"*Come Together!*": Interactions of Language Networks and Multilingual Communities on Twitter

Nabeel Albishry[1]*, Theo Tryfonas[1], and Tom Crick[2]

[1] Department of Computer Science, University of Bristol, UK
{n.albishry,theo.tryfonas}@bristol.ac.uk
[2] Department of Computing, Cardiff Metropolitan University, Cardiff, UK
tcrick@cardiffmet.ac.uk

Abstract. Emerging tools and methodologies are providing insight into the factors that promote the propagation of information in online social networks following significant activities, such as high-profile international social or societal events. This paper presents an extensible approach for analysing how different language communities engage and interact on the social networking platform Twitter via an analysis of the Eurovision Song Contest held in Stockholm, Sweden, in May 2016. By utilising language information from user profiles (N=1,226,959) and status updates (N=7,926,746) to identify and categorise communities, our approach is able to categorise these interactions, as well as construct network graphs to provide further insight on these multilingual communities. The results show that multilingualism is positively correlated with activity whilst negatively correlated with posting in the user's own language.

1 Introduction

Despite the widespread use of Twitter globally – with 328 million monthly active users as of the first quarter of 2017 – little research has investigated the differences amongst users of various languages; there is a tendency to assume that the behaviours of English users generalise to other language users [15]. Language has featured as a facet of research on the geographies of Twitter networks [26], especially whether offline geography still matters in online social networks [18]. Linguistic-inspired studies have been performed on hashtags [11], as well as the volume and proportional of tweets in English and Arabic, as part of an analysis of the Arab Spring [5]. Nevertheless, language is clearly a vital component of affiliation and discourse on the web [30], with the creation and curation of emerging multilingual networks and communities, representing well-established creative and cultural norms, including for minority languages such as Welsh [14], as well as investigations into the economics of linguistic diversity [13].

* This work has been supported by a doctoral research scholarship for Nabeel Albishry from King Abdulaziz University, Kingdom of Saudi Arabia.

In the social network analysis domain, centrality measures such as degrees, betweenness, clustering coefficient, modularity and cliques have been used in various projects to measure influence or detect the emergence of new communities [24, 29]. These measures provide the ability to assess network graphs that are constructed from collected data (for example, tweets). Selection of these centrality measures is dependent on the goal of the analysis; for example, the degree of a node helps to identify nodes with high number of connections within the network [3, 21, 25]. In a representation of a real-world network, this metric may help to identify highly connected persons, such as political leaders, sports stars or celebrities, who are potential "information spreaders" [4, 8, 31].

Clustering users in communities has been an important factor in social networking analysis, with a particular focus on clustering users based on their locations. However, for the sake of anonymity, many users tend not to disclose information about their identity, such as locations [16]; looking at Twitter, it has also been reported in the literature that geotagged tweets are generally low in number [19, 22, 27], the exponential growth in social media over the past decade has been joined by the rise of location as a central organising theme [20] of how users engage with online information services and, more importantly, with each other [1, 7, 10]. The work here examines the correlation between multilingualism of users and their associated activity.

The remainder of this paper is organised as follows: in Section 2 we introduce the methodology and key language themes; Section 4 present the 2016 Eurovision Song Context case study, along with an analysis of the key data and results; Section 5 concludes the paper with a wider discussion and a summary of the potential application of our approach.

2 Methodology

The primary purpose of this study is to identify and define an extensible analytical approach for examining language uses, communities, and diversity on Twitter. The approach is based on network graphs and their properties, such as indegree, outdegree, and edge weights. Graphs are generated from language settings in users' profiles and those for statuses. First, we construct user graphs to analyse interactions and multilingualism at the level of individual users. Then, from the user graphs, we produce language communities graph that groups users based on common languages.

3 Language Entities

To generate the required graphs, we need three essential entities from each status; user ID, user profile language, and status language[3]. Those values can be extracted from *[status]['user']['id']*, *[status]['user']['lang']*, and *[status]['lang']*,

[3] For Twitter, *status* may also be referred to as *post*, or *tweet*.

respectively. It is important to note that the focus of this work is on the analytical approach, not necessarily the accuracy of language detection; therefore we assume that language of tweets are correctly identified via the user profile. For profiles, users are expected to pick a language for their settings. Nevertheless, their language entity may show as the initial placeholder text "*Select Language...*" or a translated version that may provide information to the user's native language community.

3.1 Network Graphs

For this study, we need to generate two different graphs; one is based on individual users and their posting activity, while the other combines users into language communities. In the context of this study, all graphs must be directed to provide correct measurements, as demonstrated in Figure 1.

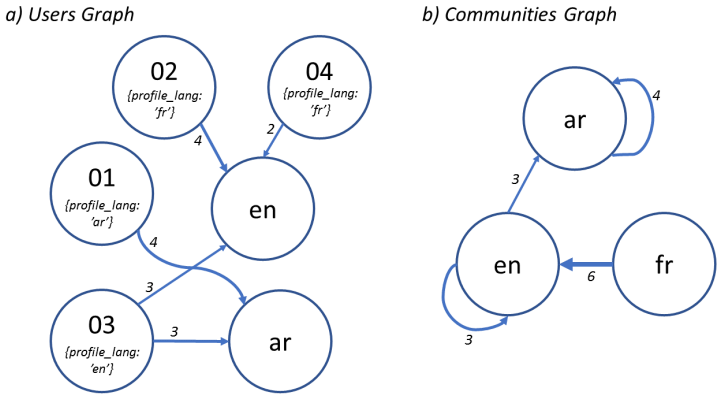

Fig. 1. Examples of simple models of language graphs

User Graph This graph represents the core structure for our analysis. As shown in Figure 1(a), nodes in this graph are of two types; users and posting language. Each posted tweet resulted in two nodes, one represents the user with profile language setting added to the node as the attribute '{*profile_lang:xx*}'. The other node represents language of the tweet. Edges link users with the posting languages they used, and their weight (thickness) measures the number of tweets that have been posted by the user (the starting node) in the target language (ending node). In the example above, the profile language setting for user '*03*' is '*en*', they posted three tweets in '*en*' and three in '*ar*' (Arabic). This graph will be referred to as the *user graph*.

Communities Graph This second graph is derived from the user graph and has one type of node to represent language community, as shown in Figure 1(b).

For each user node we generate one node from the '{*profile_lang:xx*}' attribute, and another node from the posting language to which it is connected. This resulted in combining all users of the same profile language into one node, with edge connecting to posting language and its weight measuring their activity. Theoretically, each tweet results in two language nodes, one for the user profile, and the other for language of the tweet. In our example above, users with '*fr*' (French) profiles have generated six tweets in '*en*'. In the case of '*ar*' node, we can see that users of the profile language as '*ar*' have posted four tweets in '*ar*' only – in graph terminology this is referred to as 'self-loop'; we will refer to this graph as the *communities graph*.

Throughout the paper, we refer to language communities in two ways; *profile community* to perceive the language as user profile settings, whereas *posting community* refers to the language as tweeting settings.

3.2 Measures

In this section, we will discuss how graph measures can be used to make deductions about users, associated community languages, posting language activity, and how different language communities are linked to each other. These measures and their interpretations, in the context of this study are as follows:

- *Indegree*: number of incoming edges;
- *Outdegree*: number of outgoing edges;
- *Edge Weight*: number of tweets on edge;
- *Weighted indegree*: total weights of incoming edges;
- *Weighted outdegree*: total weights of outgoing edges.

User Graph Properties User nodes have *indegree*=0, and posting languages have *outdegree*=0; these two properties will be used to distinguish between nodes. Both outdegree of user nodes and indegree of posting languages must be greater than 0. The edge weight indicates the number of tweets associated with both end nodes. Referring to the example in Figure 1(a), we can see that user '*03*' has indegree of 0 (user identifier), outdegree of 2 (number of languages he used), and weighted outdegree of 6 (total number of tweets posted). Also, in the same figure, we can see that for '*en*' posting language, it has outdegree of 0 (language nodes identifier), indegree of 3 (number of users posted in this language), and weighted indegree of 9 (total number of tweets posted); Table 1 presents main properties of this graph.

	User Language	
Indegree	0	>0
Outdegree	>0	0
Edge Weight	#Tweets	

Table 1. Node properties in user graph

Communities Graph Properties As discussed in Section 3.1, this graph is extracted from the user graph and contains one type of node: language community nodes. Nodes in this graph represent languages as profile language settings, posting language, or both. However, as the graph is directed, we can identify if a community node is for profile or posts by measuring the *indegree* and *outdegree* properties. Positive indegree implies posting language, and positive outdegree indicates profile language settings. Figure 1(b) shows three language communities, two nodes appear as posting and profile nodes, while one node exists as a profile only node. The node '*ar*', for example, has outdegree of 1 and indegree of 2. In other words, at least one user has their profile language settings as '*ar*', and at least two users have posted in '*ar*'. In terms of edge weights, we can say that there are seven tweets posted in '*ar*' language, originated from two different profile language communities. For the '*fr*' node, we can see only outdegree, which means this language community exists as a profile-only node as no user posted in '*fr*'; these measures are summarised in Table 2.

	Community Node
Indegree	posts
Outdegree	profiles
Edge Weight	#Tweets

Table 2. Node properties in communities graph

4 Case Study and Discussion

Next we will explore the analysis of `#Eurovision` datasets from the 2016 Eurovision Song Contest, based on the techniques presented in Section 2. Using the *user graph* and *communities graph*, we conduct analyses on multilingualism, activities and user behaviours in posting in different languages[4].

4.1 Case Study: 2016 Eurovision Song Contest

The 2016 Eurovision Song Contest[5] took place in May in Stockholm, Sweden, with the motto of "*Come Together!*" (as referred to in the title of this paper). There were 32 countries taking part, with two semi-finals taking place on 12 and 14 May, with 26 countries qualifying for the final on 16 May. This year's contest was perceived by many commentators to be tense and politically motivated, especially with Ukraine eventually winning the final [28]. Varying analyses see the contest as being influenced by political conflicts, friendships or cultural bias [2, 6, 9, 12], with a range of news articles explicitly discussing the possibly biased results [17]. Twitter activity was very high throughout the event on the primary

[4] In this context, *different language* refers to tweet's language that is different to the user profile language settings.

[5] `https://www.eurovision.tv/page/stockholm-2016/all-participants`

`#Eurovision` hashtag, with close to 8 million statuses, produced by nearly 1.25 million users.

The study focuses on original statuses (tweets) as the basic entity, as we wish to measure posting behaviour, not reactions. Preliminary analysis shows that they account for 48% of the total activity, of which 4% tweets with an '*unidentified*' language were eliminated. As for profiles, all users have chosen language preferences and no profile was found with the default language settings.

4.2 Multilingualism

The outdegree in the user graph shows the number of languages a user used; observing the outdegree of user nodes in the users graph revealed 20 groups of outdegree, ranging from 1 to 25. Figure 2 shows these groups, size of users and activities. Although 85% of users are monolingual, their activity accounts for 47% of all tweets. Additionally, while the average activity of users is five posts per user, monolingual users were the least active ones, scoring an average of two tweets per user. We found that 18% of tweets were in different languages, with a strong correlation between multilingualism and likelihood of using different languages.

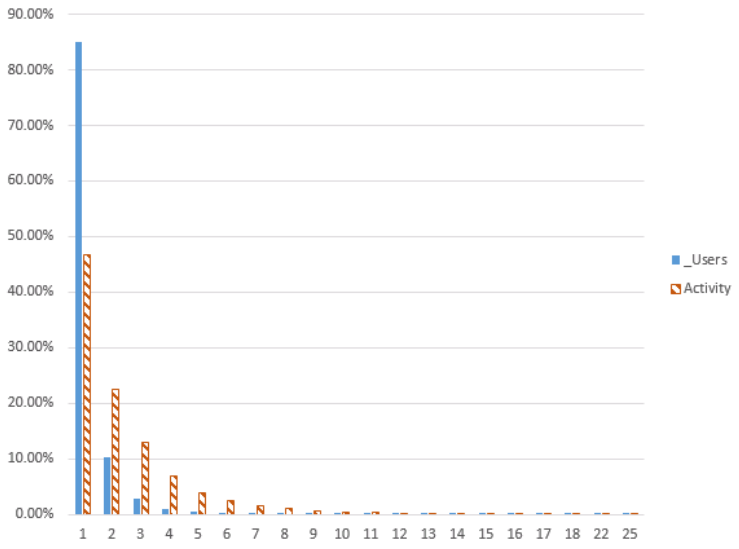

Fig. 2. Multilingual communities on `#Eurovision` and their associated activities.

We used the user graph to generate two communities graphs; the first will be used to explore language communities amongst monolingual users, while the other includes language communities for multilingual users only.

4.3 Monolingual Communities

This graph includes 63 language communities: 15 languages exist as profile-only and have not been used in any post, while 12 were used in posting but never show as a profile language. Moreover, about 13% of monolingual users used different languages in posts which form 10% of tweets in monolingual communities. Hence, strongest relationships exist as a self-loop, as discussed in Section 3.1.

To explore the relationships between language communities, we remove all self-loop edges from the graph. The resultant graph shows that monolingual users with '*en*' as profile language have posted in 47 other languages, causing 43% of tweeting activity, and that 48 other profile communities used '*en*' language in posting 43%. Also, we found that the strongest relationship (edge weight), 9% of activity, is when '*en*' profiles post in '*es*' (Spanish). A further interesting case to mention involve the '*el*' (Greek) and '*ru*' (Russian) languages. Although the number of profile communities that used '*ru*' is more than twice compared to the number of those that used '*el*', they were significantly lower in terms of activity.

4.4 Multilingual Communities

Although multilingual users form 15% of all users in the dataset, they generated 53% of tweeting activity. There are 48 language communities in this graph, 13 languages as profile-only, and 10 as posting languages. With self-loop edges excluded, activity in different languages measured 24% of multilingual users tweets. Also, we found that the strongest relations existed between the '*es*' profile community and the '*en*' posting language, which is the opposite to the monolingual case.

4.5 Visualisation

In Figure 3, we present two communities graphs; the size of the node represents weighted indegree of community; how much a language was used in tweeting, and darkness reflects weighted outdegree; participation from users of language community. Edges link between *profile* and *posting* communities, and their thickness indicates the number of tweets posted.

Whilst Figure 3(a) shows all language communities together, 3(b) presents a filtered graph. This filtered graph depicts relationships amongst language communities that scored high in weighted indegree and outdegree. Also, we eliminated users with activity lower than the overall average (five tweets/user), and generated the communities graph from the remainder.

5 Conclusions

This paper has presented an extensible approach for identifying interactions within language communities using a high-profile real-world case study – the 2016 Eurovision Song Contest – and its associated engagement and interactions on Twitter. This approach utilises network graph properties to explore

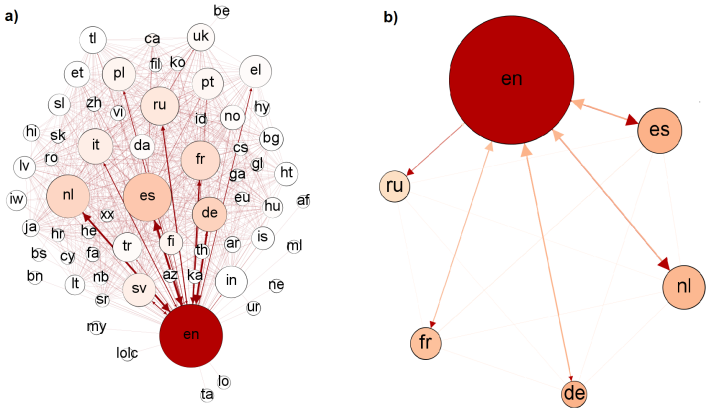

Fig. 3. Language Communities Graphs for `#Eurovision`.

the behaviour of monolingual and multilingual users. Surprisingly, even though monolingual users formed the largest proportion of users, they were less active than multilingual users. The results also confirmed that higher proportions of user multilingualism implies further distance from their profile language. In the profile community, large number of participants does not necessarily imply high language diversity, as a single post in other language is enough to take the community to a higher level of multilingualism. Therefore, filtering out those users with low activity would improve measurement accuracy. In a few cases, we witnessed users participating in a significant number of languages, up to 25 different languages. Such extreme cases may be interesting to investigate for possible spammer/false account detection or for sociolinguistics in more moderate cases (e.g. 2-5 languages).

The graph measures of users may be useful in confirming their association with language community, without the need to crawl their entire Twitter timeline. Although language settings for user profiles may indicate interface preference only, we found persistent activity in the same language across the users, especially for monolingual users.

A possible scenario for governments, politicians or campaigners would be to use this method to measure to what extent other languages are used within a profile community. It may also show how users associate themselves with one community in their profile while using other languages. Monitoring unusual activity for secondary languages may help to uncover important messages or opinions that could not be openly expressed, for a variety of reasons, to the rest of the profile community. This framework may also be extended to measure reactions via retweeting and replying using a variety of natural language processing and sentiment analysis techniques [23], to provide a different perspective for influence analysis.

Bibliography

[1] B. Blamey, T. Crick, and G. Oatley. 'The First Day of Summer': Parsing Temporal Expressions with Distributed Semantics. In *Research and Development in Intelligent Systems XXX*, pages 389–402. Springer, 2013.

[2] M. Blangiardo and G. Baio. Evidence of bias in the Eurovision song contest: modelling the votes using Bayesian hierarchical models. *J. of Appl. Statistics*, 41(10):2312–2322, 2014.

[3] S. P. Borgatti and M. G. Everett. Models of core/periphery structures. *Social Networks*, 21(4):375–395, 2000.

[4] J. Borge-Holthoefer, A. Rivero, and Y. Moreno. Locating privileged spreaders on an online social network. *Phys. Rev. E*, 85(066123), 2012.

[5] A. Bruns, T. Highfield, and J. Burgess. The Arab Spring and Social Media Audiences: English and Arabic Twitter Users and Their Networks. *American Behavioral Scientist*, 57(7):871–898, 2013.

[6] O. Budzinski and J. Pannicke. Culturally biased voting in the Eurovision Song Contest: Do national contests differ? *J. of Cultural Economics*, pages 1–36, 2016.

[7] J. Caverlee, Z. Cheng, D. Z. Sui, and K. Yeswanth Kamath. Towards Geo-Social Intelligence: Mining, Analyzing, and Leveraging Geospatial Footprints in Social Media. *IEEE Data Eng. Bull.*, 36(3):33–41, 2013.

[8] M. Cha, F. Benevenuto, H. Haddadi, and K. Gummadi. The World of Connections and Information Flow in Twitter. *IEEE Trans. on Systems, Man, and Cybernetics*, 42(4):991–998, 2012.

[9] N. Charron. Impartiality, friendship-networks and voting behavior: Evidence from voting patterns in the Eurovision Song Contest. *Social Networks*, 35(3):484–497, 2013.

[10] Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In *Proc. 19th ACM Conf. on Information and Knowledge Management*, pages 759–768, 2010.

[11] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. Gonçalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach. In *Proc. Workshop on Languages in Social Media*, pages 58–65, 2011.

[12] V. Ginsburgh and A. G. Noury. The Eurovision Song Contest. Is voting political or cultural? *Euro. J. of Political Economy*, 24(1):41–52, 2008.

[13] V. Gisnburgh and S. Weber. *How Many Languages Do We Need? The Economics of Linguistic Diversity*. Princeton University Press, 2011.

[14] E. Gruffydd Jones and E. Uribe-Jongbloed, editors. *Social Media and Minority Languages: Convergence and the Creative Industries*. Multilingual Matters Ltd, 2013.

[15] L. Hong, G. Convertino, and E. H. Chi. Language Matters In Twitter: A Large Scale Study. In *Proc. 5th Int. AAAI Conf. on Web and Social Media*, 2011.

[16] R. Kang, S. Brown, and S. Kiesler. Why do people seek anonymity on the internet?: informing policy and design. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pages 2657–2666, 2013.

[17] A. Kirk, J. Kempster, and S. Franco. Eurovision 2016: How does country bias affect the result? `http://www.telegraph.co.uk/music/news/eurovision-2016-how-country-bias-affects-the-result`, May 2016. (accessed 2017-04-01).

[18] J. Kulshrestha, F. Kooti, A. Nikravesh, and K. P. Gummadi. Geographic Dissection of the Twitter Network. In *Proc. 6th Int. AAAI Conf. on Web and Social Media*, 2012.

[19] S. Kumar, F. Morstatter, and H. Liu. *Twitter Data Analytics*. Springer, 2014.

[20] Y. Liang, J. Caverlee, Z. Cheng, and K. Y. Kamath. How big is the crowd?: event and location based population modeling in social media. In *Proc. 24th ACM Conf. on Hypertext and Social Media*, pages 99–108, 2013.

[21] W. Liu, M. Pellegrini, and X. Wang. Detecting Communities Based on Network Topology. *Scientific Reports*, 4(5739), 2014.

[22] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the Sample Good Enough? Comparing Data from Twitters Streaming API with Twitters Firehose. In *Proc. 7th Int. AAAI Conf. on Web and Social Media*, pages 400–408, 2013.

[23] M. Mostafa, T. Crick, A. C. Calderon, and G. Oatley. Incorporating Emotion and Personality-Based Analysis in User-Centered Modelling. In *Research and Development in Intelligent Systems XXXIII*. Springer, 2016.

[24] G. Oatley and T. Crick. Measuring UK Crime Gangs: A Social Network Problem. *Social Network Analysis & Mining*, 5(1):33, 2015.

[25] M. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha. Core-Periphery Structure in Networks. *SIAM J. on Appl. Math.*, 74(1):167–190, 2014.

[26] Y. Takhteyev, A. Gruzd, and B. Wellman. Geography of Twitter networks. *Social Networks*, 34(1):73–81, 2012.

[27] L. Tan, S. Ponnam, P. Gillham, B. Edwards, and E. Johnson. Analyzing the impact of social media on social movements: A computational study on Twitter and the Occupy Wall Street movement. In *Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining*, 2013.

[28] The Telegraph. Eurovision 2016: Furious Russia demands boycott of Ukraine over Jamala's 'anti-Kremlin' song. `http://www.telegraph.co.uk/news/2016/05/15/eurovision-2016-furious-russia-demands-boycott-of-ukraine-over-j`, May 2016. (accessed 2017-04-01).

[29] A. Willis, A. Fisher, and I. Lvov. Mapping networks of influence: tracking Twitter conversations through time and space. *Participations: J. of Audience & Reception Stud.*, 12(1):494–530, 2015.

[30] M. Zappavigna and J. R. Martin. *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. Continnuum, 2012.

[31] J.-X. Zhang, D.-B. Chen, Q. Dong, and Z.-D. Zhao. Identifying a set of influential spreaders in complex networks. *Scientific Reports*, 6(27823), 2016.