Language innovation and change in on-line social networks

PLEASE CITE THE PUBLISHED VERSION

http://dx.doi.org/10.1145/2700171.2804449

PUBLISHER

© ACM

VERSION

AM (Accepted Manuscript)

REPOSITORY RECORD

Kershaw, Daniel, Matthew Rowe, and Patrick K. Stacey. 2019. “Language Innovation and Change in On-line Social Networks”. figshare. https://hdl.handle.net/2134/22662.

Language Innovation and Change in On-line Social Networks

Daniel Kershaw
Highwire CDT
Lancaster University
d.kershaw1@lancaster.ac.uk

Matthew Rowe
School of Computing and Communication
Lancaster University
m.rowe@lancaster.ac.uk

Patrick Stacey
Managment Science
Lancaster University
p.stacey@lancaster.ac.uk

ABSTRACT

Language is fundamental to human communication, though throughout the course of history language has constantly evolved. This can currently be seen in the changing forms of colloquial language in various on-line social networks (OSN's). These innovations in language are even making it into every day life with the recent inclusion of 'lol' and 'rofl' into modern dictionaries. Changes and varying forms of language pose challenges to both academics and people in business when attempting to asses and communicate with different communities.

In this Ph.D, we aim to forecast online language change through the use of predictive and descriptive methodologies. Through using data sets mined from a number of OSNs, we aim to develop generalizable models and theories for assessing and predicting such language changes. We frame this work in structuration theory will allow for a structured analysis of the agent (user), the social structure and the dynamics between them. We draw on state of the art work and methods, including the development of neural nets to analysis language use and network and community classification to uncover social structures. Preliminary results have identified statistically significant innovations usage across communities across a number of OSN's, this was done though operationlizing known linguistic models of innovation acceptance.

Categories and Subject Descriptors

D.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*
; D.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering, Information filtering*

General Terms

Language Change, Prediction, Modeling

Keywords

OSN, Language, Evolution, Innovation, Change

1. INTRODUCTION

Language is a faculty of human life that people take for granted; it allows for the communications of ideas, thoughts and emotions from one person to another or a group of people. However even within language there is variation, this can be seen through the regional variation in English thought the UK, however these are in constant flux though numerous pressure and constraints in there usage [10].

The aim of this PhD is to answer the following question 'How can one forecast language change in on-line social media, and if so what are the factors that it depends on'. By using the on-line social networks (OSN's) as the medium allows for a in-depth analysis into the patterns of communication between people.

The study of variation in language has transitional been the endeavor of linguistics; famous studies from Lobov showed variation in pronunciations of English across classes in New York [16]. However these studies where time consuming requiring interviews, transcriptions and hand analysis of data; through the use of computers the cost of performing this work has decreased, though there has been a limited investigation into a continuous time series analysis of language change.

This work ultimately looks into language change/evolution; this is a term that not only draws attention to the difference in the states of a language at two points in time, but also gives an in-depth look at which components within the language have altered and the reasons for these alterations. By separating the language change into structural (e.g. grammar and word formation) and none structural components (e.g. content that the language is used in, user latent variables) the term allows for the explanation of linguistic variation that cannot be solely explained by the structure of the language itself [4].

The impact of this work though are not only limited to the academic fields. Social sense making in on-line social media is an ever growing field, though one of the limiting factors is the ever changing nature of language used by different communities on-line, thus by understanding how and when language changes should allow for a greater success rate within the field. Marketers draw on the understanding of the consumers who they are trying to target; successful campaigns in recent times have lead to brand terms embedded into every day language; Google, Facebook and iPhone. The importance of understanding OSN's for marketers has been seen in the implementation and development of a large body of work in understanding and predicting influential users within the network that can aid the dissemination of a message of campaigned.

2. PROBLEM

Ultimately this research is aiming to answer one over arching research question: "How can one forecast language change in on-line social media, and if so what are the factors that it depends on?".

To break this work into three core question we look to Giddens [12]. Though structuation theory he stated that social structure is produced and reproduced through the actions and reaction of agents; thus structure and agency are inextricably linked. The actions of the agents though can be anything, including verbal and written communications. Structure of actions though are not physical constrained, only exists as memory traces within each agent, thus the structure it's self reflexive. As with and action, language is formed by the individual and the social structure the individual sits within, for this reason the over arching question can be looked at though analyse three components.

- The agent
- The social structure
- The interplay which happens between social structure and the agent

Question 1 How can we detect language change in on-line social networks? This question can be seen as focusing purely on the agent in [12] structuration theory; it will be used to explore the individual user's language innovations in on-line communication. This work will initially draw upon theoretical frameworks for the forms of language innovations that were developed by [1], as well as more recent work for the computational detection of these innovations [8].

Question 2 How does community influence language change? In contrast to the first question, this aims to look at the social structure in the system, as opposed to the individual agent. However the methods developed for the first question will be equally useful for this question. Initially it will look into the diffusion of language innovation though varying forms of community structure, social ties and reinforcement; pulling on work such as [20, 21].

Question 3 What is the role of social constructs in language innovation and use within on-line social networks? Social structure and the agent are brought together within this final question. It is aiming to model and understand the dynamics of language innovation in the relationship of the individual and social structure, drawing on Glidden's assertion that: "[a]lthough language only exists in those instances where we speak or write it, people react strongly against others who disregard its rules and conventions [13]"

The question will be used to explore issues of power and solidarity within language and language innovation. It will also be used to identify key influencer's and users that gain greater power in comparison to others, in much the same way as marketers attempt to identify key users in a network to maximise message diffusion. Ultimately it will look at how people change and adapt their language in situations, and how this can be utilized to detect events within the network.

3. STATE OF THE ART

There has been growing interest in studying language change and evolution through the use of computational means. Computational models have shown that traditional language diffusion models (gravity and wave) can be applied to on-line social media data, showing new terms diffusing over the geographical landscape of the USA [11]. This work also identifies correlations between demographic data, geography and language styles. Though the pre-filtering to identify candidate innovations was performed over the whole data set, this then meant that words specialised to smaller communities would have been push out in favour for innovations in larger communities.

Social factors including age and gender have been shown to have a strong influence on communication styles in on-line discourse [19]; age of a user can be predicted though the use of language models such as variation in topics and emoticon usage. Gender of Twitter users was predicted again though the use of emoticons and variations in punctuation [19]. Though again there was limited acknowledgement of the communities of practice, and generalisation of the population as a whole.

It has been showed that though assessing the morphological characteristics of word blends introduced in OSNs means that the source words of the blends can be determined [8]. However it is also the change of meaning that heavily influence language change, large scale semantic changes have been shown in the Google N-gram corpus and social media data sets [15]. Again both these studies generalized to a whole population, without identifying the meaning of works is dependent on the community that is using them.

As mentioned it is not only the individual that changes language, but the interactions and roles within a community that influence the change. Social roles of users within OSNs have been studied in earnest (though not looking at language). Through assessing and automatically classifying interacting patterns within Reddit [5], models were able to predict 'answer' roles within Reddit; and showed that the users roles transcended multiple communities within the network, meaning users maintain the same interaction patterns within different communities and potentially different networks. Though this was limited to highly specialist communities that had highly dynamic interactions on a specific topics. Again through the use of topic specific networks, opinion leaders where identified and assessed for there reach within the network [23] and ignored the dynamics of user roles over time.

As inferred through out this work language change and evolution is dependent on the dynamics of the social network. The dynamics of social network has been shown [22] to highly influence the diffusion and propagation of news and memes thought on-line and offline social networks, with the rate of diffusion being a factor of; time, network structure, randomness and numerous other factors. Through time series and feature based classification one is able to identify and predict the success or failure of meme diffusion though a social network, this was done by identifying communities, and thus the audience size, network structure, and speed of growth. However this only has the ability to detect static meme diffusion, though the use of NLP systems and fuzzy matching the evolution of news reports and options is able to be seen to propagate though social network, showing that blog propagation of news events peaks 2hrs after that of main stream news [17]. Though this was not on a word level, and needed the whole article to identify similar content.

4. PROPOSED APPROACH

The main focus of this work is to forecast language change in OSNs (Twitter, Reddit); this though brings a number of challenges, this section will address the approach that will be taken in performing this research. Each research question will require different approaches, for this reason we list each question and the approach we propose to take.

Question 1: The initial question will be looking at the agent and there usage of innovations. First innovation will need to be identified; by using the BNC (British National Corpus) as a gold standered of the English Language one can infer that if a word is an innovation if it falls outside of the BNC and is composed of all alphanumeric chars. To classify if an innovation is a morphological change a number of methods proposed for identifying innovations such as word blends in OSN's and methods used within text normalisation for abbreviation detection REF. Semantic is a more complicated, this will be assessed in a number of ways; basic semantic changes can be assessed though word correlation and distribution metrics REF.

Question 2: The second question will look at communities acceptance and rejection of innovations; this will use the latent features mentioned above; patterns of innovations will be inferred though the usage of temporal topics models per community, along with morphological features such as charater-grams. Though then modeling survival and diffusion of innovation we aim to show innovations dependency on community and what allows for an innovation to pass though communities.

Question 3: The final question will look identifying dynamics the agent and the community, this will ultimately aim to predict the diffusion of new innovations though a network by looking at the agents that are using the innovations. Though combining identifying influential agents by assessing innovation diffusion paths, the agents communication patters internal and external to the community, along with community adoption patters we aim to predict the speed and range of the diffusion of innovations.

Finally the data that is being currently mined is large (400Gb+ currently) and to process the data there is going to be a number of different stages and tools needed. For this reason large data analytics distributed systems are going to be used. Code will be developed in scalable manner, utilizing know frameworks such as Hadoop and Spark to name a few. Ultimately this will lead to the development of a scalable framework to aid in future research, including tools for network analysis, time series analysis, and NLP at scale.

5. METHODOLOGY

The following section will discusses the methodology that will be used during the process of research as discussed in this work. The methodology applied is that of mix methods, however is post positivist within it epistemology; this will be used to apply theories, build hypothesis, and test the operationalization through the approach mentioned above on the selected data sets.

Within this work the theories of language change and social networks will be drawn upon. These theories will come from the fields of linguistics for the process and pressures of language change and evolutions [9], but also from management science for grounded theory in the formation of social systems though structuation theory [12], and explaining the dynamics of OSNs through the use of social reinforcement [6] and homophiliy [2]. Though the combination of theories from management science on network and social dynamics which known observation of the theories under different circumstances on on-line soial network, such as hash tag [7] and meme [17] propagation. We hypothesise that the grounded theories can be applied to the detection of language innovations in on-line social networks, and the forecast of these innovations. Verification will mainly happen though offline validation, this will be done though apply the same models across social networks and comparing different results. On-line validation could be done on deltas of dictionaries as they are updated, though this could be infrequent and not a reliable method.

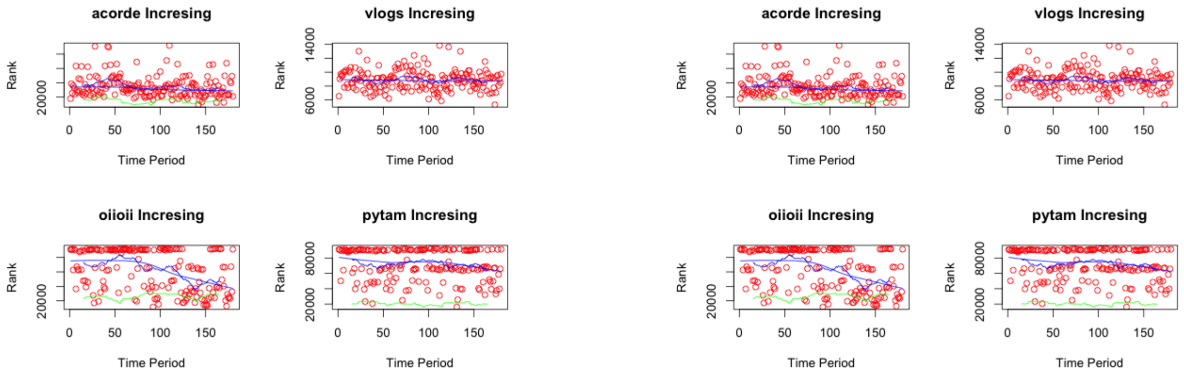
6. RESULTS

The following section will discusses results from research already published, and on going efforts.

Current research being performed looks at attempting to answer questions one and two. This work has applied has applied two widely cited models of language acceptance; Barnhart's VFRGT [3] and Metcalf's FUDGE scale [18] to attempt to classify innovations and accepted innovation. By identifying innovations as words not in the BNC, and detecting statistical significant changes in frequency of these over time has shown variations in innovation patters across two social networks (Reddit and Twitter), though when looking at communities within these networks one can see variations innovations based on geography or intrest based network subreddits. This variation in language innovation and geography was also seen in previous published work [14] which showed variation in language around the consumption of alcohol on twitter.

7. CONCLUSIONS AND FUTURE WORK

In summery we aim to model and forecast language change and innovations within OSN's. Initial analysis and framing of the problem has been done so in a heavily grounded framework, that frames the problem in such as way that allows for a structured analysis of the three components of social interactions; the agent, the network and the interplay of the two. By critiquing state of the art work in relation



(a) 1a

(b) 1b

Figure 1: plots of....

to the three components, has allow for this work to be position in a gap that is novel and relevant within the fields of research. The proposed approach identifies the need large scale pre-processing to identify innovations, along with performing time-series based assessments of the dynamics of user and network. A deductive methodology with on-line and off-line validation is also applied, allowing for conformation of results from though the use of a rigorous method of inquiry.

Future work in answering the questions will be focused around modeling the dynamics of OSNs and how the networks themselves affect the probability of acceptance or rejection of the innovations. This will use time series and interaction analysis, identifying which factors of a network that affect the possibility of that community accepting or rejecting the innovation.

8. ACKNOWLEDGEMENTS

This work is funded by the Digital Economy programme (RCUK Grant EP/G037582/1), which supports the High-Wire Centre for Doctoral Training (`http://highwire.lancaster.ac.uk`).

9. REFERENCES

[1] J. Algeo. Where Do All the New Words Come from? 55(4):264–277, Dec. 1980.

[2] J. Alstott, S. Madnick, and C. Velu. Homophily and the Speed of Social Mobilization: The Effect of Acquired and Ascribed Traits. *arXiv.org*, Apr. 2014.

[3] D. K. Barnhart. A Calculus for New Words. 28(1):132–138, 2007.

[4] R. K. Blot. *Language and Social Identity*. Greenwood Publishing Group, Jan. 2003.

[5] C. Buntain and J. Golbeck. Identifying social roles in reddit using network structure. In *WWW Companion '14: Proceedings of the companion publication of the 23rd international conference on World wide web companion*. International World Wide Web Conferences Steering Committee, Apr. 2014.

[6] D. Centola. The Spread of Behavior in an Online Social Network Experiment. 329(5):1194, Sept. 2010.

[7] H.-C. Chang. A new perspective on Twitter hashtag use: diffusion of innovation theory. In *ASIS&T '10: Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*. American Society for Information Science, Oct. 2010.

[8] C. P. Cook. Exploiting Linguistic Knowledge to Infer Properties of Neologisms, 2010.

[9] W. Croft. *Explaining Language Change*. An Evolutionary Approach. Pearson Education, Jan. 2000.

[10] W. Croft. Mixed languages and acts of identity: An evolutionary approach William Croft. *The mixed language debate: Theoretical and empirical . . .*, 2003.

[11] J. Eisenstein, N. A. Smith, and E. P. Xing. Discovering sociolinguistic associations with structured sparsity. In *HLT '11: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, June 2011.

[12] A. Giddens. *The Giddens Reader*. Stanford University Press, Jan. 1993.

[13] A. Giddens and C. Pierson. Conversations with Anthony Giddens: Making sense of modernity, 1998.

[14] D. Kershaw, M. Rowe, and P. Stacey. Towards tracking and analysing regional alcohol consumption patterns in the UK through the use of social media. *WebSci*, pages 220–228, 2014.

[15] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena. Statistically Significant Detection of Linguistic Change. *arXiv.org*, page 3315, Nov. 2014.

[16] W. Labov. *The social stratification of English in New York city*. Cambridge University Press, 2006.

[17] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506, New York, New York, USA, June 2009. ACM Request Permissions.

[18] A. A. Metcalf. *Predicting New Words*. The Secrets of Their Success. Houghton Mifflin Harcourt, 2004.

[19] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *SMUC '10: Proceedings of the 2nd international workshop on*

Search and mining user-generated contents. ACM Request Permissions, Oct. 2010.

[20] I. Sahin. Detailed Review of Rogers' Diffusion of Innovations Theory and Educational Technology-Related Studies Based on Rogers' Theory. *Online Submission*, 5(2), Mar. 2006.

[21] H. Tajifel. *Differentiation Between Social Groups.* Academic Press, Inc, Jan. 1979.

[22] L. Weng and Y.-Y. Ahn. Predicting Successful Memes using Network and Community Structure. *arXiv.org*, page 6199, Mar. 2014.

[23] Y. Zhao, G. Wang, P. S. Yu, S. Liu, and S. Zhang. Inferring social roles and statuses in social networks. In *KDD '13: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 695, New York, New York, USA, Aug. 2013. ACM Request Permissions.