Is Stack Overflow in Portuguese more attractive for Brazilian users?

Angela Lozano*, Bogdan Vasilescu†, Miguel Botto Tobar‡§, Weslley Torres§, and Alexander Serebrenik§,
*Vrije Universiteit Brussel, Brussels, Belgium
†Carniege Mellon University, Pittsburgh, PA, USA
‡University of Guayaquil, Guayaquil, Ecuador
§Eindhoven University of Technology, Eindhoven, The Netherlands

***Abstract*—The abstract goes here.**

I. Introduction

Software development and maintenance are activities that often involves many concepts and reference documents. Many software aspects may be changed over time. In order to work with them and details involved in a software project, developers often need helps from one another [1]. Nowadays, developers are using online forums as a widely way to ask questions and/or answer them about different issues related software development. StackOverflow (SO) [1] is a Q&A site with more than six million registered users. SO also has the version on Portuguese, Russian and Spanish.

The main purpose of the study was to understand the motivations behind stack overflow usage, to what extent it has/can contribute to improve skills of its users. In particular, we were interested in the motivations and profiles of users whose mother tongue is not English, and in case they are bilingual their participation (or lack of participation) in the initial website (i.e., in English http://stackoverflow.com/ and in the website dedicated to their mother tongue (i.e., in Portuguese http://pt.stackoverflow.com/. The main research questions that guided our study are:

RQ1: Which is the purpose of using SOPT? Miguel ►*(asker, mostly asker, both equality, mostly answerer, answerer, no activity)*◄

RQ2: What kinds of questions are asked on SOPT and which ones are answered?

Miguel ►*related to programming languages, environment, framework, so on*◄

RQ3: What are most common problems faced related to usage?

Miguel ►*many questions, less answers?*◄

To this study, we focus on Brazilian users because they are the biggest Portuguese-speaking community on SOPT, and in addition, the Brazilian software developers have proven highly skilled in providing software solutions Miguel ►*??*◄ .

[1]http://stackoverflow.com/

II. Theory

Alexander ►*This is an attempt to position our work in the broader context*◄ English is the language of software [2]. Still, the need of translation and, broader localization, of software elements targeting end users, such as user interfaces, user manuals and support platforms, has been commonly recognized. This is, however, less obvious for documents targeting software developers. For instance, the "Java for Consumers" page[2] exists in such languages as Dutch and French, while no such counterparts exist for the "Java for Developers" page[3]. Still, Java 8 documentation has been translated, *e.g.*, to Japanese[4], documentation of PostgreSQL 9.5 into Russian[5] and novatec is a Brazilian company specialized in translating O'Reilly books into Portuguese. Going beyond translation, original software engineering books have been published, *e.g.*, in Alexander ►*We need examples; in Russian I can find only translations and textbooks*◄ . Moreover, online developers' communities exists, *e.g.*, in Spanish[6] and French[7], while StackOverflow (SO)[8] in addition to English supports equivalent Q&A platforms in, *e.g.*, Portuguese, Russian and Japanese.

The question hence arises of the function of those non-English original or translated information sources in the developers' communities. Do they empower developers by providing them with access to technological documentation or impair their abilities not only by not encouraging them to learn English but also by encouraging them to rely on resources in their own language that—due to the popularity of English at expense of other languages—might be scarce, erroneous or out-dated? Are those information sources still relevant anno 2016 despite the progress made in automatic translation?

This discussion is clearly related to the question of the role of English as a neutral *lingua franca* or as a mechanism of domination [3], [4]. Alexander ►*Both Tardy [3] and most chapters in Ammon's book [4] do not discuss technology but*

[2]https://www.java.com/download/
[3]http://www.oracle.com/technetwork/indexes/downloads/index.html?ssSourceSiteId=ocomen
[4]http://docs.oracle.com/javase/jp/8/docs/api/
[5]https://postgrespro.ru/docs/postgresql/9.5/index.html
[6]http://www.lawebdelprogramador.com/
[7]http://www.developpez.net/forums/
[8]https://stackoverflow.com/

science; still this seems to be closely related.◄ **Alexander** *►One of the claims in science is that papers written in languages other than English or published in non-English-speaking venues are "invisible", they are not cited etc. Can something similar be claimed for software engineering? I know that Lua has been created in Brazil and Python in the Netherlands. Did this somehow affect their adoption?◄* Indeed, if English is seen as a necessary and neutral lingua franca, then technological solutions such as automatic translation should be encouraged as they have the potential of alleviating scarcity of the non-English resources or their tardiness. If, however, English is seen as a domination mechanism **Alexander** *►here I wanted to say something like "the developers need tools to oppose this dominance" but then I've started doubting whether this is true.◄* **Alexander** *►Carmel [2] explains why English is the dominant language in software. "In addition to these nine better known U.S. competitive advantages, two culturally linked assertions are presented that have received scant attention vis-á-vis competitive analysis. First, the industrial evolution of software development is at an immature stage still a cottage industry practiced by craftsmen in a cultural milieu of artisans and thus does not track other global high-technology trends. Second, packaged software is part of the copyright industry (e.g., film and music) in which United States-based firms have a sustained advantage. While manufacturing capabilities are significant for technology industries, culturally related factors, such as creativity, are more important for copyright industries. The U.S. "culture of software" which helps explain U.S. hegemony, is introduced and discussed. The three elements of this culture are the culture of individuals as manifested by the individualistic computer hacker; the entrepreneurial culture and its risk-taking ethos; and the software development culture with its embrace of ad hoc, innovation-driven development as opposed to routinized, production-driven development.◄* **Alexander** *►Lutz [5] discusses challenges related to English as lingua franca (ELF) in Siemens.◄* **Alexander** *►House [6] explains why ELF does not threaten translation (common source in globalization etc).◄* **Alexander** *►Fewer [7] claims that the language is not the only barrier and that (at least in the academic context) we have an "academic imperialism" that ignores any kind of non-American science.◄*

III. Methodology

A. Data Extraction

The data extraction has been performed on March 6, 2016, and included data from November 2013 to February 2016 from the Stack Exchange (SE) data dump[9] and for this study, we considered the Portuguese version (SOPT). The XML files corresponding to the tags, users, and posts were transferred to a MySql database, through a R function per type of file (i.e., posts, users, and tags).

B. Data Preprocesing

We cleansed the tables eliminating few users were due to lack of data. None of these users had AccountId (i.e., user identifier for all stackExchange websites), LastAccessDate, WebsiteUrl, Location, UpVotes, DownVotes or Age. All of these users have the same display name (i.e., "a25bedc5-3d09-41b8-82fb-ea6c353d75ae"), and whenever they have a ProfileImageUrl, it is the same[10]. These accounts were created at different times from November 2015 to February 2016. We could not come up with a plausible reason for these anonymous users having the same display name but no other data, they do not seem to have anything in common. In total 3 SOPT users have been eliminated.

We focused on Brazilian users thus to identify their location we used `countryNameManager`[11].

C. Searching Process

Consequently, the locations were identified, and before starting the search a group of students were selected based on whether they had experience in use SO (in Spanish version) or GitHub, as detailed below:

The search started doing a manual inspection by each user profile based on *userId*, i.e. http://pt.stackoverflow.com/users/1919/, where **1919** is the *userId*. On the user profile, we looked the email address, if it was not available, we checked whether the user has a GitHub account or a personal web page.

- With Github account we found out the email address below *userName* if it was not on, we used a browser extension gitDiscovered[12] to discover the email address or we checked his/her public activity looking for at he/she did a git command [13], and then we searched the email address using a Github API [14] by *userName*.
- With the personal web page, we searched the email address on the section "about me" or *sobre me* in the Portuguese language.
- If none of the above, we used GitHub and searched by userName from SOPT, and compared profile picture, location, skills between SOPT and search results on GitHub, and then we followed the steps above mentioned with GitHub account, in order to find the user and get the email address.

In order to ensure the accuracy of results, we selected a random group of 15 users each 500 profiles and searched using the steps above mentioned. Whether we found new email addresses or missing information, we chose another one random group and applied the manual inspection again.

IV. Interview

In order to understand how Brazilians use the Portuguese version of StackOverflow, we decided to conduct a semi-structured interview because **Weslley** *► I WILL EXPLAIN THE REASON ABOUT WE CHOOSE THIS KIND OF INTERVIEW◄* . We interviewed 4 Brazilians developers who work in different regions of Brazil. One of these developers never used the

[9]https://archive.org/details/stackexchange

[10]https://www.gravatar.com/avatar/?s=128&d=identicon&r=PG&f=1

[11]https://github.com/tue-mdse/countryNameManager

[12]https://gitdiscovered.com/

[13]https://git-scm.com/docs/git-push

[14]https://api.github.com/users/userName/events/public

Portuguese version of StackOverflow, but we interviewed him just to get his point of view about the Portuguese version of StackOverflow. All of these interviews were conducted in Portuguese then they were translated to English. Both the Portuguese and English versions of the interviews can be downloaded in Weslley ►*Add the address*◄ .

Brazil is a big country and it might have different Weslley ►*I will add some word here that I dont know it yet =)*◄ for software development. To try to cover this diversity, we used the social media to call for developers who would like to participate of the interview. We select developers from Santa Catarina, São Paulo, Brasília and Pernambuco; south, southeast, center-west and northeast of Brazil, respectively. Weslley ►*I will check the international names of these States*◄

We recorded the audio of our first interview, however during the transcription process we realized that this was not the best approach to follow because it took too long Weslley ►*I will think in another sentence/word*◄ , this process can take up to eight hours per hour of audio as described by Hove et al.[8]. Thus, we decided to conduct the other interviews using the Skype chat.

It is not the first time that a instant message tool was used to conduct an interview [9]. This approach has been discussed in the social sciences [10], [11] Weslley ►*I will improve this paragraph*◄

talk about barriers [12].. Igor Thesis is about " Supporting newcomers to overcome the barriers to contribute to open source software projects "

metodology by [13]

The guidelines; there is no right/wrong answer; better record the interview. [14]

In our first interview, we interviewed a developer from Sao Paulo (Southeast of Brazil). The interviewed was made in , Subject 1

Subject 2

Subject 3

Subject 4

summary of out finds: Weslley ►*I will remove the names...*◄

- All of them complained about the Portuguese content. They think the English version is more complete.
- Marcio are not interesting in help the community.
- Alex and Karina said that if they had an account they would help others. Alex have an account but he did not have it when he had the chance to help.
- They do not make the search on SO, first they use google, then the google send them to SO.
- Marcio thinks that on-line translation tool is good enough, so he can use the English version without any problems.
- Karina and Marcio always find a solution for their problems, so they never had to make any question on SO.
- Alex thinks that the Portuguese version will soon not be necessary anymore, because he thinks that English is essential for those who work in the IT.
- Giovanni and Marcio thinks that some people do not use the PT version because they dont know that there is a PT version. Karina thinks that some people do not use it, because of the poor PT content.
- Giovanni prefer be more active in the PT version because it is new and needs more help. And Alex prefer be more active in the EN version because (according to him) English is the official language for software development.

Text from Igor's Theses.. remove it

"All the interviews followed a semi-structured script and were conducted using textual based chat tools, like Google Talk. We chose this mean once the participants are used to this kind of tool for their professional and personal activities. The interviews were conducted following three different scripts, used according to the participant?s profiles. The scripts were validated during the pilot interviews and by one specialist in qualitative studies, and one specialist in Open Source Software"

"We understand that the use of textual chat as the interview means can be considered a threat. The possibility of context change and the execution of parallel activities that distract the interviewees can be a negative aspect of using this mean. The use of Instant Messengers has been discussed in the social sciences (Opdenakker, 2006; Hinchcliffe and Gavin, 2009), and they point out that there is a set of positive effects of using these tools. In our case, we chose to use this means once the participants are used to the environment (they could choose the IM that they were more used to), and electronic means are the default (and preferred) way of communication in OSS projects."

Things to write:

- The type of interview, how it was conducted, why we did this way

The Criteria we used to select people - Brazilians from industry that use Stack Overflow in Portuguese.

- show the guideline - The guidelines; Make it clear that we said this to the interviewed: there is no right/wrong answer and if we could record the interview...

- write that all everyone that was interviewed agreed about use chat - skype -

V. Results

We identify the location of 7.264 users of SOPT which corresponds to 27% of its users. As we foresaw, most of the users of SOPT are located at Portuguese-speaking countries, in particular in Brazil (see Table I). Although there is a wide range of non-Portuguese speaking countries users, when looking at percentages these countries only represent 2

For half of the users whose location was identified, we could identify their gender (65%). Females are an overwhelming minority (4% SOPT users).

References

[1] S. Wang, D. Lo, and L. Jiang, "An Empirical Study on Developer Interactions in StackOverflow," *Proceedings of the ACM Symposium on Applied Computing*, pp. 1019–1024, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2480557{\%}5Cnhttp://www.scopus.com/inward/record.url?eid=2-s2.0-84877961377{\&}partnerID=40{\&}md5=71971adda488466a7c3fce96954f1eca

Country	Total
Brazil*	5954
Portugal*	599
United States	220
United Kingdom	80
Canada	44
France	25
Germany	42
India	30
The Netherlands	20
Mozambique*	14
Angola*	8
Cape Verde*	4
Other non Portuguese countries	224
None	19415

Table I
USER'S LOCATION IN SOPT. PORTUGUESE SPEAKING COUNTRIES ARE MARKED WITH AN ASTERISK.

ACKNOWLEDGMENT

The authors would like to thank...

[2] E. Carmel, "American hegemony in packaged software trade and the "culture of software"," *The Information Society*, vol. 13, no. 1, pp. 125–142, 1997. [Online]. Available: http://dx.doi.org/10.1080/019722497129322

[3] C. Tardy, "The role of english in scientific communication: lingua franca or tyrannosaurus rex?" *Journal of English for Academic Purposes*, vol. 3, no. 3, pp. 247–269, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1475158503000717

[4] U. Ammon, *The Dominance of English as a Language of Science: Effects on Other Languages and Language Communities*, ser. Contributions to the Sociology of Language [CSL]. De Gruyter, 2001. [Online]. Available: https://books.google.be/books?id=-qkUlGnAs0kC

[5] B. Lutz, "Linguistic challenges in global software development: Lessons learned in an international sw development division," in *2009 Fourth IEEE International Conference on Global Software Engineering*, July 2009, pp. 249–253.

[6] J. House, "English as a global lingua franca: A threat to multilingual communication and translation?" *Language Teaching*, vol. 47, no. 3, pp. 363–376, 007 2014. [Online]. Available: https://www.cambridge.org/core/article/english-as-a-global-lingua-franca-a-threat-to-multilingual-communication-and-translation/96BB816D14D24AE0313B4739D1FF12BE

[7] G. Fewer, "Beyond the language barrier," *Nature*, vol. 385, no. 6619, pp. 764–764, 2 1997. [Online]. Available: http://dx.doi.org/10.1038/385764a0

[8] S. E. Hove and B. Anda, "Experiences from conducting semi-structured interviews in empirical software engineering research," in *Proceedings of the 11th IEEE International Software Metrics Symposium*, ser. METRICS '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 23–. [Online]. Available: http://dx.doi.org/10.1109/METRICS.2005.24

[9] I. F. Steinmacher, "Supporting newcomers to overcome the barriers to contribute to open source software projects," Ph.D. dissertation, University of SÃ£o Paulo, 2 2015.

[10] V. Hinchcliffe and H. Gavin, "Social and Virtual Networks: Evaluating Synchronous Online Interviewing Using Instant Messenger," *The Qualitative Report*, vol. 14, no. 2, 2009. [Online]. Available: http://www.nova.edu/ssss/QR/QR14-2/hinchcliffe.pdf

[11] R. Opdenakker, "Advantages and disadvantages of four interview techniques in qualitative research," *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, vol. 7, no. 4, 2006. [Online]. Available: http://www.qualitative-research.net/index.php/fqs/article/view/175

[12] D. Ford, J. Smith, P. J. Guo, and C. Parnin, "Paradise unplugged: Identifying barriers for female participation on stack overflow," *International Symposium on the Foundations of Software Engineering (FSE)*, 2016.

[13] C. M. Gerpheide, R. R. Schiffelers, and A. Serebrenik, "Assessing and improving quality of qvto model transformations," *Software Quality Journal*, vol. 24, no. 3, pp. 797–834, Sep. 2016. [Online]. Available: http://dx.doi.org/10.1007/s11219-015-9280-8

[14] C. B. Seaman, "Qualitative methods in empirical studies of software engineering," *IEEE Trans. Softw. Eng.*, vol. 25, no. 4, pp. 557–572, Jul. 1999. [Online]. Available: http://dx.doi.org/10.1109/32.799955