Yisi Han (2020): Understanding Developers' Linguistic Behaviors in Hierarchical Open Source Communities. In: Proceedings of the 18th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591*), DOI: 10.18420/ecscw2020_dc10*

Understanding Developers' Linguistic Behaviors in Hierarchical Open Source Communities

Yisi Han
Software Institute, Nanjing University, Nanjing 214000, P.R.China
hanysannie@gmail.com

Abstract. Online communication among developers is critical for the success of Open Source Software (OSS) communities. As a typical complex sociotechnical system, an OSS community often forms a hierarchical structure in which a few developers are elite while the rest are non-elite. The differences in social status among developers would unavoidably result in different linguistic behaviors. I sought to develop an understanding towards such differences in linguistic behaviors and how they influence project outcomes. Using data from GITHUB, I designed three interrelated studies. The first will characterize such differences at a collective level. The second will further explore the dynamic changes on linguistic behaviors with the focus on individual developers who experience non-elite and elite transitions. The last aims to quantify the impacts of linguistic behaviors on project outcomes. I also plan to build a tool to help developers achieve better communication within and across hierarchies.

Research question

The success of an Open Source Software (OSS) project depends on the collaboration and communication of its developers. Research into online

collaboration in OSS communities have repeatedly shown that communication plays an important role in collaboration (Steinmacher et al., 2019a,b; Singh et al., 2012). Effective communication depends on the proper language use among developers. The goal of this study is to understand communication among developers in OSS communities from the perspective of people's linguistic behaviors in the lens of social status. I focuses on the large OSS projects hosted on GITHUB which is one of the most widely adopted platforms by OSS projects. Projects hosted on GITHUB often form a hierarchical structure of developers where some users are elite and the rest are non-elite(Wang et al., 2020). The differences in social status would result in different linguistic language behaviors. This study aims to identify the differences in linguistic behaviors and characterize how they influence project outcomes. Data of 20 projects in GITHUB were collected including information of developers and the texts of all communication.

To find the differences in linguistic behaviors, I designed two studies to analyze the linguistic differences of developers in two aspects. The first (Study I) focuses on the linguistic behaviors of developers in different status. This study will provide a collective view about the linguistic differences of all developers. To further understand differences in linguistic behaviors of individual developers, the second (Study II) will explore the dynamic changes in linguistic behaviors of developers who experience non-elite and elite transitions. Based on the above two studies, I plan to conduct the third one (Study III) investigating how linguistic behaviors, with the presence of the social status system consisting of the elite and the non-elite, influence project outcomes. In this study, both OSS developers' social status and their linguistic behaviors will be taken into consideration. While prior literature, e.g., Blincoe et al. (2016); Scacchi (2004), confirms the importance of the hierarchical social status system, this study extends our knowledge of the joint impacts of social status and linguistic behaviors. Specifically, the three studies will provide answers to the following three high-level research questions:

RQ1 *What are the differences regarding the linguistic behaviors between elite and non-elite OSS developers?*

RQ2 *How one's linguistic behaviors changes if his/her social status experiences transitions (elite↔ non-elite)?*

RQ3 *How OSS developers' social status and their linguistic behaviors jointly influence project outcomes?*

In addition to the above empirical studies, I would also plan to apply the results of these studies to build a tool that helps developers achieve better communication within and across hierarchies.

Methodological approach

Developers' linguistic behaviors will be investigated with computational approaches (Pennebaker et al., 2014; Zhang et al., 2019) has been put to

investigate linguistic behaviors of different users. Based on the previous work, we may hypothesize that elite and non-elite text exhibit different linguistic behaviors in their projects. Some text analysis tools will be applied to the text of communication. Given that conversations in GITHUB are short text and their structures are not complex, Linguistic Inquiry and Word Count (LIWC) Pennebaker et al. (2001) will be the main analytical tool. LIWC has two components: the processing component and the dictionaries. The processing component is a computational analysis tool. A dictionary refers to the collection of words that define a particular category. LIWC calculates the percentages of words in text belonging to its dictionaries. LIWC detects meaning including attentional focus, emotionality, social relationships, thinking styles and individual differences Pennebaker (2011); Tausczik and Pennebaker (2010).

I plan to apply LIWC to identify the differences between elite and non-elite developers' linguistic behaviors in Study I. Similarly approaches will be applied in Study II. We would also employ several other computational linguistics techniques, for example, word embeddings. Since Study III focuses on identify relationships among several constructs (social status, linguistic behaviors, project outcomes), I will apply regression techniques.

Work to date

I had done a substantial amount of data preparations. Historical data of 20 large projects hosted in GITHUB between 2010 and 2019 has been collected. For each project, four categories of data are collected to constitute a corpus, including users' information, data of pull and request, data of commits, and data of issues. To investigate the linguistic behaviors of developers, we divided all conversations in the corpus into elite texts and non-elite texts according to the identification of users. There are 142,993 non-elite texts and 101,938 elite texts. Matching techniques were adopted to filter data to guarantee elite texts and non-elite texts are a prior balanced on any observable features so that some features would not bias the results. I have chosen four features of the text as feature set for matching: (1) the number of words in a text, (2) the number of pull and request, commit and issue, (3) the year of a text, (4) the project of a text. I used the nearest neighbor matching technique to estimate similarities of texts in the corpus. I compared p-values of four features of texts before and after the matching procedure which shows that filtered texts achieved a prior balance on the four features.

Study I is ongoing. In the next step, I applied LIWC to find differences in the linguistic behaviors of elite developers and non-elite developers. Then I used *TF-IDF* to calculate the score of each word to find the most important keywords in elite and non-elite texts. Finally, I investigated the observed patterns from the discussion of important keywords in elite and non-elite texts. The results revealed that some words like "branch", "release", elite developers use them much more times than non-elite developers; and elite developers have a clear and concise description of issues. Elite and non-elite developers have different rhetorical

patterns describing some keywords. For example, non-elite developers like using "minor" and "small" to describe "fix" while elite developers like using "address". Furthermore, some elite developers have fixed patterns about how to describe "fix" almost never change.

By the date of ECSCW 2020, I expect that Study has been concluded and a manuscript should be developed. I target CSCW'20 as the potential for this manuscript. The second and third study will be starting shortly.

Next steps

From now, I plan to take the following steps towards finishing my dissertation work:

- Refine the current work on the differences in linguistic behaviors between elite and non-elite developers.
- At the individual developer level, filter and adjust information of developers. Then analyze dynamic changes on linguistic behaviors of developers who experience transitions between non-elite and elite .
- Develop regression models to identify the relationships among social status, linguistic behaviors, and project outcomes.
- Use results of the above studies to build a tool that aims to help developers achieve better communication within and across hierarchies.

Expected contributions

This study makes the following three major contributions:

- The work extends the extant CSCW and Software Engineering literature by develop a deep understanding of communication among developers in OSS projects from the perspective of people's linguistic behaviors in the lens of social status.
- The work results a computational tool that help developers to articulate their language, thus achieve better communication within and across social hierarchies in their OSS project.
- The work provides a large labelled dataset of computer-mediated communication happened in online collaboration.

References

Blincoe, K., J. Sheoran, S. Goggins, E. Petakovic, and D. Damian (2016): 'Understanding the popular users: Following, affiliation influence and leadership on GitHub'. *Information and Software Technology*, vol. 70, pp. 30–39.

Destefanis, G., M. Ortu, S. Counsell, S. Swift, M. Marchesi, and R. Tonelli (2016): 'Software development: do good manners matter?'. *PeerJ Computer Science*, vol. 2, pp. e73.

Parra, E., S. Haiduc, and R. James (2016): 'Making a difference: an overview of humanitarian free open source systems'. In: L. K. Dillon, W. Visser, and L. Williams (eds.): *Proceedings of the 38th International Conference on Software Engineering, ICSE 2016, Austin, TX, USA, May 14-22, 2016 - Companion Volume*. pp. 731–733, ACM.

Pennebaker, J. W. (2011): 'The secret life of pronouns'. *New Scientist*, vol. 211, no. 2828, pp. 42–45.

Pennebaker, J. W., C. K. Chung, J. Frazee, G. M. Lavergne, and D. I. Beaver (2014): 'When small words foretell academic success: The case of college admissions essays'. *PloS one*, vol. 9, no. 12, pp. e115844.

Pennebaker, J. W., M. E. Francis, and R. J. Booth (2001): 'Linguistic inquiry and word count: LIWC 2001'. *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, pp. 2001.

Pennebaker, J. W. and L. A. King (1999): 'Linguistic styles: Language use as an individual difference.'. *Journal of personality and social psychology*, vol. 77, no. 6, pp. 1296.

Riehle, D. (2015): 'How Open Source Is Changing the Software Developer's Career'. *IEEE Computer*, vol. 48, no. 5, pp. 51–57.

Scacchi, W. (2004): 'Free and open source development practices in the game community'. *IEEE software*, vol. 21, no. 1, pp. 59–66.

Singh, V., S. Kathuria, and A. Johri (2012): 'Newcomer integration and learning in OSS technical support communities'. In: S. E. Poltrock, C. Simone, J. Grudin, G. Mark, and J. Riedl (eds.): *CSCW '12 Computer Supported Cooperative Work, Seattle, WA, USA, February 11-15, 2012 - Companion Volume*. pp. 215–218, ACM.

Steinmacher, I., M. A. Gerosa, T. U. Conte, and D. F. Redmiles (2019a): 'Overcoming Social Barriers When Contributing to Open Source Software Projects'. *Computer Supported Cooperative Work*, vol. 28, no. 1-2, pp. 247–290.

Steinmacher, I., C. Treude, and M. A. Gerosa (2019b): 'Let Me In: Guidelines for the Successful Onboarding of Newcomers to Open Source Projects'. *IEEE Software*, vol. 36, no. 4, pp. 41–49.

Tausczik, Y. R. and J. W. Pennebaker (2010): 'The psychological meaning of words: LIWC and computerized text analysis methods'. *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54.

Wang, Z., Y. Feng, Y. Wang, J. Jones, and D. Redmiles (2020): 'Unveiling Elite Developers' Activities in Open Source Projects'. *ACM Transactions on Software Engineering and Methodology*, vol. 29, no. 3, pp. 16:1–36.

Xu, H., Z. Zhang, C. Lin, and Z. Ding (2008): 'The Study on Innovation Mechanism of Open Source Software Community'. In: *International Conference on Wireless Communications*.

Zhang, J. S., C. Tan, and Q. Lv (2019): 'Intergroup Contact in the Wild: Characterizing Language Differences between Intergroup and Single-group Members in NBA-related Discussion Forums'. *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–35.