Towards modelling language innovation acceptance in online social networks

Towards Modelling Language Innovation Acceptance in Online Social Networks

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

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

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

ABSTRACT

Language change and innovation is constant in online and offline communication, and has led to new words entering people's lexicon and even entering modern day dictionaries, with recent additions of *'e-cig'* and *'vape'*. However the manual work required to identify these 'innovations' is both time consuming and subjective. In this work we demonstrate how such innovations in language can be identified across two different OSN's (Online Social Networks) through the operationalisation of known language acceptance models that incorporate relatively simple statistical tests. From grounding our work in language theory, we identified three statistical tests that can be applied - variation in; *frequency*, *form* and *meaning*. Each show different success rates across the two networks (Geo-bound Twitter sample and a sample of Reddit). These tests were also applied to different community levels within the two networks allowing for different innovations to be identified across different community structures over the two networks, for instance: identifying regional variation across Twitter, and variation across groupings of Subreddits, where identified example innovations included *'casualidad'* and *'cym'*.

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, Language Evolution, OSN, Twitter, Reddit

1. INTRODUCTION

Language is a faculty of human life that people take for granted; it allows for the communication of ideas, thoughts and emotions from one person to another or a group of people. Language is necessary yet fragile in that it is in constant flux through numerous pressures and constraints in usage [8]. Communicating through on-line mediums has become dominant in recent times; this itself adds extra pressure on the language being used to communicate. This pressure ultimately comes from the merging of space and time in the nature of on-line communication, i.e. having to communicate what would have been verbal and visual in-situ cues though written text in a time dependent nature [10]. This has led to an explosion of innovative and evolutionary language use in order to allow the user to communicate effectively though the medium [12].

The various forms of change seen can be referred to as language change/evolution; these are terms that not only draw 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 language change into structural (e.g. grammar and word formation) and non-structural components (e.g. the context that the language is used in) the term allows for the explanation of linguistic variation that cannot be solely explained by the structure of the language itself [5]. Such changes are therefore not so-called "free" variations but may, in fact, be correlated with extra-linguistic social features, such as social class, age, and gender. The fields of machine learning and data mining have used language (written and spoken forms) to develop and enhance many systems. However language is usually treated as a static entity, with a limited acknowledgement of the evolutionary dynamics of language. By treating language as a static entity generalisable systems are produced; however treating it as static incurs technical debt within systems through the need for retraining or modifications [29].

Assessing language evolution poses a number of challenges both from a technical perspective and theoretical perspective due to large numbers of variables that could be classified as 'language evolution', e.g. morphological and syntactic change, and variations within different communities [30]. In this paper we investigate the concept of language change and innovation in on-line social networks; this is done through grounding our work in language innovation acceptance models, together with the use of data mining techniques such as word embedding and statistical tests. To ultimately understand language change we need to examine that change over time, and what factors influence it. However this is itself challenging due to the scale of data being published and the myriad ways that language can evolve and be influenced.

WSDM'16, February 22–25, 2016, San Francisco, CA, USA.

ACM 978-1-4503-3716-8/16/02 ...$15.00.
DOI: http://dx.doi.org/10.1145/2835776.2835784 .

Thus, in this work we propose a large-scale computational model that enables language innovation to be tracked over time. This is done through applying known linguistic acceptance models to aid in the classification of accepted language innovations.
The contributions of our work are as follows:

- **Word innovation acceptance models through computation means**: An approach to the operationalization of accepted linguistical innovation acceptance models from Metcalf and Barnhart [24, 3].
- **Identification of local and global acceptance**: Showing that the computational automation of the models allows for the acceptance of an innovation to be seen within large datasets.
- **Multiple network analysis**: Language innovation and acceptance can be seen through two different datasets showing that innovation and acceptance is dependent on the community.
- **Large Corpus Analysis**: Shows the ability to perform large textual analysis with minimal pre-processing and filtering of the initial dataset; through the development of a open source scalable framework (See Section 9 for data release information)

The impact of this work is not limited to academic fields. Marketers draw on the understanding of consumers whom they are trying to target; 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 during a campaign. Hence, by understanding how language emerges and changes, one can identify key individuals who contribute to and shape such adoption.

The rest of the paper is structured as follows: Section 2 highlights state-of-the-art work in linguistic innovation detection and diffusion through computational means, along with uses of language as a feature in OSNs. Section 3 describes the linguistic innovation acceptance models that influence the process of this work. Section 4 introduces the OSNs that will be used within this work, along with the models that are developed through the influence of the language innovation models in the previous section, and Section 5 details the challenges in processing such large datasets. Section 6 presents our experiments following the application of our computational framework to identify language acceptance over the Twitter and Reddit datasets. Finally, Section 7 and 8 critiques the methods that have been applied, highlighting whether these have been effective in identifying potential innovation acceptance within the chosen datasets.

2. RELATED WORK

The following section identifies related work across the fields of NLP and Social Media analysis. All of which have seen a growing interest in assessing language change, down in part to the fact that fine-grained data is now available with ease for researchers and that increases in computing power has allowed for vast volumes to be analysed.

Through the use of stochastic sampling computational models [13] 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. Previous investigations [14] identified correlations between demographic data, geography and language styles; again through computer modelling. Though pre-filtering to identify candidate innovations was performed over the whole dataset, this then meant that words specialised to smaller communities would have been pushed out in favour of innovations in larger communities. Thus the results would have only been valid for potentially dominant communities.

Social factors including age and gender have been shown to have a strong influence on communication styles in on-line discourse; age and gender of a user can be predicted through the use of machine learning systems that have been trained using features such as variation in topics and emoticon usage [28]. Though again there was limited acknowledgement of the communities of practice, and generalisation of the population as a whole.

By assessing the morphological characteristics of word blends introduced in OSNs means that the source words were determined along with the respective definitions of the innovation [7]. Though it is also the change of meaning that heavily influences language evolution, through the use of neural nets and deep learning, large scale semantic changes have also been shown in the Google N-gram corpus and social media datasets [19]. Again both these studies generalised to a whole population, without identifying that the meaning of words 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 interaction patterns within Reddit [6], models were able to predict 'answer' roles within Reddit; and showed that user roles transcended multiple communities within the network, meaning that users maintain the same interaction patterns within different communities and potentially different networks. However, this was limited to highly specialist communities that had highly dynamic interactions on specific topics. Again, through the use of topic-specific networks, opinion leaders were identified and assessed for their reach within the network [35] and ignored the dynamics of user roles over time.

As inferred throughout our work, and as the reader will see below, language change and evolution is dependent on the dynamics of the social network. The dynamics of a social network have been shown to highly influence the diffusion and propagation of news and memes through 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 through a social network, this has been done by identifying communities, and thus the audience size, network structure, and speed of growth [31]. However this only has the ability to detect static meme diffusion, through the use of NLP systems and fuzzy matching, the evolution of news reports and opinions can be seen to propagate through social networks, showing that blog propagation of news events peaks 2hrs after that of main stream

news [22]. However, this was not on a word level, and needed the whole article to identify similar content.

There is a vast amount of research that covers aspects of language change and evolution, though each has its limitations - from limited sampling methods, over-simplified assumptions of communities within a network, and limits if using fuzzy matching to detect diffusion. In this paper we investigate, for the first time, how language evolves at both the global level and the community level. We do this through the innovative operationalisation of language acceptance models in a computational framework.

3. LANGUAGE INNOVATION ACCEPTANCE

The following section explains the grounded models from within the field of linguistics and lexicography that form the basis of computational methods developed in Section 4.3. Original studies from within linguistics looked at identifying the variations in pronunciations within New York [20], though recently the ability to access Google's Historic N-Gram dataset research have shown changes in meanings of words over time [19]. With the introduction of on-line communication there has been a multitude of innovation, from shortening '*your*' to '*ur*', combining words, and creating new abbreviations i.e. *bae*; to name a few; which come under the definition of 'innovation' and 'evolution' [10].

However the existence of an innovation does not automatically mean it is 'accepted' into a language: for it to be 'accepted' it must be acknowledged by the community that use the innovation. Thus acceptance of an innovation is defined by a community within the context of the community, ultimately drawing in the dynamics of the community and the agent as the definition of acceptance [9]. This can be related back to structuration theory where the agent and communication between the agent and community is in constant flux, being defined and redefined within each context [16].

One of the aims of linguists and lexicographers is to understand language at a point in time; this can be seen through the existence and modification of dictionaries, however a fundamental issue is how to decide when a word can be added to a dictionary. For this reason a number of models have been developed to aid the decision process; two that are widely cited are Barnhart's VFRGT [3] and Metcalf's FUDGE scale [24]. Both were developed as tools to assess and predict if new innovations introduced into a community would be maintained or lost. However both were developed to be used by lexicographers through a scoring method: hence the scoring was at the discretion of the scorer, and therefore subjective.

Metcalf stated that for a word to be accepted into a language it must first fulfil 5 points that are measured on a scale of 0 to 2, where the higher the total score the higher the probability of the word being accepted into the language:

- (F) Frequency of the word
- (U) Unobtrusiveness of the word - the word should not be used for an exotic reason
- (D) Diversity of users and situations - the variation of situations and users using the innovation
- (G) Generation of other forms and meaning - if the word starts influencing other innovations then those words have an increased chance of *success*
- (E) Endurance of the concept to which the word refers - in reference to historic meanings of the word

Barnhart proposed a similar measure for evaluating new words; again identifying time and frequency as key components, but also including the number of forms that use the word, and number of genres:

- (V) Number of forms, including variation in spelling and/or derived forms
- (F) Frequency of the word
- (R) Number of sources: e.g. newspaper, magazine, books
- (G) Number of genres the word is used within - news, poetry, spoken, blogs
- (T) Time span of the word

Both measures have time and frequency at the core of the acceptance model, though there are variations on other variables. The morphological form of the word is identified in both but for different reasons: Metcalf is concerned with the word being unobtrusive, i.e. *does it look like existing words?*; whereas Barnhart is concerned with the morphology compared to other innovations. Metcarf also identifies user situation as a key component where as Barnhart separates out source and genre.

Models such as the ones mentioned have been used in a number of studies to identify the acceptance of new words. [21] used Metcalf's scale to identify the acceptance of novel Chinese verbs, showing that they have an accuracy of 60% for predicting the acceptance of new novel verbs. One of the limitation of scoring is that it is at the discretion of the scorer, i.e. a lexicographer with intimate knowledge of the language may have have a better understanding and may give different results .

4. METHODS

In the following section we present the methods used to operationlise the theory discussed in Section 3, along with introducing the datasets that experiments were carried out upon. The methods draw prominently from the fields of data mining and NLP, but also due to the size of the data being processed from distributed computing and big data systems also.

4.1 Datasets

For this study two datasets (Twitter and Reddit) were used to investigate language innovation and acceptance. The two networks were initially selected due to their varying network structure and user dynamics. Both OSN's display fast-paced content generation, dynamic user interaction and ease of sampling a large proportion to the whole network due to the networks' public nature.

Twitter is an OSN based around the submission of short (140 chars) messages; either in response to another person's message or to broadcast a message to the followers. Twitter was sampled through the use of the Twitter streaming API;[1] this allowed for automatic delivery of up to 10% of the Twitter fire-hose.

Query strings were applied to the initial set up to limit the sample under certain constraints: a geo-location filter was

[1] `https://api.twitter.com`

applied limiting tweets to be delivered only if they contained origin coordinates from within the UK.

Table 1: Dataset Statistics

Word	Reddit	Twitter
Users	3,108,844	1,696,630
Posts	73,528,954	111,067,539
Communities	22,055	3,052
Words	15,413,783	7,304,896
Innovations	62,414	42,937
Time periods	28 weeks	37 weeks

The second dataset sampled was Reddit. Reddit is an online news and entertainment community, it is self-organised into self-regulating communities called subreddits; these generally have overarching topics such as 'Personal Finance' or 'Conservative'. In response to submitted posts users can submit comments; these comments can either be direct comments on the original post or comment on other comments. There are other features such as voting on posts and comments, though this goes out of the scope of the paper [32].

The reasons for using Reddit are that *a*) the user base is highly active,[2] *b*) the popularity of the site is relatively high [11] (though this is skewed to a younger demographic, much the same way as Twitter) and, *c*) the majority of comments and posts on the site are public .

In previous research the sampling method for Reddit focused on mining the 1000 most recent comments highly active users, thus creating a user focused dataset [26]. For this research a different sampling method was used focusing on getting a representative dataset of active engagement on subreddits. The stream of most recent comments were scraped,[3] leading to the whole set of comments on the given thread to be downloaded (the number of comments per post could range from one, to many thousands).

4.2 Data Groupings

A given dataset can be grouped in to two dimensions: time and community membership. The following section aims to explain how the grouping of the data is performed.

4.2.1 Time Grouping

To group the data by time a function $weekofyear(e)$ returns the week the Tweet or Reddit post was created on, this is based off the creation time of the data point. We define time groupings as deriving a set as follows:

$$E_k = \{e : weekofyear(e) = k, e \in E\} \quad (1)$$

Where k is the number of weeks since the first item collected within each dataset. and E is the set of all entities (Reddit posts or Twitter tweets).

4.2.2 Community Detection

The datasets (Reddit and Twitter) pose two separate challenges due to the nature of the networks; geography bound (Twitter) and interest bound (Reddit); the aim is to associate a post to a community.

For the Reddit dataset community membership of a post was inferred through decomposition of the Reddit graph into *meta-interest communities*. This was performed through the use of backbone network decomposition; by assessing if the edge between two subreddits is statistical significant based on number of users that comment on both subreddits [26]; by applying an α cut of 0.05 for the significance between nodes allowed for the graph to adhere to a power law distribution [26]. The resulting communities of related subreddits were computed through the use of the Louvain community detection algorithm [4]. This resulted in the dataset being broken down into on three community levels; local (the subreddit), regional (collection of subreddits) and global (all subreddits).

As the Twitter dataset was geographically bound from within the UK this meant that Tweets could be clustered through the use of the longitude and latitude associated with each tweet. There are four Geo-location groups within the UK; National, Regional, Post Code District and Post Code, e.g. a tweet from Post Code LA1 would appear in the LA1 set, LA set, which is itself part of the North West set, which is in turn part of the national set. To compute this a kd-tree data structure was implemented in Java that allowed for quick nearest-neighbour look-up [23]; this was used to find the shortest distance between a tweet and the centroid of a postcode.

Comparisons can be taken between the varying community definitions across Reddit and Twitter. One could say that a low-level community defined by a postcode *LA1* could be compared to a subreddit (the lowest community in Reddit), potentially containing a greater convergence on topic and language used; whereas a higher level community could be classed as showing the 'general' patterns that are global understood across all sub communities.

To group data into their relevant communities a function $community(e)$ that returns the set of entities in a community r:

$$E_r = \{e : community(e) = r, e \in E\} \quad (2)$$

Where $e \in E$ is a given entity (Tweet or Reddit post) in the set of entities (E), and R is the set of all possible communities for the given dataset, for which r belongs to.

4.3 Operationalisation

This research aims to operationalise the acceptance models proposed by Metcalf and Barnhart (Section 3) to show the existence of language change within OSN's, thus creating an equivalent computational model that maintains the properties and heuristics proposed in both. Operationalising FUDGE and VFRGT models in a technical sense means identifying a number of variables that are believed to subsume the properties of each heuristic variable proposed in each model. Though some of this work has been been done before; through the use of normalised frequencies or words in a dataset in replace of frequency, the majority of the metrics are novel.

4.3.1 Variation in Frequency

As stated in both FUDGE and VERGH, frequency of innovations is a core indication that the word could have been accepted into language.

Using the relative frequency as a proxy for popularity is becoming a standard analysis in much of linguistic research, this can be seen in trend detection on Twitter [18, 13]. Even though this is simple it can show insights into what is happening within a dataset in a relatively easy manner; however

[2] `http://www.reddit.com/about/`
[3] `http://www.reddit.com/new/`

it can potentially be misleading if used as the only form of analysis.

Through the use of variation in frequency we aim to identify a statistically significant change in the frequency of innovations; this will allow for the identification of a potential activation point where innovations have gained the attention/adoption of the population. For this we use a uni-gram language model, for each time period (t). For a given time period (t) the probability of a word (w) being used is proportional to the whole dataset within the same time period.

$$T(w,t) = \frac{|w \in C_t|}{|C_t|} \tag{3}$$

Where C_t is a bag of words at time t, $|C_t|$ is the size of the corpus for time period (t) and $|w \in C_t|$ is the frequency of the word within the same time period.

To apply this to all time periods in the data set we apply the function $T(w,t)$ to each time period t.

$$\tau_w^f = \{T(w,t) : t = [0,...,n]\} \tag{4}$$

4.3.2 Diversity of form

As identified by both Metcalf and Barnhart variation in form of a word is key to acceptance as this shows that the innovation may have entered people's vocabulary and that they feel people will understand concepts conveyed through varying the morphological form of the word. The variation in form can take a number of modes; from dropping a letter, to attaching a prefix or suffix in written communication or variation in tone when conveying the word.

For simplicity of this work variation is assessed as the probability of prefix and suffix addition to an innovation. This allows one to see if there is morphological variation in the term. Two lists of common prefix and suffix were taken from the (OED) Oxford English Dictionary, these included; 'ing', 'homo' and 'hetro'. As with the previous feature we aim to generate a time signal for the probability of prefix (τ_w^P) and suffix (τ_w^S). Where P and S are the list of prefix and suffix respectively:

$$MP(w,t,P) = \frac{\sum_p^P |beginswith(w,p,C_t)|}{|C_t|} \tag{5}$$

And the same is applied for suffixes:

$$MS(w,t,S) = \frac{\sum_s^S |endswith(w,s,C_t)|}{|C_t|} \tag{6}$$

Where MS and MP are functions that take a word w, at a time period t and a list of prefixes P or a list of suffixes respectively S, and produces the probability of prefix or suffix addition. The functions $startswith(w,p)$ and $endwith(w,s)$ are both indicator functions that return a 1 or 0. To then convert this into a time series τ_P^w and τ_S^w one on applies the function to each time period in the dataset, such as:

$$\tau_w^P = \{MP(w,t,P) : t = [0,...,n]\} \tag{7}$$

and

$$\tau_w^S = \{MS(w,t,S) : t = [0,...,n]\} \tag{8}$$

4.3.3 Diversity of meaning

The final measure aims to combine *generation of other form and meaning* and *endurance of meaning*. As stated in Section 3 for innovations to enter into general usage they cannot be too specific and their meaning has to be diverse but also common. A number of ways have been developed that allow for analysis of how 'similar' the meanings of the words are, including WordNet [25]. This is a large linguistic database which represents the synonyms of words within a graph structure. Such systems like WordNet though cannot be used within this type of research as it requires the innovation to be within the database which by definition an innovation would not be. For this reason we propose a new method that relies on word co-occurrence and embedding of the word within its own dataset.

Recently within the fields of data mining and NLP there has been growing work using neural-net based techniques for learning the vector representations of words; these have been used for a number of tasks such as POS tagging and Machine Translations [34]. [1] proposed `word2vec`, an unsupervised method of learning the embedded dimensions of word vectors by maximising the likelihood that words are predicted from their context and vice versa. The datasets used within the work allows for both time series analysis and clustering of data on community membership, and allows for the assessment of word embeddings across communities and across time periods. We propose that for each time period (t) we apply `word2vec` to each community (c), this then provides the word embedding model ($W2V_t^c$). Then for each innovation we query the model ($W2V_t^c$) for the top 100 'similar' words, and from this the Jaccard Similarity Index (JSI) is computed as the similarity between communities. Through averaging out the similarity measure per time series, we produce a time signal (τ_w^{w2v}) which indicates the evolution of meaning of the word across communities. We define $JSI(A,B)$ as a function that takes 2 sets of words and computes the Jaccard Similarity Score:

$$JSI(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

And $W2V(w_i, 100)$ is defined to return the top-100 vector similar words for a given input word:

$$S_t^{w,r} = function(r,w,t,K) \tag{10}$$

Where $S_t^{w,r}$ is the vector of words (of length K) for the given period (t) for the community (r) for a given word (w). Therefore to compute the average JSI (Jaccard Similarity Index) for a given time period (t) a Cartesian product is taken across all communities (R) with the JSI computed for each combination, and then divided by the number of communities ($|R|$):

$$f(w,t,R) = \frac{\sum_{i,v \in R, i \neq v} JSI(s_t^{w,v}, s_t^{w,i})}{|R|\,(|R|-1)} \tag{11}$$

The aim here is compute values that show a similarity between communities while still showing variation. If the value is near 0 then it could mean that the word is too diverse for general usage (i.e. too colloquial), while a word with a value near 1 would potentially indicate that it is too specific.

To turn this into a time series one would as in the previous methods apply it to each time period in turn:

$$\tau_w^{w2v} = \{f(w,t,R) : t = [0,...,n]\} \tag{12}$$

4.3.4 Increase/Decrease Classification

The aim of this work is to identify statistically significant changes in the usage of word innovations. The initial presumption taken is that language change (e.g. change in frequency) happens in a monotonic fashion; thus by applying a Spearman's Rank to each words' time series we were able to determine a metric for the increase or decrease in popularity (the resulting distribution was normally distributed). Finally to find innovations that had changed significantly a sample was taken from outside the higher and lower 95% confidence intervals as follows: if a word appeared above the upper confidence interval then it was classified as *increasing* whereas if it appeared beneath the lower confidence interval then it was classified as *decreasing*.

To compute the spearmans value (ρ_w) for each word (w) we take the vector time series τ_w (where τ_w is modified for the given metric computed e.g. τ_w^P indicating the time series for the prefix addition of word w). We define ρ_w as the Spearman's Rank (though again it is modified for the given metric e.g. ρ_w^P for prefix addition) correlation coefficient between the time series vector of the word and t, the ordered vector of weeks since the start of the data collection:

$$\rho_w = SpearmansCorrelation(\tau_w, t) \tag{13}$$

4.3.5 Limitations

The three method proposed though do not cover all the categories proposed through the VFRGT and FUDGE frameworks, this was due to the aim of the research looking into the ability to perform an acceptance mode. The missing categories will be covered in future work such as analysis by genre as we feel the topic requires a more concentrated exploration.

5. COMPUTATIONAL METHODS

The following section describes the computational methods used to implement the metrics on the datasets collected.

5.1 Technical Setup

As with most research analysing social media, the size of the data is a constraining factor. A number of systems were evaluated, such as Hadoop and equivalents, and ultimately Spark [33] was chosen as the processing engine - this allowed for a more interactive interrogation of the data. The specialised tools used throughout our work are as follows:

- TwitterNLP [27]: was used to tokenize tweets with its specialist tokonizer.
- Spark 1.3.1 [33]: allowed for high performance in memory processing of all datasets.

5.2 Pre-Processing

As previously stated the datasets where chosen due to them having different characteristics, for this reason each needs to go through a pre-processing pipeline to allow for analysis using a standardised data format. The initial textual data contained a lot of noise such as hashtags, usernames and HTTP links; through using TwitterNLP's POS tagger [27], we were able to identify tokens such as *hashtags* and *mentions*, which were then removed before the analytic stage. A second level of cleansing was also applied: through using regex long pattern repetitions of the same letter were truncated down to just three characters, e.g. *soooooooo* would be normalised to *soo* (as done in [17]).

The premise of this work was to identify the acceptance of innovations, thus a word must be classified as an innovation or not an innovation. To do this a word was classified as an innovation if it did not appear within the BNC (British National Corpus) [2]. The BNC was chosen to be the baseline for British Language as it is one of the most comprehensive studies of British English Language in recent times, taking its sources not only from books, but also newspapers, written communication and oral discourse transcripts.

Though the datasets have been filtered through cleaning and selection there is still a large proportion of words that could be classified as noise, therefore to *counteract* this potential noise we defined a word as an innovation if it had been used at least 100 times across the whole dataset being considered.

6. EXPERIMENT

The following section will apply the computational methods that have been developed in Section 4.3 to the datasets that have been discussed in Section 4.1.

6.1 Variation in Frequency

Frequency as stated is used as a proxy to determine the popularity of a term over a period of time; by combining this with the mined communities allows for an analysis of different growth and decay of innovations across the networks.

Figure 3 shows the growth and decay of innovation across the two networks on a global level, with noticeable differences are apparent; both are normal though Reddit has a lower variance of 0.00558 compared to 0.083585, and is positively skewed compared to Twitter.

Initially sampling was performed on the global level of the network, with Fig 1 and 4 showing the top and bottom 5 Spearman's Rank for the respective datasets. At this high level of the network, one can state that the innovations growing appear to be highly colloquial and potential originating from on-line games e.g. 'scrims' and 'cooldown'. However when analysing the community structure on the level below (Regions) one can see a number of variations across the UK (Table 2): for instance, in the top-5 innovations for both Wales (North and South) we see terms such as 'bootyfull' (see Figure 7) and 'cyw', these are colloquial but are not derived from the computing gaming community.

Even though words such as 'selfi' appear in the top-5 of 'Greater London' (Table 2) it would appear that when contrasting against its normalised frequencies across other regions (Figure 10) that it is more prevalent in the Channel Islands. However this is due to the Channel Islands having a lower population compared to Greater London, and as a result a new word in the Channel Islands has a greater impact compared to Greater London.

Contrasting the networks and community levels using the distributions and terms sampled one can make a number of informed observations pertaining to the dynamics of the networks. Because Reddit is topic-focused and has a more structured community, this could lead to faster innovation uptakes. This contrasts with Twitter, where community is inferred though geographical locations, and where an innovation is exposed to a larger audience - as a result an innovation could have a higher chance of being rejected, as the nature of conversation has to be more diverse and have a

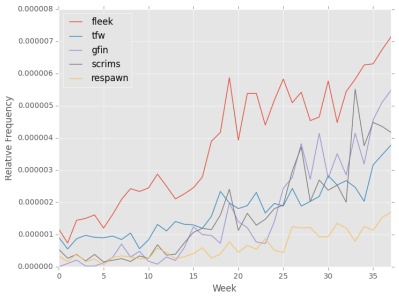

Figure 1: Popularity Increasing - Twitter (Global)

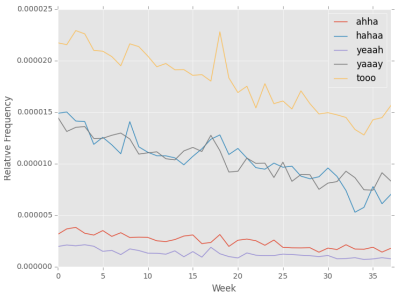

Figure 2: Popularity Decreasing - Twitter (Global)

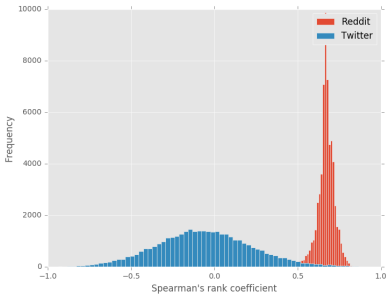

Figure 3: Spearman Rank Distribution - Twitter & Reddit (Global)

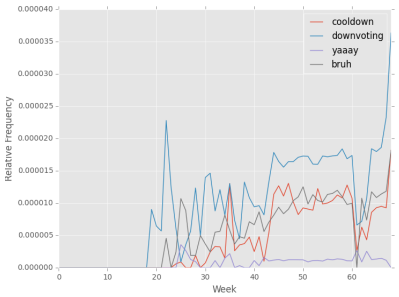

Figure 4: Popularity Increasing - Reddit (Global)

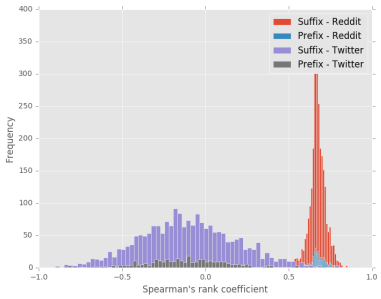

Figure 5: Spearman Rank Dist' (Form Variation) - Twitter & Reddit (Global)

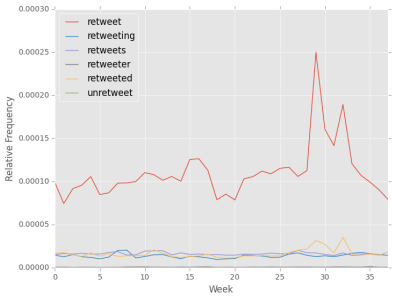

Figure 6: Relative Frequency of form variation of 'retweet' - Twitter (Global)

Table 2: Top 5 innovations per Region (Sample)

Region	Top 5 Words
Channel Islands	mmpressure, cwind, gusthumidity, bisous, capte
Greater London	casualidad, escocs, selfi, trabajando, tambem
Home Counties	cdk, contam, equinix, kaminey, bekah
North West	sitego, ziferblat, alura, dzis, prostu
Northern Ireland	gfin, pentatonix, fcking, mitchs, mphgentle
Scotland (South & Central)	cloudynight, mphrain, cwind, medpace, vuckic
Wales (North)	parklife, inout, torylib, loveshop, cyw (Coming your way)
Wales (South)	lollol, bootyful, gennith, juga, kaspas

Table 3: Twitter (Global) sample increase

Word	P	Definition
fleek	0.940586	"Eyebrows on point","Eyebrows on fleek"
tfw	0.914754	"That Feel When" (Acronym)
scrims	0.869088	"In oonline gaming a scrim is a practice match. A scrim can be any online game, but notable in Counter-strike."
respawn	0.887351	"Also known as spawn, respawn is a gaming term used to describe the action of a coming back to life after being killed."

more diverse audience. Within the top-5 innovation for regions within the UK from Twitter (Table 2), one can see a large proportion of dubious 'innovations' such as *gusthumidity*, this can be put down to Twitter weather stations being potentially introduced during data collection; these devices tweet out the weather on a frequent bases, thus giving the appearance of an innovation increasing over time.

6.2 Variation in Form

A metric for variation in form was computed through averaging the probability per time series of all the varying forms of a given word. The same sampling method was applied as the previous method where a Spearman's Rank was applied across each time series and then ranked.

Looking at the distribution of the two metrics (prefix and suffix) for Twitter and Reddit (Figure 5) shows the similarity in spread for all innovations, however there is a noticeable difference with suffix addition which has a higher frequency compared to prefix additions. By sampling the 95% confidence interval one is able to see examples of words with varying prefix and suffix addition (see Table 5): some of these innovations are timely such as UKIP (as the dataset was mined during a UK general election), whereas the growth in innovations around the word 'vape' appears to be in line with an increase in the UK population taking up the habit [15].

The majority of innovations though were the addition of 's' to the end of a word (73%) followed by 'ing'. 67% of innovations only had one variation in form, as the number

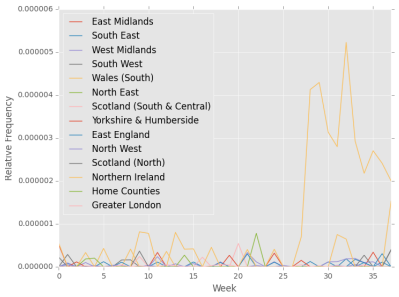

Figure 7: Regional variation in the word 'bootyful'- Twitter (Regional)

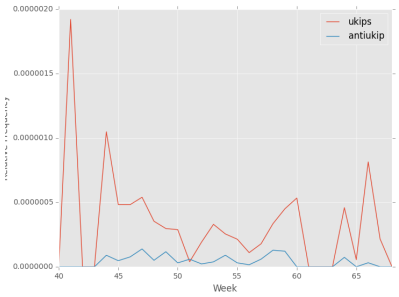

Figure 8: UKIP variation in form - Reddit (Global)

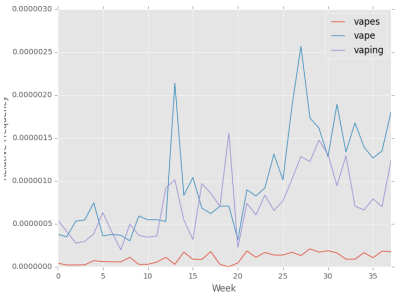

Figure 9: Relative Frequency of form variation of 'vape' - Twitter (Global)

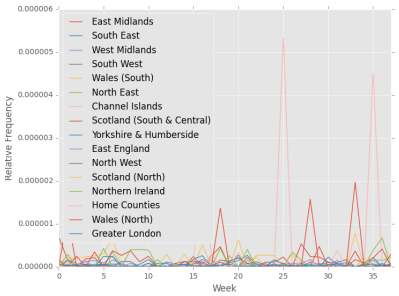

Figure 10: Regional variation in the word 'selfi'- Twitter (Regional)

Table 4: Reddit (Global) sample increase

Word	P	Definition
lamo	0.905310	"Someone possessing the quality of lameness"
bruh	0.869210	"is a variation of the slang term Bro that is often added as an interjection of feigned shock or disappointment"
downvoting	0.826010	"In competitions, the act of voting low other's entries for the purpose of improving oneself's entry"
cooldown	0.818814	"The time required for a spell or action to reset before it can be used again "

of forms increased the innovation appeared to become more specialised such as 'snapchat', 'snapchatted' and 'snapchatting' all referring to the same concept of using the Snapchat app, this trend was also seen for Facebook, and other online/mobile app platforms.

6.3 Variation in Meaning

The final computed feature was variation in meaning, which aimed to assess the coherence in the meaning of a word across the community structure of the two networks. The sampling method for this experiment varied from the previous two; here we assessed the meaning of previously sampled words.

Following initial testing at a week-level of granularity for the time period of each region achieved no results, as each period had insufficient data to learn the embedding. For this reason we modified the time granularity to a month, which allowed for the embedding to be learnt across a greater time period. However even with this modification limited results were achieved again: this ultimately resulted from the sparse words that were sampled from the methods above, where words even though classified as an innovation did not appear across *all* the communities, but when they did they they appeared at a *low* rank and thus the learned embedding, from the `word2vec` function, generated sparse words within the context of the innovation.

This could be seen when assessing the innovation 'fangirl',[4] one would think that this would have a common embedding as can be seen across the Internet, however the embedded dimensions for the innovation in the East Midlands include 'people' and 'swag' where as in East England 'itunes' and 'diesel' potentially words meaning what the areas are 'fangirling' over.

Table 5: Sample of Variation of Form

Network	Word	Example
Twitter	retweet	retweeting, retweets, retweeted, retweeter, unretweet
	vape	valpes
	esport	esports
Reddit	overclock	overclockable, overclocking, overclocker, overclocks, overclocked
	ukip	ukips, antiukip
	facepalm	facepalmes, facepalming, facepalms

7. DISCUSSION

The following section breaks down the four research contributions of this work and discusses how they have been fulfilled, along with the limitations of each.

We have shown that through relatively straightforward statistical and computational models that one is able to identify and determine the potential acceptance of a word into a 'language'. However, it should be stated that while on a global level it appears to be heavily influenced by the innovation of highly active communities, this could be seen in the predominant growth of gaming terms in the Twitter global analysis. This could come from one of two issues with the analysis: the use of frequency-based methods where a

[4]A female fan, especially one who is obsessive about comics, film, music, or science fiction.

sudden spike in usage highly affects the results, along with uneven community sizes across the datasets e.g. Twitter Greater London has a larger set of words than that of the Channel Islands.

Contrasts between the two networks (Reddit and Twitter) showed some interesting differences. These were highlighted on the regional levels, where Twitter's innovations appeared to be bound be the geo-location (e.g. bootyful), and Reddits' bound by the topic of meta subreddit groups (e.g. cooldown).

It should be noted though that this form of research is only sampling the population on-line, thus is not a real-world representation. A representational sample could be achieved through smoothing methods applied to this work, though as language is defined through the community that use the language what is innately being researched is the language of the Internet (Internet Linguistics) and potentially not the language of physical population.

Using Spearman's Rank may not have been the best sampling method for assessing the change in usage of an innovation. It does reveal intriguing insights into innovation growth, however it appeared to favour 'bursty' low-ranking words. Alternate ranking methods could have been used instead of frequency, such as applying a natural rank to each word in a time series, and then performing the experiments over the resulting data, this would have removed some of the 'burstiness' of words as an increase in rank would be linear and not dependent on the size of the corpus (e.g. normalised frequency). However, the measure did not detect innovations that were already present from the beginning of the data collection, only indicating innovations that appeared during the research.

Identification of variation in forms appeared to be successful, as it filtered out words that have only been used once. The produced results appeared to correlate with tacit knowledge of communities on Reddit and events within the UK. However, it may have been better to use a Levenshtein distance measure between words to assess their varying forms, as this would have allowed for a greater detection of none standard usage of the innovation than constraining it to just the use of a fixed list - this will be explored in our future work.

Limited results were achieved for determining a consensus of meaning, this came from the issues of innovations being used sparsely across communities, thus for the majority of regions the word may not have been used means that no embedded meaning could be computed. This could indicate that a word has not reached a global level of acceptance, or that the sampling method as stated before only detects 'bursty' innovations that are too locally used.

8. CONCLUSION

In this work we demonstrated that through the use of relatively simple statistical tests one is able to use known linguistic models to assess language and its change in on-line social networks. We have shown that when the methods are applied to two on-line social networks, they can show variation in innovations usage and persistence; this can be seen in the increase in words such as 'vape' and 'retweet'. We have also shown that these methods can be applied to the individual communities that make up the networks, where we have shown how varying community structure has potentially different language dynamics.

This work has implications further afield than the perceived linguistics and on-line social network analysis, with it having potential value in recommender systems and digital humanities. Further work will look into identifying the dynamics of language innovations within the context of users, along with the influence communities have over language and innovation diffusion.

9. DATA ACCESS STATEMENTS

All data and code created during this research are openly available from Lancaster University data archive at `http://dx.doi.org/10.17635/lancaster/researchdata/46`.

10. ACKNOWLEDGEMENTS

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

11. REFERENCES

[1] Distributional Semantics Resources for Biomedical Text Processing. pages 1–5, Nov. 2013.

[2] G. Aston and L. Burnard. *The BNC handbook: exploring the British National Corpus with SARA*. Capstone, 1998.

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

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, physics.soc-ph(10), Oct. 2008.

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

[6] 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.

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

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

[9] W. Croft. Evolution: Language Use and the Evolution of Languages. *The Language Phenomenon*, (Chapter 5):93–120, 2013.

[10] D. Crystal. *Language and the Internet*. Cambridge University Press, Sept. 2001.

[11] M. Duggan and A. Smith. 6% of online adults are reddit users. *Pew Internet & American Life Project*, 2013.

[12] J. Eisenstein. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, 2013.

[13] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. Mapping the geographical diffusion of new words. *arXiv.org*, page 5268, Oct. 2012.

[14] 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.

[15] S. L. Emery, L. Vera, J. Huang, and G. Szczypka. Wanna know about vaping? Patterns of message exposure, seeking and sharing information about e-cigarettes across media platforms. *Tobacco Control*, 23(Supplement 3):17–25, July 2014.

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

[17] B. Han, P. Cook, and T. Baldwin. Lexical Normalization for Social Media Text. *Acm Transactions on Intelligent Systems and Technology*, 4(1):–27, Jan. 2013.

[18] 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.

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

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

[21] S. L. Lai and V. T. Ng. Collaborative discovery of Chinese neologisms in social media. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 4107–4112. IEEE, 2014.

[22] 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.

[23] S. Maneewongvatana and D. M. Mount. On the Efficiency of Nearest Neighbor Searching with Data Clustered in Lower Dimensions. In *Computational Science — ICCS 2001*, pages 842–851. Springer Berlin Heidelberg, Berlin, Heidelberg, July 2001.

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

[25] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, Nov. 1995.

[26] R. S. Olson and Z. P. Neal. Navigating the massive world of reddit: Using backbone networks to map user interests in social media. *arXiv.org*, page 3387, Dec. 2013.

[27] O. Owoputi, B. O'Connor, C. Dye, K. Gimpel, N. Schneider, and N. A. Smith. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters .

[28] 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.

[29] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young. *Machine Learning: The High-Interest Credit Card of Technical Debt*. 2003.

[30] L. Trask. *Language Change*. Routledge, June 2013.

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

[32] T. Weninger, X. A. Zhu, and J. Han. An exploration of discussion threads in social news sites: A case study of the Reddit community. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 579–583, 2013.

[33] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster Computing with Working Sets . pages 1–7, May 2010.

[34] X. Zhang and Y. LeCun. Text Understanding from Scratch. *arXiv.org*, page 1710, Feb. 2015.

[35] 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.