Community-specific language in online citizen science forums: a corpus-driven diachronic study

Claudia Viggiano (University of Portsmouth, UK)

Citizen Science can be defined as the collaboration between professional scientists and interested members of the public who together carry out scientific research (Socientize Project, 2013); it is considered to be a form of 'crowdsourcing' research. Online forms of citizen science take place on citizen science platforms, where volunteers usually tag, classify or annotate data. Currently the largest such citizen science platform, Zooniverse (www.zooniverse.org) is an umbrella website which hosts over 50 scientific projects, ranging from astronomy, e.g. Galaxy Zoo (www.galaxyzoo.org) to zoology, e.g. Chimp&See (www.chimpandsee.org). Each of the Zooniverse projects include a 'Talk' section where volunteers discuss science, ask scientists or moderators for help on the tasks they are carrying out, comment on their findings, or get to know one another. The present study as reported on in this paper is part of a wider project that uses corpus linguistics methods and tools to analyse the language of citizen science communities and, in particular, to investigate the relationship between online community building and short-term language change.

Many early and pioneering studies in sociolinguistics used a diachronic approach considering variation in language use according to social characteristics such as social class (Labov, 1966) or gender (Cheshire, 1982), doing so in the context of face-to-face communities, demonstrating, as in Trudgill (1974) how gender affects dialect in different social classes. Similarly, a recent trend in corpus linguistics (e.g. Aarts et al. 2013) is to study short-term diachronic change to investigate current change—changes in the language that have taken place over relatively short spans of time, using corpora. The present paper builds on these two approaches to diachronic change and group affiliation to analyse online citizen science communities.

Lave and Wenger (1991) defined a 'community of practice' as a group of people getting together to undertake certain tasks; Zooniverse communities, on the other hand, are an example of a virtual community of practice (Stewart, 2010). Due to their task-based nature and collaborative knowledge creation, they can be studied using the Community of Inquiry framework (Garrison, 2006), which identifies communities whose collective goal is based on empirical enquiry, and is often used to assess the success, and measure the learning outcomes, of pedagogic or e-learning platforms; for example Goertzen & Kristjansson (2007) found that the learning process of students on a distance learning programme was collaborative and deeply dependent on interpersonal engagement among participants. Accordingly, 'social presence,' the ability to participate personally and authentically in the community and to be perceived as salient and 'real' by others (Nichols 2009), is a part of the aforementioned community of inquiry framework. As such, there is an implication that the stronger the social presence and therefore the sense of community, the more productive said community is in achieving its goals. Lander (2015) identifies three different types of descriptors of social presence in the language used in online learning communities: (i) affective responses such as the

expression of feelings and emotions and the use of humour (I'm sorry I've been like this all day, I was up ALL night working on college work), (ii) cohesive responses such as greetings and inclusive pronouns (hello my beautiful galaxy friends), and (iii) interactive responses such as asking questions and quoting other members' posts (Allow me to pose a question in response to your question! Yes, answering a question with a question LOL!).[1]

With such a framework in mind, the present work employs a 6 million-word corpus collected across 43 Zooniverse projects (Williams & Viggiano, 2016) to analyse short-term diachronic change and how language shifts therein are influenced by in-group dynamics; in particular, this paper looks at the introduction of new, community-specific terms: these are typically new lexical items or expressions that are introduced by one or more members, often from a semantic field related to the tasks undertaken by the community, and are used by members to signal their social presence and participation within the group. Hence, this paper looks at how the introduction and collaborative adoption of the terms by the community can be seen as cause and effect of social presence: on the one hand, the use of 'in-group' language is a marker of the participant's strong social presence; on the other, a participant's strong social presence can foster the creation of new 'in-group' language. The corpus downloaded from the Zooniverse website includes timestamp and poster information, which allow for the tracking of changes across time within the 'Talk' forums and to identify which forum members 'lead' change, and how the use of such terms can be, in itself, a marker of social presence.

Following a corpus-driven approach (Tognini-Bonelli, 2001), corpus query software Sketch Engine (Kilgarriff et al., 2014) is used to extract community-specific lexical items; in order to do so, a triangulation method of keyword and term analysis is employed: so as to determine lexical items unique to it, the Zooniverse corpus is set against three different sets of reference corpora—a general corpus, the New Model Corpus (https://www.sketchengine.co.uk/wp-content/uploads/New-Model-Corpus.pdf), an online corpus, enTenTen13, and an online scientific corpus, ScienceBlogs (both available on www.sketchengine.co.uk). Among the key items are terms with which members of Zooniverse self-identify as part of the community, thus creating group cohesion. For example, users self-identify as "zooites," a term that is used to greet and address fellow volunteers:

(1) Ahhh.. good afternoon dear fellow zooites ! :-* I've had a crazy few days and only just had a chance to catch up a bit.. (after snoozing on the sofa for a while!!) :)

Occasionally, the word is used to identify members as central or 'ordinary' "zooites," therefore addressing an issue with expertise and status within the community:

(2) an ordinary zooite can notice something unusual in Radio Galaxy Zoo, and comment on it. […] So you, other ordinary zooites, too can likely make stunning finds! :)

[1] 1 All three examples are drawn from my Zooniverse corpus (see below here).

Owing to the earlier mentioned time-stamping of posts as a form of metadata in the collected Zooniverse corpus, it is also possible to track the first few occurrences of the word, and how its use and meaning were collaboratively established with time and through discourse, displaying a strong interpersonal environment whereby meaning can be explicitly renegotiated:

(3) why are we "Users"? Surely there's a better term? Perhaps " zooites "? or "members" (we have to register and sign in)?
(4) You're all zooites in my eyes, but that term might confuse newbies! And "members" sounds a bit like something you pay for

Tracking the emergence, adoption and retention of community-specific language over time is a step towards the understanding of the collaborative creation of meaning in virtual communities, an aspect of the interpersonal dimension of meaning with which discourse studies informed by Systemic Functional Linguistics are concerned. In this regard, it builds on what Bednarek (2010) calls the instantiation of meaning—that is, how meaning serves as a means of both production and reproduction. As will be demonstrated in the paper, different such novel terms are picked up, maintained and re-negotiated in different ways according to matters of social presence (e.g. the initiating participant's familiarity to the community, measured by the frequency and regularity of their posts). The mutual relationship between social presence and the creation and use of community-specific language can therefore be seen as a strong component of the community of inquiry, adding a 'creative' dimension to the social presence framework which Lander (2015) and Goertzen & Kristjansson (2007) have found to be a central factor in the success of online communities.

References

Aarts, B., Close, J., & Wallis, S. (Eds.). (2013). *The verb phrase in English: Investigating recent language change with corpora*. Cambridge University Press.

Bednarek, M. (2010). Corpus linguistics and systemic functional linguistics: Interpersonal meaning, identity and bonding in popular culture. In Bednarek, M. & Martin J.R. (eds.) *New Discourse on Language: Functional Perspectives on Multimodality, Identity, and Affiliation* (pp. 237-266). London: Continuum.

Cheshire, J. (1982). Variation in an English dialect: A sociolinguistic study. *Cambridge Studies in Linguistics London*, 37.

Garrison, D. R. (2006). Online collaboration principles. *Journal of Asynchronous Learning Networks*, 10(1), 25–34.

Goertzen, P., & Kristjánsson, C. (2007). Interpersonal dimensions of community in graduate online learning: Exploring social presence through the lens of Systemic Functional Linguistics. *The Internet and Higher Education*, 10(3), 212-230.

Kilgarriff, A., et al. (2014). The Sketch Engine: ten years on. *Lexicography*, 1–30. Available at https://www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2014.pdf

Labov, W. (2006). *The social stratification of English in New York city*. Cambridge University Press.

Lander, J. (2015). Building community in online discussion: A case study of moderator strategies. *Linguistics and Education*, 29, 107–120.

Lave, J, Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Nichols, M. (2009). *Online discourse. E-Primer series*. Available at http://tinyurl.com/gpej524.

Socientize Project (2013). *Green Paper on Citizen Science: Citizen Science for Europe - Towards a better society of empowered citizens and enhanced research*. Socientize consortium. Available at: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=4121.

Stewart, T. (2010). Online communities. Editorial. *Behaviour and Information Technology*, 29(6), 555–556.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Trudgill, P. (1974). *The social differentiation of English in Norwich (Vol. 13)*. CUP Archive.

Williams, J. & Viggiano, C. (2016). *Capturing the Zoo: A system for downloading, preparing, and managing corpus data from online forums*. Talk at Corpus Linguistics in the South, University of Sussex, UK, 27/2/2016.