the different network variables—except for the proximity to the outside of the community—were performed using the Python library NetworKit [36].

Clustering coefficient. The local clustering coefficient is the proportion of existing edges between the neighbours of a node among all possible edges. It is a measure whose values are between 0 and 1, and which therefore reflects the degree of openness of a user's network. A clustering coefficient of 0 means that the neighbours of user u have no ties with each other, while a clustering coefficient of 1 would mean that all its neighbours also have ties to each other. Thus, the higher a user's clustering coefficient, the closer his or her egocentric network is from a clique, i.e. a cohesive subgroup.

People belonging to dense sub-groups of the network with strong ties uniting their members will generally show more linguistic conservatism and be more resistant to change, and their adoption of an innovative variant is crucial to their maintenance within the community [6]. To demonstrate the relationship between maintaining vernacular norms and belonging to such a group, [22] have measured the strength of integration of nodes into their local group. Other studies have instead mobilised the notion of strength of ties—measured either by

remaining as close as possible to its initial definition [7,21], or by inferring it from the interconnection of nodes [12,14] -, generally to highlight the stronger influence of strong ties. Nevertheless, the strength of ties, network density and overlap of egocentric networks are very closely interconnected concepts. In a network, dense sub-groups with strong links between their members generally go hand in hand with overlapping egocentric networks [5]. The local clustering coefficient therefore seemed to us to be an easier measure to implement on a large network such as ours, and one that indicates, to a certain measure, of whether an individual belongs to a closely linked sub-group.

In this way, user with a very closed network is similar to an individual with strong ties, evolving within a more closed sub-group, and therefore less exposed to innovations coming from outside.

PageRank score. The PageRank score of a user u is a measure of the prestige of an individual. This measure depends both on the number of incoming ties of u, but also on whether these incoming ties themselves have a high PageRank score. That is to say, a user followed by many people, who are themselves followed by a large number of people, will a priori have a higher PageRank score than a user followed by a larger number of people, but who are themselves followed by very few people.

Applied to our network of Twitter users, we consider this measure to reflect a user's overall popularity level. This measure of popularity can to some extent be transposed, on a much larger scale, to the notion of prestige as used by Labov in his description of the leaders of linguistic change in Philadelphia [3,4]. In addition, the higher a user's PageRank score, the more likely it is that the content they produce will be exposed to a greater number of people.

Centrality measure. The measure of centrality for a user here corresponds to their centrality within the community to which they belong. The more central an individual is to his community, the more he acts as a "bridge" between its members. To calculate this score, it was therefore first necessary to detect the communities within our user network. To do this, we used the parallel implementation of the Louvain method [37] proposed by NetworKit, which allows us to identify the most densely connected groups in the network. As this method is non-overlapping, it implies that a user can only belong to one community. This shows that the great majority of the network's users belong to large communities, most of which have hundreds of thousands of individuals.

Betweenness centrality defines the centrality of a node as the number of times it is on the shortest path between two other nodes in the network. As the complexity of its computation increases strongly with the size of the network, we use approximate centrality measures for communities with more than 10,000 nodes, and exact centrality for the remaining, smaller communities. We use for this the parallel implementation of the KADABRA algorithm [38,39] provided by NetworKit. For each community, we calculate the centrality measures of its users by considering the network as an undirected graph. Since the centrality scores obtained in this way depends on the size of the community, they are not comparable from one community to another. For this reason, for each community, the set of centrality values obtained for each of its users has been standardised so that the median of this set is equal to 0 and the interquartile range (IQR = Q3—Q1) to 1. The scaled centrality measures can then be compared between users from different communities. It should be noted that we observe a slight correlation between the centrality measures thus obtained and the PageRank scores (Spearman correlation: 0.59).

While [14] have explored several measures of centrality to define the importance of a node in their social network, we will focus exclusively on betweenness centrality. In addition to the fact that the size of our network—more than 300 million ties—does not reasonably allow us to calculate all possible network measures, we believe that this measure is the one that comes

closest to centrality in Labov's sense. For Labov, the notion of centrality refers to important people in their local community, who are often mentioned by the other inhabitants of the block, and who are strongly involved in local life [4]. These people therefore act as a bridge within their local community, which is what betweenness centrality allows us to measure at the scale of the communities in our network.

Proximity to the community outside. We designed the last network variable, that indicates how fast a user is able to get in touch with a different community than his own. More precisely, from each node in the network, 10,000 random walks are performed, and for each of them we keep the number of steps that it was necessary to take before arriving in another community. The average of these 10,000 values thus obtained constitutes the final score attributed to the user for this variable.

The smaller the average number of steps of a user, the more directly he is in contact with another community. However, if he is located close to another community, this does not mean that he is more isolated in his own. The same user can have a central position within his community, but still have quick connections with people outside the community. We also observe a Spearman correlation of only -0.14 between these two variables.

This measure of proximity to the community outside is intended to reflect in part the profile of innovators described by Milroy [5], who are likely to bring innovations to their community through more direct ties with other communities.

Each of the users in the corpus is therefore characterized according to this set of four network variables giving information about the degree of openness of their egocentric network, their relative prestige, their centrality within their community, and their proximity to the outside of the community.

Comparison of the distributions of the different network variables at the three diffusion phases and prediction

Characterisation of words. Contrary to what was previously initiated in [30], we do not aggregate all the users who have used a word of a given category (e.g. buzz) at a given phase of diffusion, but each word is characterized independently. We take the view that although the set of words making up a category of lexical innovation (buzz or change) follows a global dynamic, each word nevertheless has its own dynamics. Users of innovations such as morphological derivations may not be exactly the same as users of phonetic spellings or lengthenings. Analyzing the distributions for each variable at the word level rather than aggregating users by innovation type allows us to avoid overlooking the different dynamics that may exist within the same category of innovations.

For each diffusion phase—innovation, propagation, fixation or decline—and for each network variable, we characterize each buzz and change in the following way: for each word w and each network variable v, we retrieve the months corresponding to the diffusion phase p considered. We then retrieve all users u who adopted w for the first time during the period covered by p. Then, for each of these adopters, we retrieve the value of v that corresponds to it. At this stage, we have a set of values of v, corresponding to those of all the users who adopted w in phase p. The value of v that will be attributed to w will then be the median of this set, as the distribution of the different network variables does not follow a normal distribution. More formally, the value of a network variable v associated with a word w at phase p can be noted:

$$v_{w,p} = Med(v_{u_1}, v_{u_2}, \ldots v_{u_n})$$

Finally, each of the words in each phase is represented by a four-dimensional vector corresponding to the clustering coefficient, the PageRank score, the centrality, and the average number of steps to exit the community.

For control words, the same procedure is used but without distinguishing the different phases of diffusion.

Univariate tests. One of our goals is to characterize the actors of change. This is addressed by comparing phase by phase the distribution of the network variables of the three groups: changes, buzzes and control words. To check the significance of our observations, we use non-parametric tests, given the non-normality of the distributions. More precisely, we use the Kruskal-Wallis test which tests the null hypothesis that the population median of all groups is equal, and then as a post-hoc test the Dunn's test which allows us to compare each pair of distributions. We applied the Bonferroni adjustment to the Dunn's test to correct the significance level. In both cases, we set the significance threshold to $p<0.05$.

Predicting the fate of lexical innovations. We then tried to predict the fate of lexical innovations before their trajectory stabilizes or declines, i.e. as early as the innovation or propagation phase. To do this, we train a logistic regression model using the scikit-learn library on all the lexical innovations in our dataset—i.e. the 141 changes and 251 buzzes. This involves training a model for binary classification: the variable to be predicted is the type of lexical innovation: buzz *vs* change. The explanatory variables are the median values of the set of adopters of each word for each network variable. A first prediction is made with the data characterizing each word in the innovation phase, and a second with the data from the propagation phase.

To ensure that the model results are not biased by the greater number of buzzes than changes, the dataset is reduced to balanced classes by randomly selecting as many buzzes as there are changes. The data is also standardized before training the model, so that all medians are 0 and the IQR is 1. It is then split into training and test data representing 75% and 25% of the data respectively—this represents a training set of 211 items for a test set of 71 items. Given the small number of inputs and the fact that only 60% of the buzzes is considered, we train 10,000 models in this way varying the buzzes. Thus, in the training phase, the changes will always be the same, but the buzzes will vary systematically.

We then evaluate the quality of the prediction on the data in the innovation phase, and then in the propagation phase, by retrieving for each of the 10,000 models the following evaluation metrics: the area under the ROC curve (now AUC), the precision, and the confusion matrices. An AUC score lies between 0 and 1. If it is 0.5, it means that the model predicts as well as the hazard. The precision, also between 0 and 1, corresponds to the average rate of correct predictions. Finally, the confusion matrices give the distribution of true and false positives and true and false negatives. More precisely, we will carry out a Fisher test on each of the matrices obtained to ensure that this distribution is not due to hazard.

The scripts used to calculate the network variables, characterize the words, create the group of control words, and perform the univariate tests and predictions are available at [28].

Results

We will first ask whether and how the network characteristics of the individuals who adopt lexical innovations differ from those of the users of the control words composing our control group at the different phases of diffusion. At the same time, we will extend this questioning to the level of lexical innovations and ask which network characteristics are the most discriminating between changes and buzzes, always considering the timing of their diffusion. Secondly, we will try to find out whether it is possible to predict the fate of lexical innovations simply based on the four network characteristics of their adopters, described in the previous section.

Comparison of distributions

The figure below (Fig 4) shows the different distributions of median values used to characterize each word by type, by network variables and by phase of diffusion; each point thus represents a word, and each distribution a word category. Lexical innovations are shown in blue and green, representing changes and buzzes respectively, and control words in yellow. The distribution of the latter does not vary from one phase to another, since we cannot distinguish between different phases.

The results of the univariate tests performed on each set and each pair of distributions are presented in Fig 5, which should therefore be systematically compared with the distributions commented in Fig 4. Non-significant results are indicated by a hatched background. A yellow

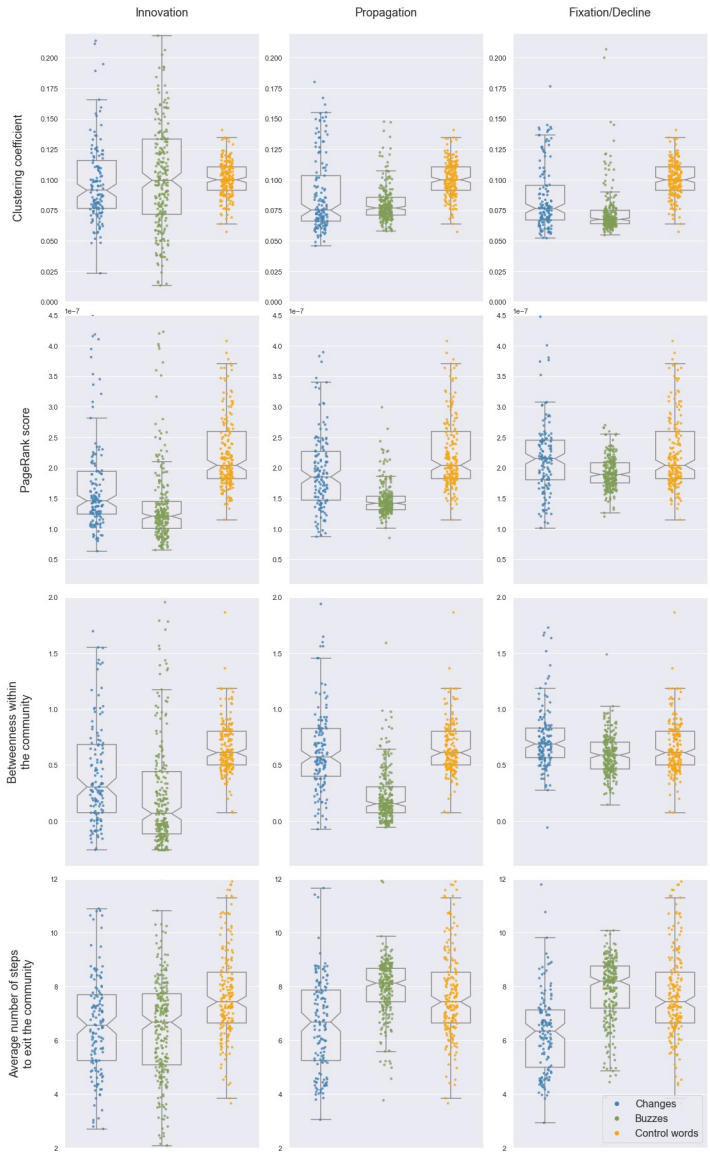

Fig 4. Distributions of median values. Distributions of median values characterizing each word by type (in blue the changes, in green the buzzes, and in yellow the control words), by network variable (rows) and by diffusion phase (columns).

https://doi.org/10.1371/journal.pcsy.0000005.g004

		INNOVATION			PROPAGATION			FIXATION/DECLINE		
	A - B	Control words - Changes	Control words - Buzzes	Changes - Buzzes	Control words - Changes	Control words - Buzzes	Changes - Buzzes	Control words - Changes	Control words - Buzzes	Changes - Buzzes
CLUSTERING COEFFICIENT	Kruskal-Wallis p-value				1,08E-28			6,85E-44		
	Dunn's p-value				6,94E-17	3,50E-26		3,60E-14	2,15E-44	1,29E-05
PAGERANK SCORE	Kruskal-Wallis p-value	1,47E-45			1,25E-51			9,87E-09		
	Dunn's p-value	2,02E-14	4,21E-46	5,59E-06	2,29E-06	6,01E-51	1,13E-16		1,40E-07	1,20E-05
BETWEENNESS WITHIN THE COMMUNITY	Kruskal-Wallis p-value	2,23E-31			1,05E-56			2,66E-07		
	Dunn's p-value	5,62E-09	4,75E-32	2,99E-05		1,47E-49	4,69E-30	2,75E-02	8,02E-03	1,72E-07
AVERAGE NUMBER OF STEPS TO EXIT THE COMMUNITY	Kruskal-Wallis p-value	2,72E-10			4,43E-16			1,30E-23		
	Dunn's p-value	1,12E-06	1,88E-09		6,64E-06	3,88E-04	1,43E-16	3,30E-12	6,27E-03	5,49E-24

Not significant A > B A < B

Fig 5. P-values obtained from the different univariate tests.

https://doi.org/10.1371/journal.pcsy.0000005.g005

background indicates that the values in distribution A (top) are globally higher than the values of distribution B (bottom); a green background indicates the opposite. For example, the p-value obtained with Dunn's test for the centrality of adopters in the fixation phase is 0.0275 and is therefore significant since it is lower than the significance threshold set at 0.05. The green background means that lower centrality values are more often observed for users of control words than for users who adopted a change in the fixation phase.

A first element that can be noted is that the correlation observed between the PageRank scores and user centrality measures emerges particularly well when we look at the graph of distributions, as their dynamics are similar for each category over the diffusion phases. While these variables may seem redundant from this point of view, a Spearman correlation of 0.62 for all phases considered (Fig 6) indicates a positive but moderate correlation. Indeed, it is quite possible to have high prestige but low centrality, as is the case for node 16 in the Fig 1, given that the centrality of a user is calculated in relation to the community to which he belongs. In the same way, a user can be not very central to his community while being very isolated from other communities, like nodes 1, 3 or 17 in Fig 1, or conversely be not very central but almost immediately in contact with other communities, like nodes 5 or 13 for example.

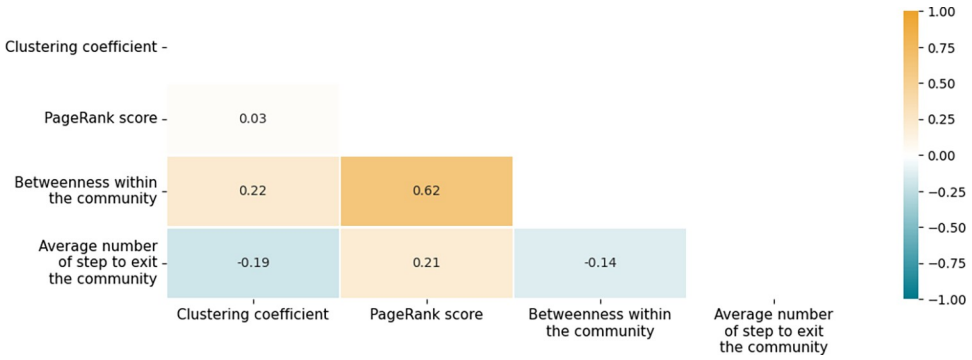

Fig 6. Spearman correlations. Spearman correlations between the different variables—all phases and all types of words considered.

https://doi.org/10.1371/journal.pcsy.0000005.g006

Finally, a user can also be central to his or her community while, on average, being in contact with other communities relatively quickly—or not (node 10)—and have a relatively open (node 8) or closed (node 9) egocentric network.

In the innovation phase, we do not observe significant differences between the distributions of the clustering coefficient. In the propagation and fixation phase, however, a distinction is observed, lexical innovations having lower clustering coefficient than control words. No evolution is observed between these two phases. Thus, the first adopters of lexical innovations do not differ in the degree of openness of their own network.

If the PageRank scores of lexical innovations are significantly lower than those of control words during the first two phases of diffusion, this difference decreases during the fixation phase regarding changes. Users who adopt lexical innovations during the first two phases of diffusion are less prestigious than normal, particularly regarding the buzzes, whose values remain in the same range from one phase to the next, whereas those of the changes gradually approach those of the control words until they reach their level in the fixation phase. It should be noted that although the buzz adopters have significantly lower PageRank scores than the other two categories in the fixation phase, they are nevertheless higher than those observed in the two previous phases.

While lexical innovations have significantly lower centrality measures than control words in the innovation phase, in the propagation phase the changes stand out from the buzzes by reaching users as central as those of the control words—no significant difference being observed between these two distributions -, while the distribution of buzz adopters remains significantly lower. While the latter, like the PageRank scores, rises in the fixation phase, it remains slightly lower than the other two. The distribution of centrality measures for change adopters is even higher than that of control words. Thus, from the propagation phase onwards, changes, unlike buzz, are adopted by more central users, which would a priori facilitate their diffusion within the community.

In the innovation phase, the distributions of the average number of steps of the lexical innovations are lower than those of the control words, while not being distinguished from each other. The lexical innovations are therefore initially adopted by users who can generally reach outside their community more quickly, which facilitates their subsequent dissemination. Indeed, when we look at the distribution of these values in the propagation phase, the distribution of changes has not really changed, whereas the distribution of buzzes increases significantly, until it is positioned at a higher level than that of the control words. While the position of the distributions remains almost identical in the fixation phase, that of the changes is concentrated around lower values.

What emerges from these observations is that the first adopters of both successful changes and unsuccessful buzzes have similar network profiles. These innovators tend to be less prestigious and more peripheral compared to average users. This effect is even more pronounced for buzzes. Innovators can also reach outside their communities more easily. This likely helps facilitate the future diffusion of these new terms. This similarity fades in the propagation phase, where changes succeed in reaching much more prestigious and central users than buzzes, while maintaining a rapid proximity to the outside of the community, which should facilitate their circulation within the community but also outside it. Buzzes, on the other hand, continue to spread, but do not manage to reach more central or prestigious individuals, on the contrary. Moreover, as they are adopted at this phase by users who are less directly connected to different communities, the circulation between them will probably be obstructed later. The fixation phase confirms the dynamics of the changes, which are therefore adopted by people who are as prestigious as the users of control words, slightly more central, but also with an even more direct proximity to the outside of the community than in the propagation phase.

While the distribution of prestige and centrality values of adopters during this phase tends to realign with those of changes and control words during their decline phase, buzzes continue to be adopted by less central and less prestigious people, and with a more laborious contact with the outside of their community. Finally, if the distribution of clustering coefficients is discriminating between lexical innovations and control words, the fact of being adopted by users with a more open network is characteristic of innovations in the last two phases of diffusion.

Prediction of the fate of lexical innovations. We can now wonder whether these differences we observe between the distributions of median values of adopters of lexical innovations are sufficiently discriminating to allow us to predict, in the innovation or propagation phases, whether a lexical innovation will maintain in the linguistic community, and become a change or, on the contrary, whether its use will eventually decline, thus becoming a buzz.

Fig 7 shows the results obtained for the precision of the 10,000 prediction models trained by logistic regression, first on the values attributed to changes and buzzes in the innovation phase, in green, and then on those in the propagation phase, in blue. Fig 8 allows us to visualize the results of the AUC scores in the same way. Prediction made from the innovation phase are imprecise, with an average precision of 0.56 and an average AUC score of 0.61. If in general the models do slightly better than chance, these scores show that it is not possible to predict the fate of lexical innovations in the innovation phase.

If we look at these scores more closely with the confusion matrices resulting from these models trained on the innovation phase data, we can see that, in general, buzzes are slightly easier to predict than changes at this stage, with an average of 61% of buzzes correctly predicted (true positives) versus 52% of changes (true negatives). However, Fisher's exact tests on these matrices provide p-values greater than 0.05 in almost 80% of cases, which means that when we observe imbalances in the distribution of true/false positives and negatives, these are mostly non-significant. On average, these p-values are around 0.34. However, for the confusion matrices resulting from the models trained using the propagation phase data, this average p-value of the Fisher exact tests is now 8.3e-05 and only 0.03% of the observed ratios between percentages of true/false positives and negatives are non-significant. At this stage, buzzes still seem to be slightly easier to predict than changes, with an average of 83% of buzzes correctly predicted compared to about 79% for changes.

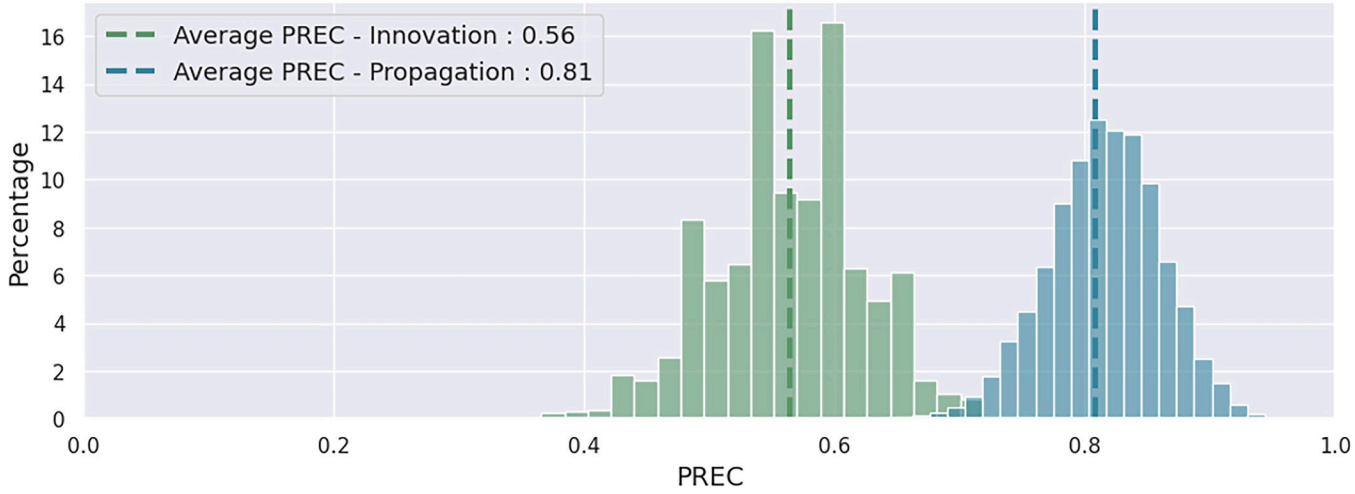

Fig 7. Precision. Precision obtained by the logistic regression models trained on the 10,000 datasets in the innovation phase (green) and in the propagation phase (blue).

https://doi.org/10.1371/journal.pcsy.0000005.g007

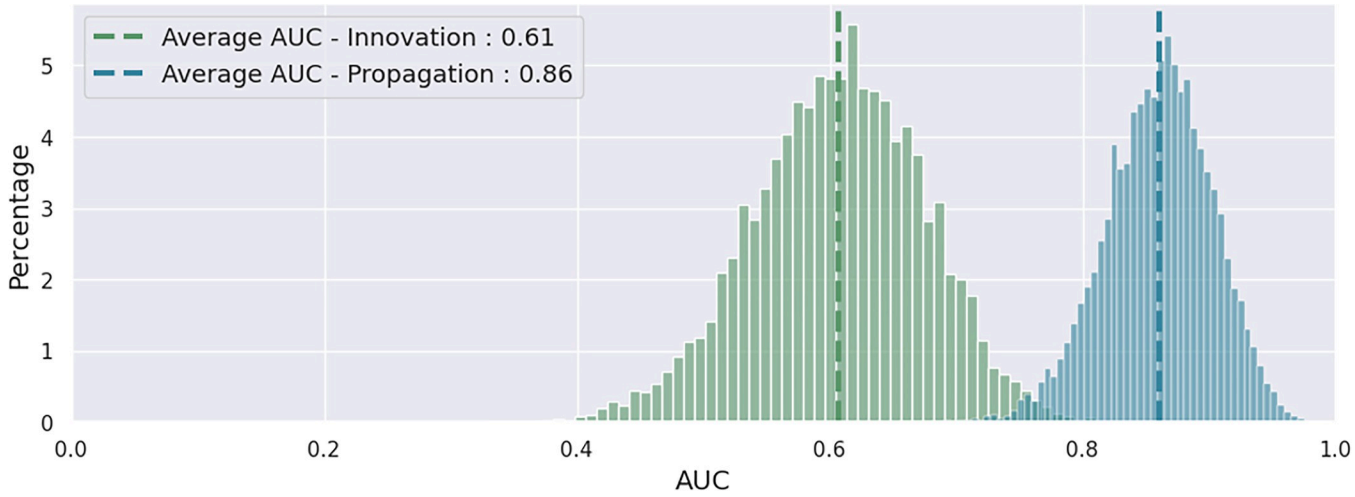

Fig 8. AUC scores. AUC scores obtained by the logistic regression models trained on the 10,000 data sets in the innovation phase (green) and in the propagation phase (blue).

https://doi.org/10.1371/journal.pcsy.0000005.g008

This significant improvement in prediction quality when using the propagation phase data is confirmed by an improvement of precision from 0.56 to 0.81, and an average AUC score of 0.86, which confirms that the classification of lexical innovations as buzz or change at this stage leaves little to chance.

In summary, it would appear that despite the significant differences in the positioning of the buzz and change distributions observed in the innovation phase for PageRank score and centrality measures, it is not possible at this stage to predict what a lexical innovation will become in the future based only on the network characteristics of its first adopters. However, when we rely on the network characteristics of the adopters of innovations in the propagation phase, it becomes quite possible to predict their fate. It is the position in the community and the more or less direct link with the outside of the adopters of an innovation at this stage that seem to seal their fate and favor (or not) their future stabilization.

Discussion

In this study, we examined the network characteristics of the users of our corpus and, in particular, whether it was possible to identify a 'typical profile' of adopters of lexical innovations at each of their diffusion phases. We also wondered whether this profile is different according to the type of lexical innovation, i.e. whether adopters of changes differ from adopters of buzzes. In this way, we seek to highlight the process of diffusion of lexical innovations and the factors in terms of network structure that contribute to their success or failure in a linguistic community, and to determine whether these large-scale results are consistent with or different from those obtained by the field surveys conducted in traditional variationist sociolinguistics, notably by Lesley and James Milroy as well as by William Labov.

First, we established that the initial adopters of lexical innovations are users with relatively similar network characteristics, regardless of whether these innovations later succeed or fail. Contrary to what we might expect, these individuals do not have a more open or closed personal network than average. They have the possibility to be in contact with other communities more quickly, without being central in their own community, nor prestigious within the global network. As such, these observations are largely transposable to those made by Milroy & Milroy [5] who define innovators as being more peripheral and having ties in several

communities. Although Milroy & Milroy [5] referred to local communities (different parts of Belfast), it is possible to transpose these observations to users during the innovation period. First, their position in the community is less central and therefore a priori more peripheral; and second, they maintain ties with at least two communities—on a much larger scale—that of social media, comprising several million users and that are defined not spatially but in relation to areas with a higher density of ties than in the rest of the network. Moreover, ties are inherently different from those maintained by the inhabitants of a city, for example.

While we did not characterize users in terms of strong- or weak-tied users, the local clustering coefficient as a measure of the degree of openness of a user's network captures to some extent a similar reality. In the propagation phase, clustering coefficients of change adopters are lower than average, which is consistent with the findings of previous work. However, we do not observe evidence that the changes were subsequently adopted by people who could be described as strongly connected, or at least belonging to a more closed subgroup, which would increase the likelihood that an innovation would spread in the community [14]. On the contrary, the degree of openness of adopters of changes seems to be higher as the diffusion of these changes progresses. However, nothing suggest that they were not taken up by a few individuals belonging to more closed subgroups, but not in sufficient numbers for this to be reflected in our results. Further studies on the strength of the ties between the users of our network and their degree of embeddedness would be desirable in order to be able to study in more detail the impact of this variable on the establishment of changes in the linguistic community.

While a profile of early adopters emerges in the first diffusion phase for lexical innovations, it is not yet possible at this stage to know whether they will become buzzes or changes, as our low prediction results in the innovation phase indicate. However, in the propagation phase, i.e. when the rate of adoption of buzzes and changes increases exponentially, we can identify a characteristic profile for users who adopt changes or buzzes. The success or failure of an innovation seems to depend on the combination of several factors. On the one hand, changes are adopted by individuals with a PageRank score that is always lower than normal, but much higher than those of buzzes. We can suggest that the adoption of lexical innovations by individuals with very low visibility implies that buzzes have a much lower frequency of exposure than changes at this stage. Repeated exposure to a term can in some cases have a significant effect on its adoption [40]. In addition, it appears that the first phase of diffusion of changes has an average duration of 18.5 months compared with 6.5 months for buzzes, i.e. almost three times longer. Changes therefore generally remain in circulation longer before entering their growth phase, which also increases the chances of being exposed to them. Thus, the higher exposure of future changes, being longer in circulation and adopted by people whose tweets are more likely to be made visible to a larger number of users, surely increases the likelihood of some changes being adopted in the future.

Next, the changes are characterized by adopters who are relatively central to their community, or at least as central as those in our control group, and located closer to other communities, whereas the opposite pattern emerges from the future buzzes. Indeed, the latter are characterized by adopters who are still very peripheral in their community and have much more distant ties to other communities. The fact that changes are adopted at this phase by users who are central to their community, acting as a bridge within it and thus facilitating their diffusion, but who also have a more direct proximity to individuals from other communities is directly in line with the observations made by Labov [4] in his Philadelphia survey when he describes the leaders of change. The adoption of innovations during the propagation phase by prestigious and central individuals, having direct ties outside the community, predicts their success. Meanwhile, innovations that do not spread to prestigious, central users tend to fail.

Our prediction results confirm that the profiles of early adopters influence the ultimate fate of new terms.

In the fixation phase, where the fate of lexical innovations is already sealed, the prestige of their adopters reaches that of our control group, when their centrality even exceeds it. Conversely, the average number of steps required to reach the outside of the community is even lower than in the previous phases. The observations of high measures of centrality within the community and immediate proximity to the outside of the community may be reminiscent of the conditions for adoption of an innovation described by Milroy & Milroy [5], i.e. for a variant to become established within a community, it is necessary that it has been adopted by people central to it, who themselves will only risk adopting the variant if it is already widely used at the margins of the community. That said, it should be noted that Milroy & Milroy [5] were describing adoption within a local community, whereas in our case we do not know whether the adoption of the innovation takes place within a single community or within the overall linguistic community of our corpus. It would also be interesting, when looking at the conditions for the success or failure of an innovation, to determine whether the fact that an innovation has succeeded in reaching several communities is a determining factor in the success of its diffusion, as Würschinger [17] finds for example. It is partly for this reason that it would be welcome in a future work to further develop the one started on communities, both by finely characterizing them, but also by observing the circulation of innovations within and between them.

One point to which we must turn our attention, and which has not been studied in this work, is the role played by the category of lexical innovation. The lexical innovations we have detected cover several categories and do not seem to be homogeneously distributed between buzzes and changes. While we find borrowings, morphological derivations, lengthenings, truncations, phonetic spellings, etc. in both types of innovation, it is immediately apparent that a greater number of lengthenings are observed in buzzes, for example, while more neologisms designating new realities or practices are present in changes. Words that have a greater communicative utility, that fill a semantic gap or that can also be used in spoken language are more likely to be maintained over time [41], as well as words used in a wider range of linguistic contexts [42]. The nature of the word itself therefore has a certain impact on its chances of survival and would be an interesting factor to consider in future research. Finally, it has been shown that other factors, notably demographic and geographical [11], play an important role in the diffusion of innovations. In future research, it would be interesting to consider all of these factors, both intra- and extra-linguistic, in order to refine and complete the results presented here on the impact of the position of speakers in the network on the diffusion of innovations.

To conclude, our study found similar general diffusion patterns for lexical innovations as previous sociolinguistic studies [4,5]. Those studies focused on phonetic innovations, localized communities, and surveys of hundreds. In contrast, our research examined lexical innovations at scale across millions of social media users. It should be noted, however, that it is easier for a speaker to act at the lexical level than at the phonological or morphosyntactic level, for example. This is because, once acquired, speakers generally do not change the way they pronounce, just as they are less likely to change their grammar. On the contrary, lexical variables are easier to manipulate, and are also more conscious and therefore more likely to be linked to identity issues. However, although they are less malleable, the other types of variables are not hermetic to change—even if this generally involves a longer time span. Thus, the question remains open as to whether the underlying mechanisms are the same and whether the influence of the network factors highlighted here can be generalized to non-lexical variables.

Acknowledgments

We gratefully acknowledge the support of the Centre Blaise Pascal's IT test platform at ENS de Lyon (Lyon, France) for the computing facilities. The platform operates the SIDUS solution [43] developed by Emmanuel Quemener.

Author Contributions

Conceptualization: Louise Tarrade, Jean-Pierre Chevrot, Jean-Philippe Magué.

Data curation: Louise Tarrade, Jean-Philippe Magué.

Funding acquisition: Jean-Pierre Chevrot, Jean-Philippe Magué.

Methodology: Louise Tarrade, Jean-Philippe Magué.

Project administration: Jean-Philippe Magué.

Supervision: Jean-Pierre Chevrot, Jean-Philippe Magué.

Validation: Jean-Pierre Chevrot, Jean-Philippe Magué.

Visualization: Louise Tarrade.

Writing – original draft: Louise Tarrade.

Writing – review & editing: Jean-Pierre Chevrot, Jean-Philippe Magué.

References

1. Marchello-Nizia C, Combettes B, Prévost S, Scheer T, editors. Grande Grammaire Historique du Français (GGHF). De Gruyter; 2020. https://doi.org/10.1515/9783110348194
2. Weinreich U, Labov W, Herzog MI. Empirical foundations for a theory of language change. WP Lehmann-Y Malkiel (Hrsgg), Directions for Historical Linguistics, Austin/London. 1968.
3. Labov W. The social origins of sound change. Locating Language in Time and Space. Academic Press New York; 1980. pp. 251–265.
4. Labov W. Principles of linguistic change. Vol. 2: Social factors. Digital print. Malden, Mass.: Blackwell; 2006.
5. Milroy J, Milroy L. Linguistic change, social network and speaker innovation. Journal of linguistics. 1985; 21: 339–384.
6. Milroy L. Language and social networks. 2nd ed. Oxford, UK; New York, NY, USA: B. Blackwell; 1987.
7. Granovetter MS. The Strength of Weak Ties. American Journal of Sociology. 1973; 78: 1360–1380. https://doi.org/10.1086/225469
8. Nguyen D, Doğruöz AS, Rosé CP, De Jong F. Computational Sociolinguistics: A Survey. Computational Linguistics. 2016; 42: 537–593. https://doi.org/10.1162/COLI_a_00258
9. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. Preis T, editor. PLoS ONE. 2013; 8: e73791. https://doi.org/10.1371/journal.pone.0073791 PMID: 24086296
10. Bamman D, Eisenstein J, Schnoebelen T. Gender identity and lexical variation in social media. J Sociolinguistics. 2014; 18: 135–160. https://doi.org/10.1111/josl.12080
11. Eisenstein J, O'Connor B, Smith NA, Xing EP. Diffusion of Lexical Change in Social Media. Berwick RC, editor. PLoS ONE. 2014; 9: e113114. https://doi.org/10.1371/journal.pone.0113114 PMID: 25409166
12. Goel R, Soni S, Goyal N, Paparrizos J, Wallach H, Diaz F, et al. The social dynamics of language change in online networks. International conference on social informatics. Springer; 2016. pp. 41–57.
13. Grieve J, Nini A, Guo D. Analyzing lexical emergence in Modern American English online. English Language and Linguistics. 2017; 21: 99–127. https://doi.org/10.1017/S1360674316000113
14. Del Tredici M, Fernández R. The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities. arXiv:180605838 [cs]. 2018 [cited 19 Nov 2020]. Available: http://arxiv.org/abs/1806.05838

15. Hovy D, Rahimi A, Baldwin T, Brooke J. Visualizing Regional Language Variation Across Europe on Twitter. In: Brunn SD, Kehrein R, editors. Handbook of the Changing World Language Map. Cham: Springer International Publishing; 2020. pp. 3719–3742. https://doi.org/10.1007/978-3-030-02438-3_175
16. Shoemark PJ. Discovering and analysing lexical variation in social media text. 2020.
17. Würschinger Q. Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter. Front Artif Intell. 2021; 4: 648583. https://doi.org/10.3389/frai.2021.648583 PMID: 34790894
18. Keidar D, Opedal A, Jin Z, Sachan M. Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang. 2022 [cited 24 Mar 2022]. https://doi.org/10.48550/ARXIV.2203.04651
19. Gadet F. Changement linguistique: Langage et société. 2021;Hors série: 41–46. https://doi.org/10.3917/ls.hs01.0042
20. Lorenz-Spreen P, Mønsted BM, Hövel P, Lehmann S. Accelerating dynamics of collective attention. Nat Commun. 2019; 10: 1759. https://doi.org/10.1038/s41467-019-09311-w PMID: 30988286
21. Bakshy E, Rosenn I, Marlow C, Adamic L. The role of social networks in information diffusion. Proceedings of the 21st international conference on World Wide Web. 2012. pp. 519–528.
22. Dodsworth R, Benton RA. Social network cohesion and the retreat from Southern vowels in Raleigh. Language in Society. 2017; 46: 371.
23. Tamburrini N, Cinnirella M, Jansen VAA, Bryden J. Twitter users change word usage according to conversation-partner social identity. Social Networks. 2015; 40: 84–89. https://doi.org/10.1016/j.socnet.2014.07.004
24. Fagyal Z, Swarup S, Escobar AM, Gasser L, Lakkaraju K. Centers and peripheries: Network roles in language change. Lingua. 2010; 120: 2061–2079. https://doi.org/10.1016/j.lingua.2010.02.001
25. Lev-Ari S. Social network size can influence linguistic malleability and the propagation of linguistic change. Cognition. 2018; 176: 31–39. https://doi.org/10.1016/j.cognition.2018.03.003 PMID: 29544113
26. Zhu J, Jurgens D. The structure of online social networks modulates the rate of lexical change. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics; 2021. pp. 2201–2218. https://doi.org/10.18653/v1/2021.naacl-main.178
27. Abitbol JL, Karsai M, Magué J-P, Chevrot J-P, Fleury E. Socioeconomic Dependencies of Linguistic Patterns in Twitter: a Multivariate Analysis. Proceedings of the 2018 World Wide Web Conference on World Wide Web—WWW '18. Lyon, France: ACM Press; 2018. pp. 1125–1134. https://doi.org/10.1145/3178876.3186011
28. Tarrade L. Network factors and diffusion of linguistic innovations; 2024 [cited 2024 Jul 16]. figshare [internet]. Available from: https://figshare.com/articles/software/Network_factors_and_diffusion_of_linguistic_innovations/26310976/1
29. ICAR, DANTE Inria, LIDILEM, ALMANACH. SoSweet. 2024 [cited 2024 April 10]. ORTOLANG [internet]. Available from: https://hdl.handle.net/11403/sosweet/v1
30. Tarrade L, Magué J-P, Chevrot J-P. Detecting and categorising lexical innovations in a corpus of tweets. Psychology of Language and Communication. 2022; 26: 313–329. https://doi.org/10.2478/plc-2022-15
31. Blythe RA, Croft W. S-Curves And The Mechanisms Of Propagation In Language Change. Language. 2012; 88: 269–304.
32. Rogers EM. Diffusion of innovations. 5th ed. New York: Free Press; 2003.
33. Feltgen Q, Fagard B, Nadal J-P. Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change. R Soc open sci. 2017; 4: 170830. https://doi.org/10.1098/rsos.170830 PMID: 29291074
34. Chambers JK. Patterns of variation including change. The handbook of language variation and change. 2013; 297–324.
35. Tarrade L. Detection of lexical innovations; 2024 [cited 2024 Jul 16]. figshare [internet]. Available from: https://figshare.com/articles/software/lexical_innovation_detection/26310973/2
36. Staudt CL, Sazonovs A, Meyerhenke H. NetworKit: A tool suite for large-scale complex network analysis. Net Sci. 2016; 4: 508–530. https://doi.org/10.1017/nws.2016.20
37. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment. 2008; 2008: P10008.
38. Borassi M, Natale E. KADABRA is an ADaptive Algorithm for Betweenness via Random Approximation. 2016; 18 pages. https://doi.org/10.4230/LIPICS.ESA.2016.20

39. van der Grinten A, Angriman E, Meyerhenke H. Parallel Adaptive Sampling with almost no Synchronization. arXiv; 2019. Available: http://arxiv.org/abs/1903.09422

40. Romero DM, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. Proceedings of the 20th international conference on World wide web. 2011. pp. 695–704.

41. Grieve J. Natural selection in the modern English lexicon. 2018. pp. 153–157.

42. Stewart I, Eisenstein J. Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. arXiv:170900345 [physics]. 2018 [cited 8 Dec 2021]. Available: http://arxiv.org/abs/1709.00345

43. Quemener E, Corvellec M. SIDUS—the solution for extreme deduplication of an operating system. Linux J. 2013;2013: 3:3.