The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities

Del Tredici, M.; Fernández, R.

Publication date
2018
Document Version
Final published version
Published in
The 27th International Conference on Computational Linguistics

The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities

Marco Del Tredici and **Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam
{m.deltredici|raquel.fernandez}@uva.nl

Abstract

We investigate the birth and diffusion of lexical innovations in a large dataset of online social communities. We build on sociolinguistic theories and focus on the relation between the spread of a novel term and the social role of the individuals who use it, uncovering characteristics of innovators and adopters. Finally, we perform a prediction task that allows us to anticipate whether an innovation will successfully spread within a community.

1 Introduction

Language is an incessantly evolving system, and linguistic innovations of different kind are continuously created. New variants originate in concrete communicative situations with particular speakers. After being introduced, some of them are adopted by other speakers belonging to the same community and possibly spread until they become community norms. In contrast, other innovations do not manage to make their way into the community conventions and just disappear after a certain period of time. This process raises many intriguing questions, which have been the focus of attention in Sociolinguistics since the late 1960's (Weinreich et al., 1968; Chambers and Schilling, 2013). For example, at what linguistic levels (phonology, lexicon, syntax) do innovations arise and succeed? Who are the leaders of change and what are their social characteristics? In this paper, we investigate lexical innovation in online communities, focusing on the latter type of question, by means of a large-scale data-driven approach.

We study the interplay between the birth and spread of new terms and users' social standing in large online social communities, taking as starting point hypotheses put forward in the Sociolinguistics literature (see Section 2). We present a longitudinal study on a large social media dataset, including 20 online forums and around 10 million users overall. We consider each forum as an independent social network, and investigate how lexical innovations arise and spread within a forum. We analyse the fate of approximately 8 thousand innovations (focusing on acronyms such as *lol*, phonetic spellings such as *plis*, and other linguistic phenomena usually collected under the term of *Internet slang*), and relate their spread within a social network to the role of the individuals who use them. We characterise users' roles within a community by means of a novel, theoretically-motivated *tie-strength* measure, combined with well-known centrality measures.

We show that (1) innovators (users who introduce a new term) are central members of a community, connected to many other users but with relatively low tie-strength, and (2) strong-tie users (who belong to cliques or sub-groups within the community) effectively contribute to the dissemination of a new term. This pattern is surprisingly consistent across the 20 online communities under investigation. In addition, we show that, by solely using information on speakers' tie strength as predictor variable, we can anticipate whether an innovation will successfully spread within a community.

Our work yields new theoretical insights into the applicability of sociolinguistic theories to online settings. It also has practical significance for NLP systems encountering novel terms in social media: While new terms that do not succeed in becoming community norms may safely be treated as out-of-vocabulary words, a better understanding of the dissemination process of novel terms beyond their

Proceedings of the 27th International Conference on Computational Linguistics, pages 1591–1603
Santa Fe, New Mexico, USA, August 20-26, 2018.

frequency can help identify those innovations that a system should be able to deal with. Together with the in-depth analysis presented in the paper, we make available our dataset, which includes interactions scraped from 20 topic-based forums over a period of 4 to 8 years, as well as the code for preprocessing the data, extracting the dissemination trajectories of ∼8k internet slang terms, and computing the strength of the ties among users.[1]

2 Background

2.1 Sociolinguistic Theories

Our investigation of innovations in online social communities builds on Milroy's (1987) theory, according to which linguistic innovations propagate and become community norms when individuals with *strong ties* adopt them early on; however, it is usually individuals with only *weak ties* who are the actual *innovators*, i.e., those who introduce novel linguistic variants. Milroy's proposal is thus an application to Sociolinguistics of *The Strength of Weak Ties* theory put forward by Granovetter (1973) to explain the diffusion of innovations more broadly.

In Milroy's (1987) original study, the strength of a tie between two individuals is given by a combination of factors, such as the amount of time spent together, the level of intimacy, etc. As a result, strong ties are usually those connecting family members or close friends, while weak ties connect acquaintances. Individuals linked by strong ties form close-knit sub-communities or clusters, whereas weak ties act as bridges between these clusters. According to Milroy's theory, individuals without strong ties are more likely to innovate. Thanks to their bridging role, they can introduce an innovation into one or more close-knit clusters. Once introduced, the innovation can be adopted by closely connected individuals within a cluster, who quickly propagate it to others.

Milroy's theory has found confirmation both in small-scale sociolinguistic studies carried out in the protestant enclaves of Belfast (Milroy and Milroy, 1985; Milroy and Milroy, 1987) and in agent-based models with simulated data (Fagyal et al., 2010). However, it has not been tested at a large scale on actual linguistic data. Moreover, the theory contrasts with other well-established models regarding the source of novel variants, which casts some doubt on the generality of its application: Labov's studies on New York adolescent gangs and the neighbourhoods of Philadelphia (Labov, 1972; Labov, 2001) identified *leaders*, i.e., those who have strong networks and are the most popular in their communities, as the drivers of change. Similarly, Eckert's (2000) studies of adolescent groups in Detroit showed that language novelty and change is led by charismatic leaders with strong ties to the local community.

We assess the applicability of Milroy's theory to the emergence of lexical innovations in online social communities, where relationships between individuals and their exposure to new terms can only be modelled in terms of their directly observable linguistic interaction in online discussion threads.

2.2 Related Computational Work

Research on language change has recently experienced a boost within the NLP community (Hamilton et al., 2016; Frermann and Lapata, 2016). Here we focus on reviewing approaches that are directly related to the work we present in this paper, i.e., approaches that investigate linguistic innovation within relatively short time spans of a few years (rather than looking at historical time) and that do so by leveraging data from social media to create a graph of connected users. More general literature on graph modelling and innovation diffusion is out of the scope of our review.

Graphs representing social networks have been used to analyse the extralinguistic factors that drive diffusion, such as geographical (Eisenstein, 2015) and demographic variables (Eisenstein et al., 2014), as well as the position of individuals in the social network (Paolillo, 1999) and the kind of user interactions that foster the diffusion of the innovations. For example, Goel et al. (2016) and Paradowski and Jonak (2012) investigate the amount of interaction required to adopt an innovation, showing that while some innovations are adopted by a new user after multiple interactions, for others a single contact is sufficient. Rotabi et al. (2017a) relate the success of competing lexical variants to the seniority of the users who use them. This work is extended by Rotabi et al. (2017b), who introduce an inheritance graph

[1] https://github.com/marcodel13/The-Road-to-Success

to represent how speakers pass innovations (in this case called *practices*) on to others in time through real-life collaborations.

Perhaps the work that is most directly connected to ours is that of Rotabi and Kleinberg (2016), who investigate the diffusion patterns of words that experience a frequency burst at a given time, which the authors call *trends*, and analyse the level of activity of the users involved in the different phases of the diffusion process. While related, our work differs on key points: we focus on linguistic innovations rather than trends and characterise users in terms of tie strength rather than level of activity.

A parallel line of work has investigated linguistic diffusion in relation to social ties using agent-based computer simulations (Ke et al., 2008; Fagyal et al., 2010; Swarup et al., 2011). In particular, Fagyal et al. (2010) address questions similar to ours via simulated data and find that peripheral users are crucial in introducing new linguistic variants, which are then adopted and spread by central agents in the community. The current availability of large social media datasets containing real linguistic interactions allows us to study these questions with more ecological validity.

3 Methodology

We describe our dataset, the methodology we use to define the social network of a community and the social role of its members, and the procedure to identify linguistic innovations and characterise their diffusion.

3.1 Data

We use data from Reddit,[2] a popular website which includes around 1 million communities called *subreddits*. A subreddit is a topic-based forum where users can submit posts, comment on existing posts or score them. While in this paper we treat each forum independently (leaving the analysis of innovation diffusion across forums for future work), in order to conduct a large-scale analysis and draw conclusions that generalise across communities, we analyse 20 different forums that show substantial variability in terms of subject matter and size (see Table 1 for an overview). Despite their rich heterogeneity, all subreddits we consider share the following features, which are indicative of highly active and interconnected communities, where the spreading cascade process at the base of innovation diffusion finds a favourable environment (Hamilton et al., 2017; Guille et al., 2013):

- a topic that reflects a strong external interest, such as sport teams, videogames or TV series;
- small-to-medium size, which in contrast to very large subreddits such as r/news (15M users) or r/funny (18M), are less dispersive and favour tighter connections;
- high density, i.e., high ratio of existing connections over the number of potential connections;[3]

We downloaded the entire content of each subreddit from its first post to the end of 2016. We segment the data from each subreddit into consecutive time bins corresponding to one month.[4] We discard time bins with less than 200 active users, which are common during the first few months of a subreddit lifespan, and ignore any posts whose author is unknown.[5]

Next, we explain how we leverage this longitudinal data to extract information about the social role of forum members, as well as to detect linguistic innovations and characterise their diffusion.

3.2 Social Network

We create a graph representing a community's social network for each month t during the community lifespan. In these graphs, nodes are users and edges encode whether users have interacted. We consider two users to be connected by an edge if they comment within the same thread in close proximity.[6] Given

[2] `https://www.reddit.com`

[3] Hamilton et al. (2017) report density values in the range $[0.001 - 0.016]$ in their set of subreddits.

[4] We also experimented with smaller bins of one week, obtaining similar results.

[5] When users delete their account, the posts, comments, and messages submitted prior to the deletion are still visible to others, but information about the user is not available (see Reddit Privacy Policy at `https://www.redditinc.com/policies/privacy-policy`).

[6] In particular, if they are separated by at most two posts, as done e.g., by Hamilton et al. (2017).

subreddit	years	tokens	users	density	innovations
r/Android	7	158	1.03M	0.006	730
r/apple	8	89	580k	0.006	584
r/baseball	6	101	576k	0.014	520
r/beer	7	29	291k	0.008	360
r/boardgames	6	88	313k	0.004	380
r/cars	6	101	544k	0.014	605
r/FinalFantasy	4	22	137k	0.009	218
r/Guitar	7	71	387k	0.009	496
r/harrypotter	5	39	287k	0.005	227
r/hockey	7	191	847k	0.012	602
r/Liverpool	5	40	173k	0.018	314
r/Patriots	5	26	151k	0.009	231
r/pcgaming	5	52	350k	0.003	360
r/photography	8	81	353k	0.006	485
r/pokemon	6	107	1.02M	0.006	695
r/poker	6	28	104k	0.012	258
r/reddevils	4	49	186k	0.008	329
r/running	6	56	279k	0.008	367
r/StarWars	6	56	542k	0.008	381
r/subaru	5	21	187k	0.005	340

Table 1: Statistics of the subreddits in our dataset, including: years of activity considered until end of 2016; total # of tokens (in millions); total # of active users (including users who may have left the community); average ratio of network ties over all possible ties computed over all the time bins (density); total # of linguistic innovations analysed.

that arguably lexical diffusion follows a simple "contagion" model whereby a single contact is sufficient for the spreading process (Goel et al., 2016), our graphs are undirected and unweighted.[7]

Milroy's (1987) theory relates the diffusion of a linguistic innovation to the local topology of individuals within a network, distinguishing between close-knit sub-groups and individuals outside of these groups. In line with other studies such as Weng et al. (2015) and Zhao et al. (2010), we build on a measure introduced by Onnela et al. (2007), which determines the strength of the edge between two individuals i and j in terms of the overlap O_{ij} of their adjacent neighbourhoods, as follows:

$$O_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}} \quad [1]$$

where n_{ij} is the number of adjacent nodes between i and j, and k_i, k_j their respective degree, i.e., the number of edges incident to each of them. Possible values for O_{ij} are in the range $[0, 1]$, where 0 indicates no common neighbours (weakest possible connection between i and j) and 1 exactly the same adjacent neighbours (strongest possible connection between i and j).

We now leverage Equation [1] to characterize users given the strength of their connections. According to Milroy (1987), weak-tie individuals have only weak connections (are not part of close-knit clusters), while strong-tie individuals have strong connections with other users, but may also have weak connections if they are linked to weak-tie users. To capture this, we define the tie strength of each individual i as the highest value of her incident edges. That is, for all individuals j directly connected to i:

$$\textit{tie-strength}(i) := \max(O_{ij}) \quad [2]$$

Taking the maximum captures what we are after: A community member who only has weak connections will have low tie-strength and be considered a *weak-tie user*, while a member with either strong connections only or with both strong and weak connections will have high tie-strength and be considered a *strong-tie user*, who will be part of a local clique (Luce and Perry, 1949). Unlike mean or median values, which tend to be rather balanced and thus do not help to distinguish between innovators and non-innovators, taking the maximum is in line with Milroy's theory and captures the key difference between these two groups of users, as will be shown in the next section.

[7]The graphs are implemented with Python's package `networkx`.

Besides computing tie-strength as defined in Equation [2] for all users in our social graphs, we also compute their centrality values for three common measures of network centrality — degree, betweenness, and eigenvector — which are global indices of the importance of a node with respect to all other nodes in a graph (Newman, 2010).

3.3 Linguistic Innovations

We focus on *Internet slang*, a general term commonly used to refer to a range of linguistic phenomena such as abbreviations (*cu* for *see you*), acronyms (*IIRC* for *if I remember/recall correctly*) and phonetic spellings (*dat* for *that*). The motivation behind this choice is twofold: firstly, forms of this kind are very abundant and continuously introduced in online communication; and secondly, they are easier to identify and track than innovations at other levels, such as those related to meaning shift. In the present study, we do not focus on tracking the co-evolution of two variants (e.g., *dat* vs. *that*) that may be competing, but rather on analysing the emergence of new forms and their trajectories independently, in light of the tie-strength of the members who use them.

Our starting point are the terms in the dictionary available at `NoSlang.com`,[8] a comprehensive record of Internet slang that is constantly updated. After removing terms including non-alphabetic characters, we obtain a list of approximately 6k terms. For each subreddit, we only consider terms that:

- are used at least 10 times in the subreddit;
- are not present during the first 3 months of the community's existence; and
- are introduced within the initial quarter of the community lifespan.

That is, we restrict our analysis to *newly* introduced Internet slang terms that are not present from the very beginning of the community's activity, but are not introduced too late so as to be able to observe their trajectory for a substantial period of time.[9]

The number of innovations considered across all subreddits is 7962, while the number of unique innovations amounts to 1456. Most of the terms (around 76%) occur in more than one community, although no innovation is present in all the subreddits in our dataset. Thus, around 24% of innovations tracked occur in just one community — some of these are clearly topic-related, e.g., *pkemon* in /r/pokemon, while others are more general purpose abbreviations that, in principle, could appear in any community, such as *txs* (*thanks*, in /r/Android) or *omgz* (*oh my god/gosh*, in /r/subaru). Regarding frequency, while we set a minimum threshold of 10 occurrences, in practice 72% of terms occur at least 50 times on average.

Dissemination trajectory. Once introduced, innovations may have different fates: they can spread widely within the community, be used by just a small sub-group, or fail to make an impact and disappear altogether. We define the fate of a term as its *dissemination*, which we compute as the proportion of community members who use it at a given moment in time (Del Tredici and Fernández, 2017; Altmann et al., 2011). It should be stressed that dissemination differs from frequency. Although often they are highly correlated, in principle a term can have high relative frequency but low dissemination and vice versa. We quantify the diffusion of innovations in terms of their dissemination since this gives us a measure of their spread within a community. To this end, for each innovation we calculate its *dissemination trajectory* as the vector of its monthly dissemination values since the innovation was introduced.

Tie-strength trajectory. Similarly, we compute the *tie-strength trajectory* of each innovation as a vector whose features correspond to the maximum tie-strength value among the users that used it in the corresponding month. We choose the maximum value in order to test Milroy's (1987) hypothesis, according to which the crucial factor in the diffusion of an innovation is its adoption by strong-tie members (see Section 2). Considering only the maximum value provides a simple way to test whether any individual with high tie-strength has used the term in a given month.[10]

[8] `https://www.noslang.com/dictionary/`

[9] To assess whether ambiguity may be an issue, we checked whether the terms also appear in a standard English dictionary, PyDictionary. For example, the slang term *bra* for *brother* also has the standard meaning of *brasserie*. Given that under 2% of terms in our dataset could potentially be ambiguous, we decided to not treat them in any special way.

[10] The dissemination and the tie-strength vectors always have the same magnitude.

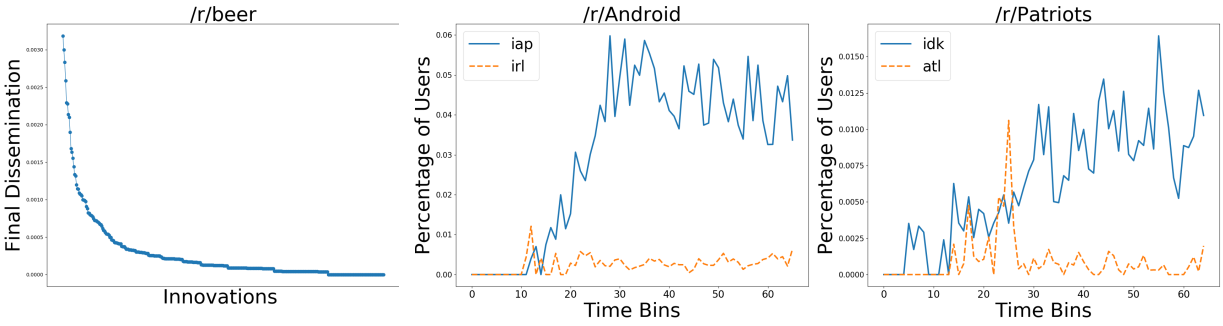

Figure 1: **Left**: Distribution of the final dissemination values in /r/beer. **Center/Right**: Examples of dissemination trajectories of *successful* (blue solid line) and *unsuccessful* (orange dashed line) innovations.

4 Empirical Observations

Our analysis reveals some patterns common to all communities, which we present in this section.

4.1 Linguistic Innovations

In order to explore the relative success of innovations, we consider the level of dissemination reached by an innovation to be the average dissemination value in the last six months.[11] Results show that the distribution of these final dissemination values is highly skewed for all the subreddits — see, for example, Figure 1 (left) for /r/beer. While a few innovations disseminate successfully (i.e., are adopted by a relatively high number of community members), most of them do not spread, and either disappear or barely appear in the last period.

In Figure 1 (center/right) we show examples of successful and unsuccessful innovations. Successful innovations such as *iap* (*In App Purchases*) and *idk* (*I don't know*) show a stable increase in dissemination after their introduction, which can reach a plateau at some point (*iap*) — thus showing the S-shaped curve typical of general processes of innovation adoption (Rogers, 2010) — or can still be ongoing at the end of the period covered by our analysis (*idk*). Unsuccessful innovations, in contrast, can either have a flat dissemination trajectory, as in the case of *irl* (*In Real Life*), indicating that the term has never experienced a spread in the community, or present a peak at some point, followed by a sudden decrease with no stable recovery, as for *atl* (*Atlanta*).

Given these observations, we formally define the classes of *successful* and *unsuccessful* innovations based on the dissemination *slope* of a term, computed as the difference between its average dissemination value in the first six months and in the last six months in the dissemination trajectory vector. We include in the *unsuccessful* class innovations with slope index ≤ 0, i.e., those with trajectories similar to the *irl* and *atl* examples in Figure 1 (center/right). In order to discard innovations with very low positive slope (i.e., those that do not disappear, but are only sporadically used) we only include in the *successful* class terms whose slope index is above the average value of the community.[12] We will make use of these two classes in the prediction experiment we present in Section 6.

4.2 Social Networks

Next, we analyse the distribution of users' tie-strength values in the social graphs derived for the communities in our dataset.[13] We find a clear pattern for all subreddits: the large majority of users have low tie-strength, with around 39% having values ≤ 0.05 and almost 50% having values ≤ 0.1; while, around 15 to 20% of users have strong tie-strength, with values ≥ 0.5. Figure 2 shows the average tie-strength value distribution computed over all the monthly graphs of all subreddits in the dataset, with probabilities calculated for bins of size 0.1 for illustration purposes.

[11] This makes our measurements more robust than taking only the very last bin.

[12] Average slope index is positive for all subreddits.

[13] A sample graph is available at `https://github.com/marcodel13/The-Road-to-Success`.

This distribution mirrors the typical power-law distribution observed for centrality measures in online communities (Mihalcea and Radev, 2011). The topological properties captured by our tie-strength measure (Equation [2]), however, are different from those captured by centrality, as already hinted at in Section 3.2. The three centrality measures considered (degree, betweenness, and eigenvector) correlate strongly with each other (Spearman's r in range 0.85–0.89).[14] But there is only a moderate correlation with tie-strength: r=0.63 degree, r=0.61 betweenness, and r=0.47 eigenvector. In addition, we observe that the three centrality measures correlate strongly with number of posts (r=[0.78–0.91]) in all subreddits, while we find low correlation between number of posts and tie-strength (r=0.31, std=0.08, $p < .05$).

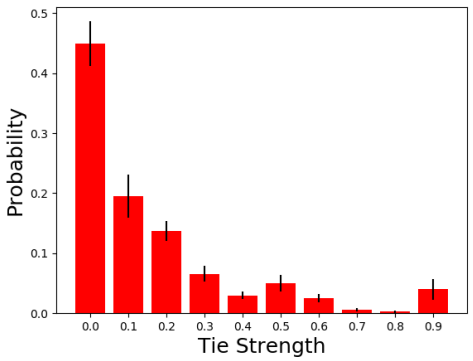

Figure 2: Average tie-strength distribution for all subreddits, with standard deviation.

These results confirm the difference between our tie-strength measure and centrality. While centrality values are *global* indices of the role of a node with respect to the entire graph (Newman, 2010), tie-strength captures the *local* topological information around a node. In the online social communities we investigate, individuals at the core of the social network, who interact with many other individuals and have high posting activity, receive high centrality values. In contrast, high tie-strength values are the signature of users who belong to small cliques, but who do not act as hubs for the entire network. We take this as confirmation that our tie-strength measure does capture key features of the social structures underpinning Milroy's (1987) theory. It remains to be seen, however, whether these social structures, as realised in large online social communities, lend support to the theory's main claims.

5 Assessing Sociolinguistic Claims

In this section, we uncover the features that characterise innovators and analyse the role of strong-tie users in the dissemination process.

5.1 Innovators

We consider *innovators* those members who introduce a new term, i.e., those who use it for the very first time in a community. In order to verify whether innovators are weak-tie users (as hypothesised by Milroy (1987)), we compare the distribution of the tie-strength values of innovators to the general tie-strength distribution in the community. Since innovations are introduced at different points in time, the general tie-strength distribution in the community is defined as the average of the months at which innovations were introduced. We compute Kullback-Leibler divergence (KL), which measures how similar the behaviour of two distributions is. We find KL values in the range 0.15–0.4, which indicates a moderate difference between the distributions. The source of this difference can be appreciated in Figure 3: the tie-strength values of innovators cluster around 0.1–0.3.[15] This contrasts with the overall tie-strength distribution: innovators tend to *not* be strong-tie users (lower blue than red bars for tie-strength ≥ 0.4 in Figure 3) and are far less likely to

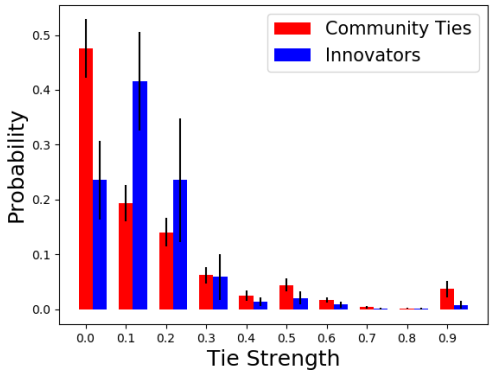

Figure 3: Comparison of the probability mass distribution of innovators' tie-strength values and avg. values of all users in all subreddits.

[14] All correlation coefficients reported are averages across time bins and subreddits and are all significant with $p < 0.05$.

[15] The same trend holds for all subreddits. Individual plots per subreddit can be found at `https://github.com/marcodel13/The-Road-to-Success`.

have very weak tie-strength than most users (lower blue than red bar for tie-strength < 0.1).

In addition to analysing tie-strength, we compare the centrality values and the posting activity of innovators and non-innovators. We observe that innovators are significantly more central than other users (for all measures considered: degree, betweenness and eigenvector) and are significantly more active in terms of number of posts than other individuals in the community (unpaired Welch's t-tests, all with $p < 0.05$).

Thus, a very robust pattern emerges across all subreddits, showing that innovators do have a particular profile in terms of their social standing: they do not belong to tightly connected cliques and occupy a central position in the network, as hubs with many connections of relatively low strength. On the one hand, this seems in line with Milroy's hypothesis, since it confirms that innovations do not arise within sub-communities that are close-knit. On the other hand, our results may also be interpreted as lending support to Labov's (1972; 2001) characterisation of innovators as *leaders*, since they occupy a core position in the network.

5.2 Strong-Tie Users and Innovation Spread

We consider strong-tie users those with a tie-strength value ≥ 0.5. By definition, these are users who are part of cliques or sub-communities within a subreddit. As mentioned in Section 4.2, strong-tie users constitute between 15 and 20% of the total number of community members. Thus, in contrast to the small scale social communities examined by Milroy and other sociolinguists, where most community members are part of some tight-knit sub-group (such as a family or a church congregation), in large online social networks strong-tie users are a minority. Despite this, strong-tie users are not isolated in remote cliques: Our analysis shows that, in all subreddits, strong-tie users are significantly more central (with respect to all, degree, betweenness, and eigenvector centrality) and have significantly more posting activity than users with weak tie-strength values ≤ 0.05, who make up the vast majority of community members — around 39% on average (unpaired Welch's t-test, $p < 0.05$). Thus, in terms of centrality, strong-tie members occupy an intermediate position in the social network, forming a loose ring around the core innovators and the majority of members, who are on the periphery and are characterised by very weak tie-strength.

To analyse the role of strong-tie users in the innovation diffusion process, we proceed as follows: We identify any time t_i in the tie-strength trajectory vector when the innovation is used by some strong-tie member for k consecutive months. We then check its average dissemination in the period up to time t_i and compare it to the average dissemination in the six months following t_{i+k-1}. We find that when $k = 1$ (i.e., when an innovation has been used by a strong-tie member only in one month) the probability that the dissemination increases in the next six months is around 50% for all subreddits — a value similar to the likelihood of dissemination increase after the usage by a weak-tie user. However, as k increases, and thus the adoption by strong-tie users becomes more stable, a future increase in dissemination becomes progressively more likely, for all the subreddits. Importantly, the same effect is not observed for weak-tie users, for whom, independently from the number of months, the probability of a future increase in dissemination is always approximately the same. Figure 4 shows examples of how the probability of dissemination changes after the adoption by either strong- or weak-tie users (see Appendix A for full results).

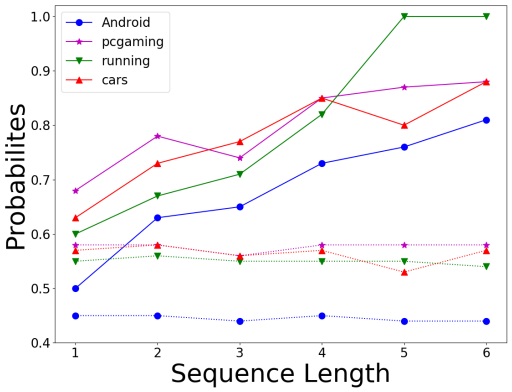

Figure 4: Probability of dissemination increase after a term is adopted by a strong-tie user (solid line) or by a weak-tie user (dotted line) for k consecutive months, computed for k in range $[1 - 6]$.

These results, thus, are consistent with Milroy's (1987) claims, and furthermore show that innovation diffusion is connected to *sustained* adoption by strong-tie community members.

6 Predicting Innovation Success

Most innovations do not succeed in becoming community norms, but some do. Here we assess whether information about the tie-strength of members who use an innovation in the first months after its introduction can predict whether it will be successful in the future. This provides further theoretical insight into the importance of tie-strength for innovation diffusion, and has practical significance by contributing to identifying new terms that NLP systems should be able to process. Our aim here is not to maximise prediction accuracy—which is likely to require taking into account several factors beyond users' tie-strength—but rather to explore whether the statistical effects we have uncovered are strong enough to have some predictive power.

We approach this as a binary classification task, making use of the distinction between *successful* and *unsuccessful* innovations defined in Section 4. We extract a subvector of length k from the tie-strength trajectory vector of innovations and use it as features for the prediction. For instance, with $k = 3$, we use the tie-strength information from the first three months of usage of a term to predict if it will be successful or not, leveraging subvectors of increasing magnitudes. We use Python's `scikit-learn` Random Forest classifier with default parameters[16] and perform 100-run cross-validation, using 90% of the data for training and 10% for testing. We compare our results against a *weighted baseline*, whereby the two labels (successful/unsuccessful) are randomly assigned taking into account their frequency in the training set. The classes are fairly balanced across subreddits, with an average proportion of 55% successful and 45% unsuccessful.

When leveraging tie-strength information from only the first 3 months of usage, we obtain F1 results that are significantly higher than the baseline for 12 out the 20 subreddits. But overall, performance remains rather low, with an average F1-score for the *successful* class of 0.62 vs. 0.58 for the baseline. Given that new terms are introduced by users with relatively low tie-strength (as shown in Section 5.1), arguably in the initial few months before a novel term is picked up by a strong-tie user, there is little difference between successful and unsuccessful innovations. With tie-strength information from the first 6 months of usage, we are able to make predictions with results significantly above baseline for 18 out 20 subreddits, with an average F1-score of 0.68. Not surprisingly, performance increases substantially when information for a longer period (first/second year of usage) is exploited, reaching an average F1-score of 0.76, significantly above the baseline for all communities. Detailed results, including precision, recall, and F1-score for each subreddit, can be found in Appendix B. In Figure 5 we graphically illustrate the results for a few subreddits, which are representative of the general trend observed.

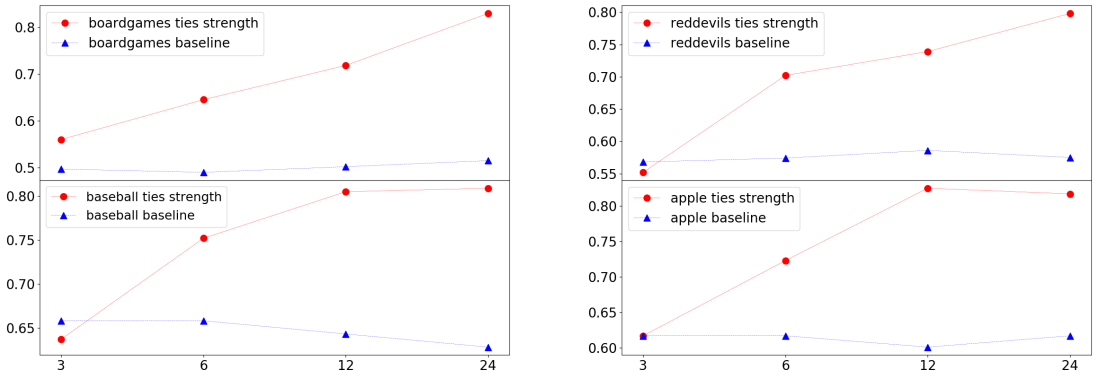

Figure 5: F-score (y-axis) for the successful class obtained with the ties-strength values of the first k months (x-axis) after the introduction of a term.

[16] `http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`.

7 Conclusions

This work has provided a large-scale analysis of the interplay between the birth and spread of new terms and users' social standing in large online social communities. Building on sociolinguistic theories — in particular, Milroy's (1987) version of *The Strength of Weak Ties* theory — we have proposed a simple measure to quantify tie-strength in Milroy's sense and have used it in combination with common centrality measures to uncover the characteristics of innovators and to assess the role of strong-tie users in the dissemination process.

Regarding innovators, our results show that they are central community members, connected to many other users but with relatively low tie-strength. As for strong-tie users, we find that in online social networks they are a small proportion of community members, organised in small cliques, and that they do play an important role in the spread of an innovation, presumably by spreading new terms introduced by innovators into their sub-groups. The patterns we have revealed are surprisingly consistent across the 20 online communities we have investigated.

Our work opens a range of interesting questions. In particular, we are looking forward to investigating how the patterns we have observed are affected by users' membership in multiple forums and how these multi-community users transfer new terms between communities.

Acknowledgements

This research has received funding from the Netherlands Organisation for Scientific Research (NWO) under VIDI grant nr. 276-89-008, *Asymmetry in Conversation*. We thank the anonymous reviewers for their comments as well as the area chairs and PC chairs of COLING 2018.

References

Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. 2011. Niche as a determinant of word fate in online groups. *PLoS ONE*, 6(5):e19009.

Jack K Chambers and Natalie Schilling. 2013. *The handbook of language variation and change*, volume 129. John Wiley & Sons.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

Penelope Eckert. 2000. *Language variation as social practice: The linguistic construction of identity in Belten High*. Wiley-Blackwell.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE*, 9(11):e113114.

Jacob Eisenstein. 2015. Identifying regional dialects in online social media. In C. Boberg, J. Nerbonne, and D. Watt, editors, *Handbook of Dialectology*. Wiley.

Zsuzsanna Fagyal, Samarth Swarup, Anna María Escobar, Les Gasser, and Kiran Lakkaraju. 2010. Centers and peripheries: Network roles in language change. *Lingua*, 120(8):2061–2079.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *International Conference on Social Informatics*, pages 41–57. Springer.

Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.

Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. 2013. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016*.

William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Proceedings of the eleventh International Conference on Web and Social Media*.

Jinyun Ke, Tao Gong, and William SY Wang. 2008. Language change and social networks. *Communications in Computational Physics*, 3(4):935–949.

William Labov. 1972. *Language in the inner city: Studies in the Black English vernacular*. University of Pennsylvania Press.

William Labov. 2001. *Principles of linguistic change: Social factors*. Blackwell.

R Duncan Luce and Albert D Perry. 1949. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116.

Rada Mihalcea and Dragomir Radev. 2011. *Graph-based natural language processing and information retrieval*. Cambridge University Press.

James Milroy and Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of linguistics*, 21(2):339–384.

James Milroy and Lesley Milroy. 1987. Belfast: change and variation in an urban vernacular. In *Sociolinguistic patterns in British English*, pages 19–36. E. Arnold.

Lesley Milroy. 1987. *Language and social networks*. Blackwell.

Mark Newman. 2010. *Networks: an introduction*. Oxford University Press.

Jukka-Pekka Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336.

John Paolillo. 1999. The virtual speech community: Social network and language variation on irc. *Journal of Computer-Mediated Communication*, 4(4):JCMC446.

Michał B Paradowski and Łukasz Jonak. 2012. Diffusion of linguistic innovation as social coordination. *Psychology of Language and Communication*, 16(2):131–142.

Everett M Rogers. 2010. *Diffusion of innovations*. Simon and Schuster.

Rahmtin Rotabi and Jon M Kleinberg. 2016. The status gradient of trends in social media. In *Proceedings of the tenth International Conference on Web and Social Media*, pages 319–328.

Rahmtin Rotabi, Cristian Danescu-Niculescu-Mizil, and Jon Kleinberg. 2017a. Competition and selection among conventions. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1361–1370. International World Wide Web Conferences Steering Committee.

Rahmtin Rotabi, Cristian Danescu-Niculescu-Mizil, and Jon Kleinberg. 2017b. Tracing the use of practices through networks of collaboration. In *Proceedings of the 11th International Conference on Web and Social Media*.

Samarth Swarup, Andrea Apolloni, and Zsuzsanna Fagyal. 2011. A model of norm emergence and innovation in language change. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 693–700. International Foundation for Autonomous Agents and Multiagent Systems.

Uriel Weinreich, William Labov, and Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In *Directions for Historical Linguistics: a Symposium*, pages 95–188. University of Texas Press.

Lilian Weng, Márton Karsai, Nicola Perra, Filippo Menczer, and Alessandro Flammini. 2015. Attention on weak ties in social and communication networks. *arXiv preprint arXiv:1505.02399*.

Jichang Zhao, Junjie Wu, and Ke Xu. 2010. Weak ties: Subtle role of information diffusion in online social networks. *Physical Review E*, 82(1):016105.

Appendix A: Probability of Dissemination Increase

The table in this appendix complements the results presented in Section 5.2.

	k=1		k=2		k= 3		k=4		k=5		k=6	
subreddit	**strong**	**weak**	**strong**	**weak**	**strong**	**weak**	**strong**	**weak**	**strong**	**weak**	**strong**	**weak**
r/Android	0.5	0.45	0.63	0.45	0.65	0.44	0.73	0.45	0.76	0.44	0.81	0.44
r/apple	0.58	0.47	0.7	0.48	0.73	0.48	0.8	0.47	0.8	0.47	0.81	0.47
r/baseball	0.62	0.55	0.73	0.55	0.71	0.54	0.76	0.56	0.75	0.55	0.75	0.54
r/beer	0.51	0.49	0.62	0.49	0.68	0.49	0.69	0.48	0.72	0.48	0.73	0.48
r/boardgames	0.71	0.6	0.83	0.6	0.8	0.58	0.88	0.6	0.7	0.56	0.9	0.6
r/cars	0.63	0.57	0.73	0.58	0.77	0.56	0.85	0.57	0.8	0.53	0.88	0.57
r/FinalFantasy	0.61	0.58	0.59	0.57	1.0	0.56	–	–	–	–	–	–
r/Guitar	0.57	0.53	0.7	0.54	0.7	0.52	0.81	0.53	0.89	0.53	0.77	0.53
r/harrypotter	0.53	0.51	0.5	0.5	1.0	0.51	–	–	–	–	–	–
r/hockey	0.74	0.61	0.8	0.61	0.89	0.62	0.76	0.58	0.83	0.61	0.94	0.62
r/Liverpool	0.61	0.52	0.61	0.51	0.67	0.53	0.67	0.49	–	–	–	–
r/Patriots	0.65	0.64	0.62	0.63	0.6	0.65	0.67	0.65	–	–	–	–
r/pcgaming	0.68	0.58	0.78	0.58	0.74	0.56	0.85	0.58	0.87	0.58	0.88	0.58
r/photography	0.65	0.57	0.73	0.57	0.82	0.57	0.84	0.58	0.75	0.57	0.73	0.57
r/pokemon	0.54	0.48	0.66	0.48	0.69	0.48	0.69	0.48	0.71	0.47	0.71	0.47
r/poker	0.57	0.57	0.7	0.57	1.0	0.57	–	–	–	–	–	-
r/reddevils	0.56	0.54	0.54	0.53	0.6	0.52	0.6	0.5	0.4	0.5	1.0	0.49
r/running	0.6	0.55	0.67	0.56	0.71	0.55	0.82	0.55	1.0	0.55	1.0	0.54
r/StarWars	0.61	0.5	0.77	0.51	0.82	0.51	0.88	0.5	0.9	0.5	0.91	0.5
r/subaru	0.49	0.53	0.53	0.51	0.92	0.52	0.8	0.48	–	–	–	–

Table 2: Probability of increase in dissemination of a linguistic innovation after being used by a **strong-tie** or a **weak-tie** user for k consecutive months. Missing values indicate no such condition was found in a community.

Appendix B: Detailed Results on Success Prediction

The following table gives detailed results for the prediction task described in Section 6. Statistical significance between tie-strength features and the weighted baseline is computed with the dependent t-test for paired samples implemented by Python's `scipy` package on results from 100 runs.

		k=3			k=6			k=12			k=24		
subreddit		**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Android	**t**	0.49	0.5	0.49	0.55	0.57	0.56	0.59	0.59	0.59	0.68	0.69	0.68
	b	0.37	0.38	0.37	0.38	0.38	0.38	0.4	0.41	0.4	0.39	0.39	0.39
apple	**t**	0.64$^{\#}$	0.6$^{\#}$	0.62$^{\#}$	0.72	0.73	0.72	0.8	0.85	0.82	0.78	0.86	0.82
	b	0.64$^{\#}$	0.62$^{\#}$	0.63$^{\#}$	0.66	0.6	0.63	0.62	0.61	0.61	0.65	0.61	0.63
baseball	**t**	0.63	0.65$^{\#}$	0.64$^{\#}$	0.71	0.79	0.75	0.73	0.89	0.8	0.74	0.89	0.81
	b	0.67	0.67$^{\#}$	0.67$^{\#}$	0.66	0.68	0.67	0.64	0.67	0.65	0.63	0.65	0.64
beer	**t**	0.52$^{\#}$	0.52$^{\#}$	0.52 $^{\#}$	0.58	0.62	0.6	0.59	0.65	0.62	0.64	0.7	0.67
	b	0.52$^{\#}$	0.51$^{\#}$	0.51$^{\#}$	0.51	0.49	0.5	0.49	0.5	0.49	0.52	0.5	0.51
boardgames	**t**	0.56$^{\#}$	0.56$^{\#}$	0.56	0.64	0.65	0.64	0.69	0.74	0.71	0.81	0.85	0.83
	b	0.52$^{\#}$	0.51$^{\#}$	0.51	0.5	0.51	0.5	0.55	0.5	0.52	0.54	0.52	0.53
cars	**t**	0.66	0.66	0.66	0.7	0.76	0.73	0.71	0.82	0.76	0.72	0.85	0.78
	b	0.6	0.61	0.6	0.6	0.61	0.6	0.6	0.6	0.6	0.58	0.59	0.58
Finalfantasy	**t**	0.67$^{\#}$	0.6$^{\#}$	0.63$^{\#}$	0.65$^{\#}$	0.7	0.67	0.76	0.84	0.8	0.79	0.86	0.82
	b	0.65$^{\#}$	0.58$^{\#}$	0.61$^{\#}$	0.63$^{\#}$	0.58	0.6	0.63	0.6	0.61	0.64	0.6	0.62
Guitar	**t**	0.68	0.65	0.66	0.68	0.76	0.72	0.71	0.81	0.76	0.75	0.85	0.8
	b	0.58	0.6	0.59	0.58	0.61	0.59	0.56	0.6	0.58	0.58	0.6	0.59
harrypotter	**t**	0.53$^{\#}$	0.55$^{\#}$	0.54	0.57	0.58	0.57	0.56	0.6	0.58	0.51	0.55	0.53
	b	0.49$^{\#}$	0.51$^{\#}$	0.5	0.47	0.48	0.47	0.48	0.52	0.5	0.45	0.48	0.46
hockey	**t**	0.72	0.75	0.73	0.72	0.79	0.75	0.68$^{\#}$	0.81	0.74	0.74	0.88	0.8
	b	0.64	0.64	0.64	0.68	0.62	0.65	0.66$^{\#}$	0.63	0.64	0.64	0.62	0.63
Liverpool	**t**	0.59	0.59	0.59	0.62	0.64	0.63	0.58$^{\#}$	0.61	0.59	0.69	0.78	0.73
	b	0.53	0.5	0.51	0.53	0.52	0.52	0.54$^{\#}$	0.51	0.52	0.53	0.53	0.53
Patriots	**t**	0.72$^{\#}$	0.74	0.73	0.8	0.81	0.8	0.81	0.88	0.84	0.81	0.92	0.86
	b	0.71$^{\#}$	0.69	0.7	0.64	0.69	0.66	0.67	0.71	0.69	0.65	0.69	0.67
pcgaming	**t**	0.78	0.7$^{\#}$	0.74	0.77	0.83	0.8	0.82	0.86	0.84	0.79	0.89	0.84
	b	0.66	0.66$^{\#}$	0.66	0.68	0.68	0.68	0.67	0.68	0.67	0.68	0.7	0.69
photography	**t**	0.72	0.64	0.68	0.74	0.72	0.73	0.74	0.8	0.77	0.77	0.85	0.81
	b	0.63	0.62	0.62	0.65	0.61	0.63	0.64	0.61	0.62	0.63	0.61	0.62
pokemon	**t**	0.65	0.67	0.66	0.58	0.66	0.62	0.62	0.77	0.69	0.68	0.79	0.73
	b	0.53	0.52	0.52	0.51	0.5	0.5	0.53	0.52	0.52	0.52	0.49	0.5
poker	**t**	0.58$^{\#}$	0.62$^{\#}$	0.6$^{\#}$	0.6$^{\#}$	0.64$^{\#}$	0.62$^{\#}$	0.58	0.63$^{\#}$	0.6$^{\#}$	0.65	0.78	0.71
	b	0.62$^{\#}$	0.6$^{\#}$	0.61$^{\#}$	0.61$^{\#}$	0.59$^{\#}$	0.6 $^{\#}$	0.63	0.59$^{\#}$	0.61$^{\#}$	0.58	0.6	0.59
reddevils	**t**	0.57$^{\#}$	0.54$^{\#}$	0.55$^{\#}$	0.68	0.72	0.7	0.7	0.78	0.74	0.74	0.86	0.8
	b	0.59$^{\#}$	0.58$^{\#}$	0.58$^{\#}$	0.58	0.59	0.58	0.58	0.61	0.59	0.59	0.59	0.59
running	**t**	0.58$^{\#}$	0.55$^{\#}$	0.56$^{\#}$	0.64	0.68	0.66	0.65	0.77	0.7	0.66	0.81	0.73
	b	0.57$^{\#}$	0.58$^{\#}$	0.57$^{\#}$	0.58	0.61	0.59	0.56	0.61	0.58	0.58	0.58	0.58
Starwars	**t**	0.63	0.63$^{\#}$	0.63$^{\#}$	0.57$^{\#}$	0.63$^{\#}$	0.6$^{\#}$	0.6$^{\#}$	0.67	0.63	0.65	0.76	0.7
	b	0.55	0.58$^{\#}$	0.56$^{\#}$	0.6$^{\#}$	0.61$^{\#}$	0.6$^{\#}$	0.56$^{\#}$	0.59	0.57	0.6	0.59	0.59
subaru	**t**	0.69	0.64$^{\#}$	0.66	0.69	0.79	0.74	0.69	0.83	0.75	0.69	0.84	0.76
	b	0.63	0.61$^{\#}$	0.62	0.61	0.61	0.61	0.61	0.61	0.61	0.62	0.59	0.6
Average	**t**	0.63	0.62	0.62	0.7	0.66	0.68	0.68	0.76	0.72	0.71	0.81	0.76
	b	0.58	0.57	0.58	0.58	0.57	0.58	0.58	0.58	0.58	0.58	0.58	0.58

Table 3: Average precision (P), recall (R), and F1-score for the *successful* class over 100 runs with 10-fold cross-validation. k= length of the tie-strength vector used for the prediction (corresponding to number of months); **t / b**= results obtained using tie-strength information and the weighted baseline, respectively. Difference between **t** and **b** is significant ($p < 0.05$) except when marked with $^{\#}$.