

Semantic Variation in Online Communities of Practice

Marco Del Tredici and Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam

{m.deltredici|raquel.fernandez}@uva.nl

Abstract

We introduce a framework for quantifying semantic variation of common words in Communities of Practice and in sets of topic-related communities. We show that while some meaning shifts are shared across related communities, others are community-specific, and therefore independent from the discussed topic. We propose such findings as evidence in favour of sociolinguistic theories of socially-driven semantic variation. Results are evaluated using an independent language modelling task. Furthermore, we investigate extralinguistic features and show that factors such as prominence and dissemination of words are related to semantic variation.

1 Introduction

In computational linguistics and NLP, variation in word meaning has mostly been studied in the abstract, as lists of possible word senses (Navigli, 2009; Yarowsky, 2010). In contrast, other neighbouring fields such as sociolinguistics and psycholinguistics have emphasised the link between semantic variation and the activities and interactions of speakers. For example, the psychologist Herbert Clark appeals to the notion of ‘common ground’ to characterise patterns of word usage: “*Word knowledge, properly viewed, divides into what I will call communal lexicons, by which I mean sets of word conventions in individual communities [...] When I meet Ann, she and I must establish as common ground which communities we both belong to simply in order to know what English words we can use with what meaning*” (Clark, 1996); while the sociolinguist Hasan Ruqaiya argues that “*there is evidence of sociosemantic variation*” which must be taken into account “*unless the concept of meaning is arbitrarily constrained*” (Hasan, 1989). The distinction between the two approaches is relevant: while the former is based on the idea that, for a given word, a finite list of discrete senses is available, the latter builds on a more dynamic concept, namely that a new meaning can emerge in any interaction among speakers, who use it in order to make communication more effective.

Understanding the intricate ways in which patterns of word use and communities of individuals are related is essential for characterising the interests and the expressive means of sub-cultures, as well as to develop NLP tools that are effective in the face of variation (Hovy, 2015; Yang and Eisenstein, 2017). In this paper, we study how word meaning (as captured by distributed vector representations) varies across and within different online communities. We take online communities, such as online discussion forums, to be excellent examples of *communities of practice* (Wenger, 2000; Eckert and McConnell-Ginet, 1992), that is, aggregates of individuals not defined by a location or a population, but rather by social engagement in some common endeavour. Using computational modelling techniques and statistical analyses, we show that community-specific conventional meanings of common word forms (as opposed to jargon) do arise and can be reliably detected, which is consistent with the theoretical standpoint of Clark (1996), among others.

The paper makes the following contributions: We adapt a model for geographically located language introduced by Bamman et al. (2014) to learn word representations for different online communities of practice. We introduce a framework for quantifying semantic variation and apply it to several Reddit sub-communities engaged in discussing two broad domains, Football and Programming. We evaluate

our framework extrinsically against a language modelling task, showing that the semantic shifts we detect on common words are strong enough to affect performance.

Our results show that distinct meaning conventions arise within communities engaged in discussing a shared domain, but also that the domain itself is not the only determinant of semantic variation: sub-communities concerned with discussing the same general topic may also develop their own conventional meanings for common words, which supports the view that the main factors driving semantic variation are local accommodation effects presumably arising during interaction.

In addition, our findings indicate that, besides frequency-related factors, the level of dissemination of a word among community members plays a key role in understanding the dynamics of meaning variants.

2 Related Work

The present investigation is related to several strands of research in computational sociolinguistics and historical linguistics. Within the former, a substantial amount of work has used NLP techniques to study correlations between linguistic variables and macro-sociological categories such as age (Nguyen et al., 2013), gender (Nguyen et al., 2014; Burger et al., 2011; Ciot et al., 2013), and other demographic factors (Eisenstein et al., 2014). A related line of research has explored the interplay between language use and social relations among community members. For example, Cassell and Tversky (2005) and Huffaker et al. (2006) investigate the correlation between linguistic features and the strength of the relations among users in newborn communities; Danescu-Niculescu-Mizil et al. (2012) and Noble and Fernández (2015) show how variations in linguistic style can provide information about power differences in social groups. Yet other related work has focused on how acceptance into existing communities is mediated by the adoption of community norms (Nguyen and Rosé, 2011; Tran and Ostendorf, 2016) and on how the process whereby linguistic innovations become norms can be leveraged to predict the permanence of a user in a community (Danescu-Niculescu-Mizil et al., 2013).

Common to all approaches mentioned above is the exploitation of language features to implement predictive models for *non-linguistic* features (such as gender, power differences, or community permanence). Less attention has been paid to investigating linguistic variation in its own right. Those approaches that do address this aspect have concentrated almost exclusively on community-specific jargon and slang, i.e., neologisms, unique acronyms and abbreviations — e.g., ‘dx’ for ‘diagnosis’ in breast cancer discussion forums (Nguyen and Rosé, 2011) or ‘scrim’ for ‘practice match’ in online gaming (Kershaw et al., 2016). They have therefore ignored the fact that social interaction among speakers often leads to semantic variation of common word forms: that is, word that “*belong to many communal lexicons, though with very different conventional meanings*” (Clark, 1996). In the present study we concentrate on precisely this type of semantic variation.

Our approach takes a synchronic perspective, i.e., we do not look into the temporal dynamics of meaning variation. Nevertheless, in terms of methodology, our work is related to computational historical linguistics. Diachronic meaning change has been studied at different time scales, from a few decades to several centuries. A variety of techniques have been explored: Latent Semantic Analysis (Sagi et al., 2011; Jatowt and Duh, 2014), topic clustering (Wijaya and Yeniterzi, 2011) and dynamic topic modelling (Frermann and Lapata, 2016). More recently, word embeddings (Mikolov et al., 2013) have proved useful for investigating meaning change over time. The most common approach consists in creating independent vector representations for consecutive time spans and then using a transformation matrix to map vectors from one space to another one (Kulkarni et al., 2015; Zhang et al., 2015; Hamilton et al., 2016). Similarly to this strand of research, our work leverages the power of word embeddings, but exploits a different approach originally introduced by Bamman et al. (2014) to account for geographical variation. As we will explain in detail in Section 4, this approach is an extension of the skip-gram vector model (Mikolov et al., 2013) that allows us to learn meaning representations per community that build upon shared representations.

Meaning variation determined by geographical location — including that of Bamman et al. (2014) — has often focused on dialectal varieties in the USA using data from Twitter (Eisenstein et al., 2010;

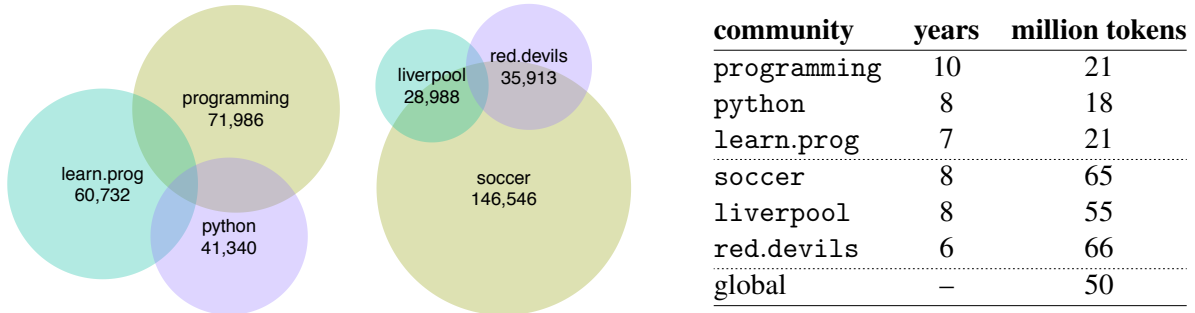


Figure 1: Left: total number of members and their overlap in the Programming and Football supra-communities. Right: main statistics (time span and number of word token) in each community dataset.

Doyle, 2014; Eisenstein et al., 2014). In contrast to this line of work, as pointed out in the introduction, we are interested in investigating semantic variation in *communities of practice* (Wenger, 2000; Eckert and McConnell-Ginet, 1992): communities defined by social engagement rather than geo-location or other demographic variables.

3 Experimental Setup

Online communities offer an unprecedented opportunity to study linguistic variation and its dynamics. For our investigation of semantic variation, we collected data from Reddit, a large on-line community which includes approximately 1 million sub-communities called ‘subreddits’.¹ A subreddit is essentially a discussion forum where individuals with a shared interest on a topic or activity interact: once a user has subscribed to a subreddit, she can post any kind of content (text, links, pictures), reply to existing posts as well as ‘upvote’ or ‘downvote’ them. Subreddits can therefore be considered *communities of practice* in the sense of (Eckert and McConnell-Ginet, 1992).

We collected data from 6 different subreddits: half of them (`programming`, `learn.programming` and `python`)² are concerned with the domain of computer programming, while the other half (`soccer`, `liverpool`, and `red.devils`) are related to the domain of sports, in particular football.³ We refer to the subreddits as *communities* and to the two groups of subreddits related by a common theme, Programming and Football (with a capital), as *supra-communities* or domains. It is important to note that Reddit does not have a hierarchical structure whereby subreddits are classified into groups. We base this grouping on the common theme and on shared membership. The communities `programming` and `soccer` have a somewhat special status as they are more general in terms of topic and larger in terms of number of members. Figure 1 shows the total number of members in each community and the pattern of shared membership within a supra-community. Over 12% and 15% of members within Programming and Football, respectively, belong to at least two communities in the respective domain. The communities within a domain are thus substantially interconnected. In contrast, the Programming and Football supra-communities share less than 2% of users.

For each community, we crawled the contents created by all members during its whole lifespan (between 6 and 10 years). In the present study, we do not make use of the longitudinal character of the corpus, i.e., we abstract away from the temporal aspect and consider each of the community datasets synchronically as a whole.⁴ Since the resulting datasets had different sizes in terms of number of tokens, we randomly subsampled some of them (those crawled from `soccer`, `programming`, and

¹<https://www.reddit.com>

²The actual names of the latter two subreddits are `learnprogramming` and `Python`; we have slightly modified the names for clarity and simplicity.

³The subreddit `liverpool` (actual name `LiverpoolFC`) consists of fans of Liverpool Football Club, while `red.devils` (actual name `reddevils`) groups fans of the Manchester United Football Club.

⁴The temporal information is a valuable feature of the corpus, which we plan to exploit in future work – see Section 7.

learn.programming) in order to make them comparable in size to the other communities within the same domain. The table in Figure 1 summarises the main statistics for each community.

Finally, in order to obtain a sample of community-independent linguistic practices, we created an additional dataset by randomly crawling posts and comments exchanged within any of the existing subreddits during January 2017. We refer to this dataset as the *global* community. The global community includes 50 million tokens from hundreds of thousands of different subreddits, contributed by more than 445k different users. Less than 1% of these users are members of the Programming and Football supra-communities. We consider the linguistic practices present in this dataset as a proxy for general language use.

This experimental setup allows us to investigate different types of semantic variation taking place at different levels: (1) meaning variants deviating from the general language and shared by communities concerned with a common domain, and (2) meaning variants specific to a community and differing both from general language and from other communities within the same domain. In the next section, we define a framework for capturing these two types of semantic shift in a precise, quantitative manner.

4 Framework

We describe the vector-space model we use to learn word representations for online communities of practice and then introduce two indices to measure semantic variation.

4.1 Vector space model

Let C be a set of *communities of practice* and let g denote the *global community*, reflecting general (community-independent) language use. We use subsets such as $D \subset C$ to denote sets of communities related by a certain *domain* (Programming and Football in the experimental setup we use here).

We adapt the model introduced by Bamman et al. (2014) for geo-located language, which in turn is an extension of the skip-gram model by Mikolov et al. (2013). The model relies on a set of contextual variables—geographical locations in the case of Bamman et al. (2014) and Kulkarni et al. (2016), and online communities of practice in our setup. Instead of using a single embedding matrix containing a single real-valued vector for every word in the vocabulary, several matrices are defined: a main matrix W , which is learned by considering all occurrences of each target word in the entire corpus, and one matrix W_c per community (including one matrix W_g for the global community). During training, given an input word w used in a message exchanged within community c , the hidden layer is calculated as the sum $h = w^\top W + w^\top W_c$. Back-propagation via stochastic gradient descent then updates both embedding matrices.

This joint parametrisation has several desirable properties: the model learns different embeddings for the same word (one per community: $w_{c1}, w_{c2}, \dots, w_g$) that are part of the same vector space and therefore can be compared to each other. Furthermore, the word representations share information across communities (via the main matrix W), which, intuitively, operates as a regularizer, thus capturing the intuition that the use of a word in a given community is not radically different from its use in other settings but rather a modulation of conventions built upon general shared common knowledge (Clark, 1996).

We tokenise the datasets described in Section 3 (no further preprocessing is applied), and create two independent vector space models for the Programming and Football supra-communities, respectively. We consider only those words that appear at least 100 times in each community dataset and learn word embeddings with 200 dimensions using L2 regularisation. The global community dataset is used in both models.

4.2 Measures of semantic variation

The model described above allows us to derive word embeddings w_c, w_g for each word w and community c in a domain D , encoding how w is used within that community and in the global community g ,

respectively.

For any two vectors $v, v' \in \mathbb{R}^k$, let $\text{sim}(v, v')$ denote their *cosine similarity*. Given two sets of communities $A, B \subseteq C \cup \{g\}$, we use $\text{Sim}_{A,B}^w$ to refer to the following multiset of similarity values for word w :⁵

$$\text{Sim}_{A,B}^w = \{\text{sim}(w_a, w_b) \mid (a, b) \in A \times B \text{ with } a \neq b\}$$

Let S and S' be two such multisets of similarity values. To measure the extent to which these values are higher in S than in S' , we use the following index, where μ and σ are the mean and the standard deviation, respectively:

$$\mathbb{I}(S, S') = [\mu(S) - \sigma(S)] - [\mu(S') + \sigma(S')]$$

We can now use this generic index to construct several specific indices to quantify different types of semantic variation.

Variation at domain level: We consider that a word w exhibits a domain-specific semantic shift if its meaning is relatively constant across communities with a common domain, while being distinct from its use in the global community. The *domain shift index* $\text{dsi}^w(D)$ captures exactly this, for a given domain $D \subset C$ and word w :

$$\text{dsi}^w(D) = \mathbb{I}(\text{Sim}_{D,D}^w, \text{Sim}_{D,\{g\}}^w)$$

For words with positive **dsi** values, the higher the index, the more pronounced their semantic shift across a domain with respect to the language use of the global community.

Variation at community level: We now want to quantify the degree to which a given word exhibits a semantic shift specific to a community, i.e., not shared by other communities concerned with the same domain D . This type of semantic variation is particularly interesting because, when present, it arguably shows that meaning variants can arise in a community independently from the topic discussed.

In particular, we focus on capturing scenarios where the meaning of a word w in a community $c \in D$ has drifted away from its general use in g , while in other domain-related communities the meaning remains closer to that observed in the global community. This is what the *community shift index* below captures, where $D \setminus \{c\}$ denotes the set of communities in domain D except for c :

$$\text{csi}_c^w(D) = \mathbb{I}(\text{Sim}_{D \setminus \{c\}, \{g\}}^w, \text{Sim}_{\{c\}, \{g\}}^w)$$

Again, for words with positive $\text{csi}_c^w(D)$ values, the higher the index, the stronger the shift in c relative to other domain-related communities.

Using the community-specific word embeddings learned with our vector space model, we compute $\text{dsi}^w(D)$ and $\text{csi}_c^w(D)$ values for all words per domain and community, respectively.

Figure 2 shows the distribution of **dsi** values for the Football domain and the **csi** values of the communities belonging to the domain.⁶ All the distributions present a common pattern: few words undergo a strong semantic shift in the domain / community (left tail of the graph), while the majority of the words present a small or null shift, corresponding to **dsi** / **csi** values included in the range between 0 and 0.2. Note that on average **dsi** values are larger than **csi** ones because, intuitively, the **dsi** captures the shift in the domain vocabulary compared directly to the global community, while **csi** represents the more subtle shifts within communities belonging to the same domain. The right tail of negative values has different interpretations for the domain and the communities. Negative values of **dsi** are assigned to the same words that have high **csi** values, i.e. words that show a strong shift in just one of the communities part of the general domain. Finally, for each community, negative **csi** values are assigned to words that undergo strong semantic shift in *another* community of the same domain.

⁵In practice, in case $A = B$, we only compute one cosine similarity value for every unordered pair rather than for every ordered pair. Observe that this does not affect either the mean or the standard deviation of the multiset.

⁶Similar results are found for the Programming domain and its communities.

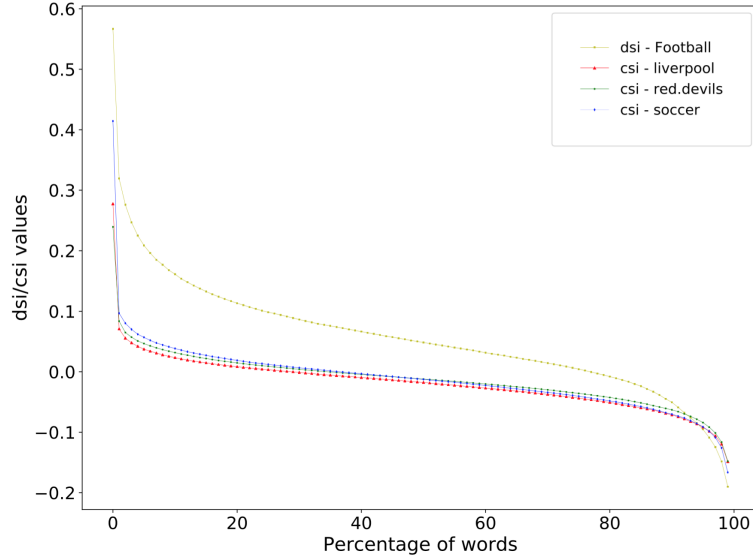


Figure 2: The **dsi** and **csi** values (on the y-axis) for the Football domain and the communities which belong to it. On the x-axis the number of words (in percentage): since we consider only words in the shared vocabulary (see Section 3) the total amount of words is the same for all the communities and for the domain.

5 Evaluation

In order to verify whether the measures proposed in the previous section capture semantic variation that is noticeable beyond cosine distances in semantic space, we evaluate them using an independent language modelling task.

5.1 Method

We implement a neural language model (NLM) using an existing encoder-decoder LSTM⁷ with 2 layers of size 200. We randomly split the dataset of each community into training (70%), validation (15%), and test (15%) sets and train one NLM per community using the word embeddings previously learned for that community with the vector space model described in Section 4.1. We train the models for 40 epochs, using Adam estimation (Kingma and Ba, 2014) for parameter update and dropout for regularisation. The same procedure is also carried out for the global community. All the community language models reached an average test perplexity between 45 and 67 on the task of predicting the upcoming word given the preceding word (window size = 1) — a performance in line with the state of the art, (e.g., Zaremba et al. (2014)).

For each domain D , we define two sets of target words: a `shift` set containing the top 10 words with the highest $\text{dsi}^w(D)$ values, and a `no.shift` set containing the 10 bottom words with the lowest positive $\text{dsi}^w(D)$ values. We do the same per community c : the `shift` set includes the ten words with the highest $\text{csi}_c^w(D)$, while the `no.shift` set includes the ten words with the lowest $\text{csi}_c^w(D)$ per community.

At test time, given a set of target words, we compute the average perplexity for each target word w on predicting $w + 1$ with the original w embeddings used for training (ppl_{train}^w) and with *alternative* embeddings for w learned from another community (ppl_{alt}^w). We then measure change in performance as relative perplexity increase:

$$\text{ppl}_{change}^w = \frac{\text{ppl}_{alt}^w - \text{ppl}_{train}^w}{\text{ppl}_{train}^w}$$

⁷https://github.com/pytorch/examples/tree/master/word_language_model

domain	shift		no.shift	
	$c \rightarrow D \setminus \{c\}$	$c \rightarrow g$	$c \rightarrow D \setminus \{c\}$	$c \rightarrow g$
Programming	6.04	64.9 (*)	5.77	9.02
Football	2.40	40.47 **	-0.78	-1.04
community	$g \rightarrow D \setminus \{c\}$	$g \rightarrow c$	$g \rightarrow D \setminus \{c\}$	$g \rightarrow c$
programming	0.24	3.87	4.92	10.05
python	6.73	26.83 **	0.68	8.83
learn.prog	11.85	56.77 *	9.57	13.28
soccer	11.32	8.92	13.33	11.93
liverpool	2.45	17.70 **	3.84	5.31
red.devils	4.98	55.84 **	2.98	5.60

Table 1: Perplexity increase medians in each setting with significance level of Wilcoxon signed-rank test (***) $p < .001$, (**) $p < .01$, (*) $p < .05$).

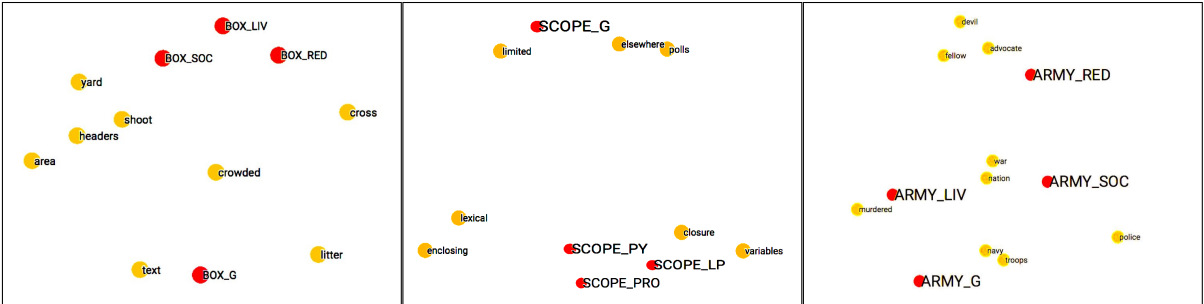


Figure 3: Two-dimensional representation in semantic space of meaning variants for words ‘box’ (high \mathbf{dsi} (Football)), ‘scope’ (high \mathbf{dsi} (Programming)), and ‘army’ (high $\mathbf{csi}_{\text{red.devils}}$ (Football)).

The rationale behind this method is the following: Regarding domain variation, we hypothesise that for *shift* words the increase in perplexity of the NLM of a given community will be significantly higher when testing on alternative embeddings belonging to the general community than on alternative embeddings belonging to another domain-related community. Regarding community-specific variation, we hypothesise that, when leveraging the NML of the global community, using embeddings of *shift* words in community c as alternative embeddings will yield significantly higher perplexity than using alternative embeddings from other communities within the same domain.⁸ In all cases, we expect that for *no.shift* words (i.e., words for which there is no semantic variation according to our indices) the change in perplexity with different embeddings will be negligible.

We evaluate these hypotheses by calculating $\text{ppl}_{\text{change}}^w$ values for *shift* and *no.shift* words and checking for significance with Wilcoxon signed-rank test.

5.2 Results

Table 1 shows an overview of the results. For conciseness, we only show the median $\text{ppl}_{\text{change}}^w$ values. Regarding domain variation, as predicted, for words with low $\mathbf{dsi}^w(D)$ values (*no.shift*), we never observe a significant difference in perplexity when different embeddings are used. In contrast, for words with high $\mathbf{dsi}^w(D)$ values (*shift*) the increase in perplexity is always significantly higher when the original embeddings of a community are substituted with those of the general language ($c \rightarrow g$), while perplexity remains reasonably stable when the alternative embeddings come from another domain-related community ($c \rightarrow D \setminus \{c\}$). This holds for both domains, Football and Programming, with the exception of the *programming* community, for which there is no significant difference in perplexity when

⁸Recall that $\mathbf{csi}_c^w(D)$ is meant to capture a meaning variant of w in c that has drifted away from w ’s use in the global community more than in other domain-related communities.

learn.programming and general language embeddings are used — indicated by (*) in Table 1.

As for community-specific variation, again we never observe a significant difference in perplexity for words with low $\text{csi}_c^w(D)$ values (no.shift). For words with high $\text{csi}_c^w(D)$ values (shift), our hypothesis is confirmed for the more specific communities liverpool, red.devils, python and learn.programming: there is a significant increase in perplexity when the embeddings from these communities are used with the global NLM ($g \rightarrow c$), which is in line with the presence of a community specific semantic variant within the domain. This is not confirmed for the more general communities programming and soccer. This latter negative result is in fact intuitive: it is unlikely that these more general and larger communities will exhibit meaning variants that are further away from general language use than the more specific, smaller communities.

Figure 2 shows some examples of meaning shift captured by our indexes. The words ‘box’ and ‘scope’ are among the ten words with the highest **dsi** for Football and Programming, respectively. As a consequence, the domain-related variants are closely located, while the variant of the general community is farther away in semantic space. Difference in meaning is also evident from the nearest neighbours.⁹ The word ‘army’ has high **csi** in the red.devils community. In the other domain-related communities, the word has meaning variants that are closer to its use in the general community. In the red.devils community, however, ‘army’ is conventionally used to denote the Manchester United fans (e.g., ‘we need all types of supporters to make the red army’), as evidenced by its closest neighbours.

6 Factors Influencing Semantic Variation

Having confirmed that the semantic shift indices proposed in Section 4.2 capture variation that is noticeable in an external language modelling task, we now turn to analysing the factors that may be related to the presence of such variation.

6.1 Features

We consider four features capturing different properties of word *forms* and investigate their effect on meaning variation:

Frequency. It is known that more frequent words have a tendency to be more polysemous (Zipf, 1949), are more semantically stable over time (Hamilton et al., 2016), and evolve at slower rates across languages (Pagel et al., 2007). Word frequency may therefore play a role in semantic variation across communities of practice. We compute word frequency as the log-scaled relative frequency of a word in a given community:

$$\text{Freq}(w, c) = \log_{10}(N_c^w/N_c)$$

where N_c is the total number of words in the sample dataset of community c and N_c^w the number of occurrences of word w in that sample. Frequency in a domain $\text{Freq}(w, D)$ is calculated equivalently.

Prominence. Many measures have been proposed to weight the prominence of a word in a language sample, including TF-IDF. Our choice here is inspired by literature on terminology extraction (Velardi and Sclano, 2007). We compute the prominence of w as its frequency in a community (N_c^w) relative to its frequency in a domain, or as its frequency in a domain (N_D^w) relative to its frequency in general language use (N_g^w):

$$\begin{aligned} \text{Pro}(w, c) &= N_c^w / (N_c^w + N_{D \setminus \{c\}}^w) \\ \text{Pro}(w, D) &= N_D^w / (N_D^w + N_g^w) \end{aligned}$$

⁹In the domain of Football, ‘box’ has come to mean the penalty area, which is associated with game actions such as ‘cross’, ‘shoot’ and ‘headers’.

	Freq	Pro	Spe	Dis
Programming	0.96 ***	0.72 ***	0.46 ***	0.37 ***
Football	1.32 ***	0.81 ***	0.52 ***	0.63 ***
python	0.12	1.0 ***	0.09	0.21 ***
learn.prog	0.02	0.63 ***	0.32 *	0.23 **
liverpool	0.26 **	0.47 **	0.18	0.38 ***
red.devils	0.16	0.42 **	0.16	0.20 *

Table 2: Effect size (Cohen’s d) and unpaired two-sample t -test significance level (***) $p < .001$, (**) $p < .01$, (*) $p < .05$) for each feature.

Community-specific jargon or slang words will typically have very high prominence. In contrast, we hypothesise that common words exhibiting semantic variation as a result of community conventions — which are our focus here — are likely to *not* be singled out by very high prominence values. Nevertheless, their level of prominence may still be a determiner of variation.

Specificity. Besides frequency-related aspects, we also want to capture the extent to which a given word w appears in a restricted set of contexts. We approximate this by computing the collocational score of every bigram containing w and then scoring them using log-likelihood ratio as association measure (Dunning, 1993; Manning and Schütze, 1999).¹⁰ We take the value of the highest ranked bigram as a proxy for the contextual specificity of w in community c ($\text{Spe}(w, c)$) or domain D ($\text{Spe}(w, D)$). The feature values are normalised to obtain scores in the range $[0, 1]$.

Dissemination. Finally, we consider the range of individuals using a given word. A priori, words with the same frequency, prominence, or contextual specificity may differ in their level of social dissemination, i.e., in the proportion of community members using them. We compute a word’s dissemination within a community c as follows:

$$\text{Dis}(w, c) = (U_c^w / U_c) \times (1 - \text{RelFreq}(w, c))$$

where U_c^w is the number of community members who use word w and U_c the total number of members in community c . Since words with very high frequencies (such as function words) will be used across the board, we weight the ratio U_c^w / U_c by the inverse of w ’s relative frequency. Dissemination in a domain $\text{Dis}(w, D)$ is calculated equivalently.

Word dissemination has been shown to be predictive of changes in word frequency over time (Altmann et al., 2011). Here we investigate whether it is a determiner of semantic variation.

6.2 Results

To investigate the role of the features introduced above, we test whether their values are significantly different in words that exhibit a strong semantic shift (words with **dsi** / **csi** values equal or larger than 2 standard deviations above the mean within a domain or community) and words with no semantic variation (with index values lower than one standard deviation above the mean).

At the domain level, we find very robust patterns for all features: the words that have undergone a strong domain shift have significantly higher frequency, prominence, contextual specificity, and social dissemination in each respective domain, Programming and Football. Table 2 shows the significance level of a unpaired two-sample t -test and the effect size for each feature.

At the community level, since the shifts for the programming and soccer communities were not validated in our extrinsic evaluation (Table 1), we do not consider these communities here. For the other 4 communities, we find a systematic pattern: words that exhibit a semantic shift particular to a community

¹⁰We used the NLTK implementation described at <http://www.nltk.org/howto/collocations.html>

are significantly *more* prominent in that community than in other domain-related communities, and *less* disseminated within that community than words that do not exhibit a shift. The significance of frequency and specificity vary per community. A summary is given in Table 2.

6.3 Qualitative analysis

As hypothesised, words with high **dsi** / **csi** values have significantly higher levels of prominence in the respective domain or community, but lower levels than jargon. For example, ‘box’ and ‘believers’, which have high **dsi** in Football and high **csi** in Liverpool, respectively, have prominence values of 0.7, in contrast to jargon terms such as ‘hat-trick’ (Pro=1 in Football) and ‘bitwise’ (Pro=1 in Programming), which are not singled out by our semantic shift indices. This confirms that our measures of semantic variation identify meaning variants of *common* words (such as ‘box’ and ‘believers’) that arise in communities of practice.

From qualitative analysis, we observe that contextual specificity, which is significantly higher in words that exhibit a variant at the domain level, can give rise to different semantic phenomena. For instance, in the case of ‘box’ (see footnote 9), we observe semantic *broadening*, a generalisation of meaning possibly as a consequence of metaphorical use. While in other cases, specificity is related to semantic *narrowing*. This holds, for instance, for ‘yellow’, which has come to mean ‘yellow card’ in the Football domain. The strength of the collocation ‘yellow card’ seems to have made possible a narrower interpretation of ‘yellow’, as in ‘Terry got a very stupid *yellow*’.

In contrast to domain-level variation, specificity and frequency do not play an important role across the board for semantic shift at the community level (see Table 2). Meaning variants that are specific to a particular community are not highly frequent and thus it is less likely that they take part in collocations (see e.g., Shin and Nation (2008)). As mentioned, we find that words with high **csi** values are less disseminated within the community. We see this as potentially related to the general process of linguistic innovation and diffusion described in Chambers and Trudgill (1998) and usually represented by a sigmoid function (see, for example, Fagyal et al. (2010)). Linguistic variants originate among and are initially adopted by a circumscribed number of members. At this stage (corresponding to the left tail of the function) few users use the innovation, which is therefore not highly disseminated in the community. Our intuition is that the **csi** index captures innovations which are in this phase. Some variants may then rapidly spread within the community (central part of the function) and possibly to other domain-related communities, until they reach a plateau, in terms of frequency of use (right tail of the function). This is the stage which is captured by our **dsi** index: the innovation, at this point, has been largely adopted, and, consequently, has a high dissemination value.

It is also possible that some community-specific semantic variants are used as identity markers (e.g., ‘army’ in red.devils or ‘believers’ in liverpool), which are then presumably not likely to spread to other communities. Such uses may be limited to members who are particularly invested in the community and thus not part of other domain-related communities, which may lead to lower dissemination (since different communities within a domain share a substantial number of members, as shown in Figure 1). These speculations, however, need to be verified with further analysis, which we leave to future work.

7 Conclusions

We have investigated meaning variation from the perspective of social engagement in online communities of practice, exploring the hypothesis that meaning conventions are not only topic dependent, but that different meanings can emerge in communities discussing the same topic. We verified our research hypothesis using a large dataset from Reddit discussion forums, and showed that our quantitative measures allow us to identify semantic variation in the use of common (non-slang) words at both domain and community levels. We evaluated our findings using an extrinsic language modelling task.

Our analysis of the factors that influence socially-driven semantic variation should be seen as a preliminary investigation, which we believe opens the door to more in-depth studies we plan to conduct

in the future. The most natural extension of the current work is an investigation of the social dynamics that lead to meaning variation: while in the present work we have shown the outcome of such dynamics, i.e. the observable meaning shift in different communities of practice, in our future work we plan to focus on the interactions among speakers, which are at the base of observable variation. Directly related to this is the consideration of the diachronic dimension, linking the presence of semantic variation to the more general dynamics of meaning change. Our aim in this direction is to consider the evolution of meaning conventions in time while taking into account the network structure of communities of practice.

In parallel, we plan to explore other aspects within the synchronic perspective, such as the relationship between semantic variation and demographic factors, e.g., geo-location, age, or gender — in particular, in light of the fact that the datasets we are using here are likely to be biased towards the language use of male speakers. Finally, we want to extend our investigation to a larger set of communities, in order to make our findings and claims more robust.

References

- Altmann, E. G., J. B. Pierrehumbert, and A. E. Motter (2011). Niche as a determinant of word fate in online groups. *PLoS ONE* 6(5), e19009.
- Bamman, D., C. Dyer, and N. A. Smith (2014, June). Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 828–834.
- Burger, J. D., J. Henderson, G. Kim, and G. Zarrella (2011). Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309. Association for Computational Linguistics.
- Cassell, J. and D. Tversky (2005). The language of online intercultural community formation. *Journal of Computer-Mediated Communication* 10(2), 00–00.
- Chambers, J. K. and P. Trudgill (1998). *Dialectology*. Cambridge University Press.
- Ciot, M., M. Sonderegger, and D. Ruths (2013). Gender inference of twitter users in non-english contexts. In *EMNLP*, pp. 1136–1145.
- Clark, H. H. (1996). *Using language*. Cambridge University Press.
- Danescu-Niculescu-Mizil, C., L. Lee, B. Pang, and J. Kleinberg (2012). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pp. 699–708. ACM.
- Danescu-Niculescu-Mizil, C., R. West, D. Jurafsky, J. Leskovec, and C. Potts (2013). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 307–318. ACM.
- Doyle, G. (2014). Mapping dialectal variation by querying social media. In *EACL*, pp. 98–106.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61–74.
- Eckert, P. and S. McConnell-Ginet (1992). Communities of practice: Where language, gender, and power all live. In K. Hall, M. Bucholtz, and B. Moonwomon (Eds.), *Locating Power, Proceedings of the 1992 Berkeley Women and Language Conference*, pp. 89–99.
- Eisenstein, J., B. O’Connor, N. A. Smith, and E. P. Xing (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287. Association for Computational Linguistics.

- Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing (2014). Diffusion of lexical change in social media. *PloS one* 9(11), e113114.
- Fagyal, Z., S. Swarup, A. M. Escobar, L. Gasser, and K. Lakkaraju (2010). Centers and peripheries: Network roles in language change. *Lingua* 120(8), 2061–2079.
- Frermann, L. and M. Lapata (2016). A bayesian model of diachronic meaning change. *TACL* 4, 31–45.
- Hamilton, W. L., J. Leskovec, and D. Jurafsky (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016*.
- Hasan, R. (1989). Semantic variation and sociolinguistics. *Australian Journal of Linguistics* 9(2), 221–275.
- Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 752–762.
- Huffaker, D., J. Jorgensen, F. Iacobelli, P. Tepper, and J. Cassell (2006). Computational measures for language similarity across time in online communities. In *Proceedings of the HLT-NAACL 2006 workshop on analyzing conversations in text and speech*, pp. 15–22. Association for Computational Linguistics.
- Jatowt, A. and K. Duh (2014). A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 229–238. IEEE Press.
- Kershaw, D., M. Rowe, and P. Stacey (2016). Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp. 553–562. ACM.
- Kingma, D. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kulkarni, V., R. Al-Rfou, B. Perozzi, and S. Skiena (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 625–635. ACM.
- Kulkarni, V., B. Perozzi, and S. Skiena (2016). Freshman or fresher? quantifying the geographic variation of language in online social media. In *ICWSM*, pp. 615–618.
- Manning, C. D. and H. Schütze (1999). *Foundations of statistical natural language processing*. MIT Press.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys* 41(2), 10.
- Nguyen, D. and C. P. Rosé (2011). Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media*, pp. 76–85. Association for Computational Linguistics.
- Nguyen, D., D. Trieschnigg, A. S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F. De Jong (2014). Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1950–1961.

- Nguyen, D.-P., R. Gravel, D. Trieschnigg, and T. Meder (2013). “How old do you think I am?” A study of language and age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI Press.
- Noble, B. and R. Fernández (2015, June). Centre stage: How social network position shapes linguistic coordination. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, Denver, Colorado, pp. 29–38. Association for Computational Linguistics.
- Pagel, M., Q. D. Atkinson, and A. Meade (2007). Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature* 449(7163), 717–720.
- Sagi, E., S. Kaufmann, and B. Clark (2011). Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, 161–183.
- Shin, D. and P. Nation (2008). Beyond single words: The most frequent collocations in spoken english. *ELT journal* 62(4), 339–348.
- Tran, T. and M. Ostendorf (2016). Characterizing the language of online communities and its relation to community reception. *arXiv preprint arXiv:1609.04779*.
- Velardi, P. and F. Sclano (2007). ”termextractor: a web application to learn the common terminology of interest groups and research communities”. In *7th Conference on Terminology and Artificial Intelligence*, pp. 85–94.
- Wenger, E. (2000). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.
- Wijaya, D. T. and R. Yeniterzi (2011). Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversity on the social web*, pp. 35–40. ACM.
- Yang, Y. and J. Eisenstein (2017). Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*.
- Yarowsky, D. (2010). Word sense disambiguation. In *Handbook of Natural Language Processing, Second Edition*, pp. 315–338. Chapman and Hall/CRC.
- Zaremba, W., I. Sutskever, and O. Vinyals (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.
- Zhang, Y., A. Jatowt, S. S. Bhowmick, and K. Tanaka (2015). Omnia mutantur, nihil interit: Connecting past with present by find-ing corresponding terms across time. In *Proc. of ACL*, pp. 645–655.
- Zipf, G. (1949). *Human Behavior and the Principle of Least Effort: an Introduction To Human Ecology*. Addison-Wesley.