In: Social Computing, Behavioral Modeling and Prediction, LNCS, Washington, D.C.: Springer, 2017

Is Word Adoption a Grassroots Process? An Analysis of Reddit Communities

Jeremy R. Cole, Moojan Ghafurian, and David Reitter

The Pennsylvania State University
University Park, Pennsylvania, USA
{jrcole,moojan,reitter}@psu.edu

Abstract This study examines how novel words originate in and disperse through online communities. It asks whether larger numbers of people or closer social ties are better environments to foster the adoption of new words. The data stem from Reddit, a large sample of web-mediated, asynchronous conversations. Reddit communities are divided by size in this study: larger communities are based on discussing general topics and have weak social ties, whereas small communities are based on discussing specific topics and have strong social ties. The analysis shows that the majority of new words are created / first adopted in larger communities.

Keywords: word adoption, dispersion, Reddit, community organization

1 Introduction

Language is a communication system that varies among speakers and is constantly changing. English is a particularly productive language: new words are invented frequently. In fact, the rate of new word formation has increased in the past century [6]. Newly introduced words might be used for only a short period of time or may last longer and contribute to large-scale language change. This process relies on speakers taking liberties with their word choice and on speaker communities that facilitate and accept the use of novel words.

Experimentally, lexical change has been studied in relatively small groups: for example, using the *naming game* paradigm. Such experiments show how new words can be created and will emerge as an agreed-upon standard in these small, artificial and temporary communities that are created for that purpose. The naming game can be seen as a model of how communities reach consensus about a communication system, naming, or generalized linguistic systems [3]. In most communities, the members successfully reach consensus in such games [4].

There are individual and group differences in the task. For instance, such differences could lead to some migratory speakers adopting new words more than others. This result was found by comparing old and new profession names in census data [9]. Further, mixing up group composition can increase the quality of the communication systems produced [4]. However, even relatively stationary members of these communities can adopt new ideas and words.

Group size also influences the convergence of communication systems. As Dall'Asta et al. [3] discuss, the time of convergence increases as the population size increases. According to this model, smaller groups are expected to converge faster (thus agree on new words faster) than larger ones. However, in the context of a social network, more centralized distribution leads to faster adoption [10]. Can both be true?

Twitter has also been used to study word adoption in the context of larger networks. The dispersion dynamics of hashtags can be surprising in that they appear to be different depending on the topic [8]. Still, the interaction of those broadcasting tweets in a public forum is not necessarily a good model of the directed communication of communities: users have weak social ties with each other, and they are not naturally partitioned in such a way as to study the nature of the communities that facilitate innovation and early adoption of new ideas or words.

In naming games the focus is on language as a shared community contract (e.g., [4, 7]); however, it does not discuss interactions across communities and its framework limits the possible number of participants. In this study, we focus on the conditions that facilitate the adoption of novel words within and across larger communities. As we will see, group size in the communities we study is, surprisingly, correlated with more rather than less innovation. As we address the features of communities that facilitate creativity or early adoption of new concepts, we examine data source stemming from delineated, but connected communities.

To explore the effect of community size on the rate of adoption of novel words, we analyze the *Reddit* dataset, as it provides us with a variety of groups of different sizes. The question we ask regarding the communities relate to their size and specificity: are words first adopted in more specific, smaller communities on Reddit and dispersed to the larger, more general communities, or does dispersion take place in the opposite direction? We think that answering this question could give valuable cues as to whether and how group size promotes linguistic and thematic innovation.

2 Methods

2.1 Data: Reddit Corpus

Our data set consists of approximately 426GB of Reddit data, ranging from the year 2012 to the year 2014.

Reddit.com is a community-driven news aggregation website that mostly contains discussions and ratings [2]. Communities on Reddit are divided into different categories (such as Education, Technology, image sharing, etc.), and each of these general categories are then divided into several subcategories, which are called *subreddits*. For example, Educational category can have science as a subreddit. Each of these subreddits can be about any given topic, general or specific. For instance, it is possible to have subreddits about science, biology,

and genetics. These three subreddits can be completely independent, as they are not organized hierarchically.

Before applying any of our analysis, we filter out comments in subreddits with a small number of users. As anyone can make a subreddit and invite their friends to join, we wanted to avoid small subreddits that may more closely resemble social networks than communities.

2.2 Defining New Words

As discussed, our data spans 2012-2014. In this sense, we came up with a simple way to determine if a word is new: its time of first use was more than a year after the first comment for the data we used. We also excluded words with a small number of uses or that did not consist entirely of alphabetic characters. Some example words can be found in Table 1. In total, we found 3550 words matching these criteria.

These words are not entirely recognizable to those not in the culture, and they vary in their novelty. For instance, Square Cash, a financial product, is frequently referred to as *squarecash* by Reddit users. Others, however, are strictly adoptions, such as Chromecast. Thus, we will refer to these as *first adoption* events.

While some of the first adoption events are origination events, all of them are a discussion of something new. The first discussion of a new idea has social consequence. In Reddit, people receive both explicit and implicit rewards for social acceptance, through the curation mechanism. Thus, first adoptions will likely occur in communities that maximize this payoff. Thus, we seek to determine what type of community that is. We will use *adoption* to refer to any usage of a new word by a subreddit, using origination or first adoption for the first subreddit to adopt it, and *later adoption* for later usages.

Word	First Adopter	Later Adopter
dogetips	dogecoin	funny
isanderkirby	AdviceAnimals	AskReddit
peshka	gaming	Warthunder
gamecribs	leagueoflegends	counterstrike
squarecash	economy	Bitcoin

Subsubreddit	Supersubreddit
UniversityOfHouston	houston
justneckbeardthings	fatpeoplestories
simpleios	iOSProgramming
DotaCR	DotA2
lisp	programming

Table 1: On the left, there are sample words with their originating subreddit, along with an example destination subreddit. On the right, there are example pairs of subsubreddits and supersubreddits.

2.3 Inducing the Structure of Reddit Subforums

Reddit's inherent organization is shallow, rather than hierarchical. Underneath the top-level hierarchy, subreddits are not formally organized. Still, the topics have a range of specificities. For instance, there could be a subreddit focused on board games in general, with a separate subreddit for specific board games,

such as Settlers of Catan or Monopoly. Our intuition suggests then, that people passionate about Monopoly are also passionate about board games in general.

We then define a *subsubreddit* and a *supersubreddit* as such: A is a subsubreddit of B if at least $N\%$ of A's members are also members of B. B is a supersubreddit of A if and only if A is a subsubreddit of B. We used $N = 25\%$, because for this value, for any subsubreddit, A of B, A was not also a supersubreddit of B. This resulted in approximately four thousand pairs of subreddits. In general, the supersubreddits will cover more general topics and involve more users, while the subsubreddits will cover more narrow topics and have fewer users. Some example pairs can be found in Table 1. As an important point, this relationship is ultimately relative. Lastly, this pairing is somewhat conservative: not all subreddits are in any pairing. Even larger subreddits were only in a relationship with intuitive relations: AskReddit's only relationship was with TrueAskReddit.

Still, these pairs are largely interesting in the context of our research question: which types of communities lead to first adoption, rather than later adoption?

2.4 Results

The results of all of the Wilcoxon Signed-Rank Tests are found in Table 2. Our first analysis focused on first adoptions that were later adopted by the other in the pair (Paired Adoptions). For origination events, this could be an adoption from the previous subreddit, rather than an external source. Regardless, it reflects words that are of interest to both subreddits in the pair, to examine how such words move through subreddits around that topic. We find that supersubreddits have significantly more first adoptions than the subsubreddits.

	V	Mean-sub	Mean-super	p-value
Paired Adoptions	8263.5	0.5083	4.9472	< 0.0001
Full First Adoptions	32158	0.3632	6.1997	< 0.0001
Full Later Adoptions	121350	6.9978	41.9935	< 0.0001
Norm(U) First Adoptions	95613	0.0001	0.0004	< 0.0001
Norm(U) Later Adoptions	850780	0.0052	0.0030	0.7078
Norm(C) First Adoptions	114840	0.00001	< 0.00001	< 0.0001
Norm(C) Later Adoptions	1233900	0.0004	0.0001	< 0.0001

Table 2: The results of the Wilcox Signed Rank Tests, Norm(U) is normalized by number of users, while Norm(C) is normalized by number of comments

Our second analysis focuses on the total numbers (Full). While more words may originate in any given super-subreddit than a sub-subreddit, it is possible that the totals tell a different story. In this analysis, we maintain the pairings, but instead look at the total number of words that originated in that subreddit and the total number of words that were adopted by that subreddit. In this analysis, there are once again reliably more first adoption events in the supersubreddits. Further, there are more later adoption events as well. This is possibly because there is simply more of all types of events, since supersubreddits are larger.

Therefore, we wanted to disentangle the effect of different population sizes for subsubreddits and supersubreddits (Norm(U)). In other words, given that more

specific communities have fewer users on average, do they still originate more words per user? In short: they do not. Even normalized for the number of users, there are more first adoptions on supersubreddits, though there are significantly fewer later adoption events. This also suggests that it's not simply the result of there being more total events per user.

Nonetheless, we examine the results normalized for the total number of comments, in case there is a non-linear effect of more users on more comments (Norm(C)). Indeed, there is, and in this analysis, the effect reverses. While the numbers are small and still fairly close, there's a robust effect where subsubreddits now have both more first adoptions and later adoptions. This means that every comment is more likely to have a first or later adoption event.

3 Discussion

The story on the dispersion of new words in online communities we present is somewhat complicated by the different direction of the effects based on the type of normalization. We think it can still be disentangled. It relies on two basic social principles. The first is that more people leads to more diverse, varied, and rich conversation. However, specialized subcommunities, such as those in the subsubreddits, have very focused conversation about specific topics.

As we can see from the results, the supersubreddits have more first adoption events in the majority of the analyses we ran. There are likely some social phenomena at play here: with more people, conversation is more varied. In this sense, in Reddit at least, adding people has a greater than linear increase in the amount of conversation. Indeed, when it comes to brainstorming more generally, which word origination could be a subset of, larger groups come up with more and better ideas [5]. Furthermore, electronic brainstorming does not suffer from production blocking with large groups [5]. In Reddit, where communication is possibly asynchronous and does not rely on a shared communication channel, we could expect that effect to be enhanced.

A possible explanation for that is that varied conversation lowers the chance that new content is considered controversial. As the subreddit's topic is broader, the range of allowable discussion topics is also broader. Alternatively, it could be that these communities rely on more centralized sources, perhaps due to the up-vote system. This would be in line with the idea that centralization causes faster dispersion [10]. On the other hand, in a more specific community, reaching outside the norm may cause disagreement in the community.

In fact, this could extend to the point where discussion in these communities, rather than be broadened, is extensively narrowed. This leads to a very small amount of focused discussion. Thus, even though fewer new words are adopted, there are more new words adopted when normalized for the number of comments. Each comment is more likely to contain a new word, due in part to the small number of new comments.

Indeed, other irregularities have been observed around the effects of large group size before, such as cooperation in social dilemmas [1]. In certain situa-

tions, large group size facilitates cooperation, while in others, it seems to hinder it. It is possible that our events are likewise split into two categories: for instance, word originations and first adoptions of external words. Nonetheless, definitively answering that question is beyond the scope of this paper.

4 Conclusion

The adoption and dispersion of new words have been studied from the perspective of social networks, small groups, and cultures, but more rarely at the level of small but overlapping communities. While studies with naming games have suggested smaller groups are more productive, social network analysis has suggested centralization is faster at innovation. We now provide evidence that both of these observations can generalize to the level of small communities, depending by what scale you measure it, corroborating previous research on both word adoption and cooperation.

References

[1] Hélène Barcelo and Valerio Capraro. "Group size effect on cooperation in one-shot social dilemmas." In: *Scientific Reports* 5.7937 (2015).

[2] Kelly Bergstrom. ""Don't feed the troll": Shutting down debate about community expectations on Reddit. com." In: *First Monday* 16.8 (2011).

[3] Luca Dall'Asta, Andrea Baronchelli, Alain Barrat, and Vittorio Loreto. "Nonequilibrium dynamics of language games on complex networks." In: *Physical Review E* 74.3 (2006), p. 036105.

[4] Nicolas Fay, Simon Garrod, and Leo Roberts. "The fitness and functionality of culturally evolved communication systems." In: *Phil Trans Royal Soc of London B: Biological Sciences* 363.1509 (2008), pp. 3553–3561.

[5] R Brent Gallupe, Alan R Dennis, William H Cooper, Joseph S Valacich, Lana M Bastianutti, and Jay F Nunamaker. "Electronic brainstorming and group size." In: *Academy of Management J* 35.2 (1992), pp. 350–369.

[6] A. Lehrer. "Neologisms." In: *Encyclopedia of Language & Linguistics*. Ed. by Keith Brown. Second Edition. Oxford: Elsevier, 2006, pp. 590–593.

[7] David Reitter and Christian Lebiere. "How groups develop a specialized domain vocabulary: A cognitive multi-agent model." In: *Cognitive Systems Research* 12.2 (2011), pp. 175–185. DOI: 10.1016/j.cogsys.2010.06.005.

[8] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter." In: *Proceedings of the 20th international conference on World Wide Web*. ACM. 2011, pp. 695–704.

[9] R Urbatsch. "Movers as early adopters of linguistic innovation." In: *Journal of Sociolinguistics* 19.3 (2015), pp. 372–390.

[10] Antti Vilpponen, Susanna Winter, and Sanna Sundqvist. "Electronic word-of-mouth in online environments: Exploring referral networks structure and adoption behavior." In: *J of Interactive Advertising* 6.2 (2006), pp. 8–77.