KeyWords of success – what words are associated with success in online Citizen Science?

Glenn Hadikin (University of Portsmouth, UK)

This paper aims to answer one primary research question: *What words are associated with success in online Citizen Science?* Citizen Science (CS) is a form of science crowdsourcing where members of the public come together but rather than give money they give their time to help various universities and other research institutions with their research. The concept arguably began in 1900 when members of the US Audubon society asked its members to count birds on Christmas day rather than participate in the tradition of shooting them (Silvertown 2009). Since then citizen scientists have been involved in helping researchers study the evolution of snails, recording air, soil and water quality and they have been conducting a survey of invasive species in the USA (Silvertown, 2009). In the last five years, there has been a burst of research activity into the phenomenon of CS, and, since 2009, the website Zooniverse.org has developed into the world's largest CS site where volunteers join projects online and help label or classify data. It is timely for the research community to ask what influences these volunteers - especially the effect of their linguistic choices.

In this work-in-progress, techniques from corpus linguistics have been combined with methods from economics to compare the language used in seven weaker CS projects with six more successful ones from the Zooniverse. Measures of success are taken from Cox et al. (2015), and include various measures of public engagement and scientific success such as peer-reviewed publications, user posts and blog posts.

The data

Zooniverse project	*Public engagement score*
Snapshot serengeti	0.614
Galaxy zoo old	0.453
Galaxy zoo new	0.416
Planet hunters	0.382
Planet four	0.25
Seafloor explorer	0.118
Ancient lives	0.071
Old weather	0.049
Milky way	0.043
Bat detective	0.041
Cyclone center	0.03
Solar stormwatch	0.027
Moon zoo	0.015

Figure 1: Public engagement score for 13 Zooniverse projects based on Cox et al. (2015)

Cox et al. (2015) use a positioning matrix to rank 18 Zooniverse projects based on 6 'contribution to science' elements and 6 'public engagement' elements. This paper makes use of their public engagement ranking criteria only. As an example, the collaboration element of public engagement is calculated as the following :

$$\frac{\text{Number of papers with citizen scientist authors}}{(\text{Project age})^2}$$

Wright et al. (2016) built corpora of the text-based discussion forums associated with 43 Zooniverse projects. The 13 shown in figure one were both available as corpora and have been given a public engagement score in Cox et al. (2015) so were selected for this study.

Corpus	Number of tokens
topsix	5 851 392
bottomseven	1 220 710

Figure 2: Number of tokens in each of the two Zooniverse corpora

Figure two shows the size of the two corpora that were used for a KeyWord analysis using WordSmith tools 6.0 (Scott, 2012). The corpus 'topsix' is simply all the forum data from the six projects with the highest engagement scores and 'bottomseven' contains all the forum data from the seven with the lowest scores. Note there is different subject matter being discussed in each project forum so that must be considered when interpreting results. The risks associated with a small reference corpus highlighted in Berber-Sardinha (2004) are also acknowledged.

KeyWords of success

KeyWord category	KeyWords
space	UNIVERSE GALAXIES GALAXY STAR STELLAR STARDUST DARKNESS SPIRAL MERGER PLANET TRANSIT MERGE REDSHIFT PLANETS ASTEROID
religion	ANGELS HEAVENS GOD
grammar	JUST WE INTO ME RE THEN IS NOT WITHIN MY YOU ITS EITHER ALL WHERE AND A CANNOT SUCH I INFACT WHY SHE YOUR IM YA
names	PADDY IZZY JOHN EDD WEEZ WAVENEY TSERING KITHARODE MEG PLUK FIRESTORM ALICE ROY LIZ TERRANCE HANNY CURTIS JOHNSON DAVE COOK ANDREW SOPHIE HIGGS SUE PLANCK PETER PAULROGERS EINSTEIN ADAMS PAT ALBERT
everyday	STAY PLAY SWEET HAIR ENGAGING PRACTICING BEER CAMERA PASSION

	UNDERSTAND MUSIC DESPERATELY CLUMPY CAR LEGS ACRONYMS SAY COLOURS HOUSE EVERYBODY LONG THANKING
internet	D URL LOL ROFL O
Zooniverse	ZOO GZ ZOOITES SERENGETI ZOOITE OBJID OOTD
time	NIGHT MOMENT FUTURE MORNING TIME EVENING TOMORROW YEARS SEASON DURATION TONIGHT FRIDAY PERIOD SATURDAY
animals	WILDEBEEST ZEBRA CAT HOUSE-SIZED CATS HYENA HORNS HARTEBEEST GAZELLE ANIMALS BUFFALO BIRD BUCK WARTHOG LION ANIMAL ZEBRAS
other science	NUCLEUS ATOM UNITS CANDIDATES SPECTRAL INHABITS ELLIPTICALS TELESCOPE QUANTUM VOLCANOES EIGENSTATE ATOMS PHYSICS CA GAS EQUILIBRIUM CLASSIFIERS PARTICLES PHOTONS
interaction	CIAO HIYA OH GOODNIGHT BYE YEAH HEY NIGHTY HEH HA
numbers	BILLION SEVENTY MILLION HUNDRED HALF BILLIONS THOUSAND
quotes	TARE TUTTARE SOHA OM TURE CLARKE ARTHUR PAEANS AMARA MILLIWAYS STRIP-MINE SERAFINOWICZ KLEE BOWING KANO JIGARO EXALTED FORSTER HOWARDS
other	AM THEORY WELL COOKIES SPIDERS SPIDER ROSE SPIN MODEL INTERACTING EXPANSION TALK IRREGULARS RING C Z MAG SM EST FERMATS IV ES B Q AGIAN K MO ARM

Table 3: Selected KeyWords when 'topsix' is compared with 'bottomseven'

Table 3 shows a selection of the 500 KeyWords that were generated when 'topsix' was compared against 'bottomseven'. The KeyWords were manually categorised into the 14 categories shown. Space is one of the largest categories with 56 KeyWords.[1] This is clearly influenced by the subject matter of one of the largest and most successful Zooniverse projects - Galaxy Zoo. Though unsurprising as KeyWords, one should not hastily overlook such a category as uninteresting, however, because it highlights what subjects the volunteers are choosing to focus on, and could potentially be used to predict the success of a new project. Similarly the animals

[1] Only a subset of the larger categories were reproduced in table three.

listed are influenced by the subject matter of a popular project - Snapshot Serengeti - but the names of animals and related words coming through as KeyWords could indicate animals that are relatively easy to identify and may provide Zooniverse leaders with useful information about how to adjust their training materials to support discussion and identification of less well-known animals.

Other sections such as *grammar*, *internet* and *interaction* are less obviously influenced by the subject matter and provide a snapshot of the linguistic behaviour of volunteers interacting and working together. The set of personal pronouns *we, me, you, I, she* (as well as *Im* and *ya* as a variant of *you*) suggest a friendly atmosphere where volunteers are engaging in personal chat as well as discussing science. Lines 1, 2 and 3 below show the use of *ya* in the three samples of text from top six.

(1) Well CIS Miami is on, so see ya all later. ;)
(2) a monocle, much cooler :D c ya Dave ;D ;D ;D
(3) you a Horlicks :) That helps ya sleep. Well it is cloudy, but

Caution is needed when it comes to interpretation of any results based on this method. I am not claiming that a project leader can simply use such words more frequently to improve interaction and better statistics in terms of volunteers classifying their data. It is possible that other factors such as regular interaction between professional scientists and volunteers keeps volunteers interested, and that any linguistic evidence of a friendly atmosphere comes as a secondary effect. The linguistic atmosphere briefly sketched out here highlights areas of variation and themes that I will discuss further in the paper. Our discussion around this topic and further work could prove valuable to the leaders of other online communities who want to understand the relationship between project goals and online interaction between site users.

References

Berber-Sardinha, T. (2000). Comparing corpora with WordSmith Tools: How large must the reference corpus be?, in *Proceedings of the Workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000).* 1-8 October 2000, Hong Kong, pp. 7–13. Also available online from https://www.aclweb.org/anthology/W/W00/W00-0902.pdf

Cox, J., Oh, E.Y., Simmons, B., Lintott, C., Masters, K., Greenhill, A., Graham, G. & Holmes, K. (2015). Defining and measuring success in online citizen science: a case study of Zooniverse projects. *Computing in Science & Engineering*, 17 (4) pp. 28-41. ISSN 1521-9615 DOI 10.1109/MCSE.2015.65

Scott, M. (2012). *WordSmith Tools* (version 6). Stroud: Lexical Analysis Software.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, 24(9) pp. 467–471. DOI 10.1016/j.tree.2009.03.017

Wright, S., Clarke, B., Hadikin, G., Saraceni, M., Viggiano, C., & Williams, J. (2016, December 19). *Language of Citizen Science*. Retrieved from

http://www.port.ac.uk/centre-for-european-and-international-studies-research/research-projects/language-of-citizen-science/