Is Stack Overflow in Portuguese attractive for Brazilian users?

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

***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: Is there any evidence of an increase of popularity in SO-PT, and a decrease of popularity in SO among portuguese speaking users?

RQ2: Are the overlapping contributors more active than those who participate only in a single site?

RQ3: Is there any differences in behaviour among overlapping users depending on the website on which they participate?

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 ►*??*◄ .

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 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

[1] http://stackoverflow.com/

[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/

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.*◄

Compared to learners with English as their home-language their peers not having English as the home language **Alexander** ►*but those are kids at school*◄ are facing additional challenges when using computer platforms and software in English [8].

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.

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 Stack Overflow, we decided to conduct a semi-structured interview because in this type of interview it is possible to collect unexpected information. We interviewed 4 Brazilians developers who work in different regions of Brazil. One of these developers never used the Portuguese version of Stack Overflow, but we interviewed him just to get his point of

[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

view about the Portuguese version of Stack Overflow. All of these interviews were conducted in Portuguese then 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 patterns in the software development. In order to try to cover this diversity, we used the social media to call developers from the industry who would like to participate of the interview. We selected developers from Brasilia, Pernambuco, Santa Catarina and Sao Paulo; center-west, northeast, southeast, south of Brazil, respectively.

We recorded the audio of our first interview, as proposed by [9], however during the transcription process we realized that it was not the best approach to follow because it took too long. This process can take approximately two times the time of the original audio recordings[10] or even worse, up to eight hours per hour of audio as described by Hove et al.[11]. Thus, we decided to conduct the other interviews using some instant message software like Skype[15].

It is not the first time that an instant message tool was used to conduct an interview [12]. However, they said that using a textual chat the responses might be distract during the interview, because they can be executing other activities at the same time and it could be a threat. In our case, we do not think it as a threat because they answered the questions very fast. Furthermore, all of the responses agreed to use an instant message tool instead of video chat, actually one of them said that s/he felt more comfortable using text instead of video. This approach has been discussed in the social sciences [13], [14], and they pointed out more gains then loss.

During the interview we tried to make it as much informal as possible, because we think that the respondents could feel more comfortable answering questions in a "friend to friend" talk than in a formal strict interview.

We followed some of the guidelines proposed by [9], for instance, at the beginning of the interview, we said that the respondents would not be evaluated, therefore there were not wrong or right answers. We were very careful to make sure the respondents could feel relax. Following is the interview guide used:

1) How do you use Stack Overflow?
   a) Have you ever made a question in Stack Overflow? If yes, in which version?
2) Why do you use Stack Overflow in Portuguese instead of the English version?
3) Do you think that there are some specifics subjects/questions that are easier to find in the Portuguese version?
   a) (In case s/he complain about the content) what do you think about creating new content? Making questions and answering them, maybe translating the content from English to Portuguese.
4) Did you find any question that you know the answer? Did you answer it? (If not) Why not?
   a) Did it get accepted ? (If not) Why do you think it was not accepted?
5) Do you feel motivated in help other people? (If not) what could be done to motivate you help them?
6) Do you know any other people who use the Portuguese version of the Stack Overflow.
   a) (If not) Why do you think you are the only one who use it?
   b) Do you think that, even people that do not speak English, are using the English version with the help of an on-line translate tool?
7) In your opinion, what is the importance of Stack Overflow having the Portuguese version?

Weslley ►*Should I describe the respondents????*◄

Weslley ►*Write something here*◄ We summarized the finds:

- All of them complained about the Portuguese content. They think the English version is more complete.
- Subject 4 is not interesting in help the community.
- Subject 3 and 4 said that if they had an account they would help others. Subject 3 have an account but he did not have it when he had the chance to help.
- They do not make the search on Stack Overflow, first they google it, then Google shows some results from Stack Overflow.
- Subject 4 thinks that on-line translation tool is good enough, so s/he can use the English version without any problems.
- Subject 2 and 4 always find a solution for their problems, so they never had to make any question on Stack Overflow.
- Subject 3 thinks that the Portuguese version will soon not be necessary anymore, because s/he thinks that English is essential for those who work in the IT.
- Subject 1 and 4 thinks that some people do not use the Portuguese version because they do not know that there is this version. Subject 2 thinks that some people do not use it, because of the poor Portuguese content.
- Subject 1 prefer be more active in the Portuguese version because it is new and needs more help. And Subject 3 prefer be more active in the English version because (according to her/him) this is the official language for software development.

V. Survey

The survey was designed and reported by following the recomendations provided [15].

A. Goal and Research Questions

B. Respondents

Our respondents consisted of Brazilian users who have accounts in both English and Portuguese versions of the Stack Overflow website. We focused on Brazilian users because most of SO-PT users are located in Brazil.

[15]https://www.skype.com

C. Survey Structure

The survey was structured in blocks which grouped the questions into six topics:

1) Background: this refers to information about the person replying such as: occupation, mother tongue, Portuguese and English skills related to writing and reading, language used. This block of 5 questions.

2) Stack Overflow in Portuguese (SO-PT): the objective was characterize the respondents information about the platform usage. In particular we collected information concerning: time as user, type and frequency of contribution, during what software development activities they have used SO-PT, information sources used instead of SO-PT, electronic translation tools and their types, and whether the participation in SO-PT is different to the english website of Stack Overfkow (SO). This block of 8 questions.

3) Stack Overflow in English (SO-EN): the refers to information regarding the platform usage. The block included questions related to items such as: time as user, type and frequency of contribution, during what software development activities they have used SO-PT, information sources used instead of SO-PT, electronic translation tools and their types. This block of 7 questions.

4) Stack Overflow usage (Portuguese vs. English): this consisted of questions such as: ways of contributing and their frequency, factors that affect the choice of version of Stack Overflow, and content-wise quality in both websites.

5) Skills with Stack Overflow (SO-EN and SO-PT): this referes to information regarding skills and their types. The block included 5 questions.

6) Barriers in Stack Overflow (SO-EN and SO-PT): the objetive was know the main barriers to do not contribute in Stack Overflow (SO-EN and SO-PT). This barriers were found in [16]. The block included 2 questions.

D. Survey Design

Miguel ►*To address the research questions formulated*◄ , we drew up a survey consisting of 6 blocks of questions, with 34 questions in all in two versions Portuguese and English. Some questions were not presented to all individuals, as they were determined by the responses provided to other questions (i.e., conditional ones). Each person therefore answered a maximum of 26 questions. The electronic copy of the survey is available online in English at: https://goo.gl/aKvdQY and in Portuguese at: https://goo.gl/O8Iasd.

Most of the questions were measured using nominal scales, and a few others were measured with Likert scales, they were opened and closed questions. Some of them also included a space for extra information.

E. Survey Execution

The procedure followed consisted of the following steps:

1) The survey was online from January 24 to February 10 of 2017, using Google Forms [17].
2) Users were invited (via email) to participate the study.
3) After the surveys had been collected, analyses were performed, aiming to answer the Miguel ►*research questions*◄ . Data analysis was based on a quantitative analysis focusing mainly on descriptive statistics and percentages of the information collected.

VI. Results

A. Miguel ►*SO-PT Users*◄

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

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.

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

B. Survey

A total of 215 Brazilian users from the 1050 responded to the survey during the time it was online. This result is significant because of the difficulty normally involved in obtaining such a large quantity of individuals suitable for making up a target population.

1) Background: The majority of the respondents (92%) are ICT professionals, and only 8% have other professions such as students and academics. The ICT role most frequently played by the participants is that of software engineer/developer/system or database administrator (83%), and data scientist/machine learning developer/statistics or maths developer (3%). The remaining roles (technical coordinator, electronics technician, head of software development, web designer, software architect, etc.) are performed by less than 6% each. This fact indicates that the ICT professionals use Stack Overflow to ask/answer issues related to their software development activities.

We asked the respondents subjects related to: mother tongue, writing and reading skills in both Portuguese and English, and language used in their daily activities. The mother tongue of the respondents is Portuguese (99%), and the

remaining 1% is Finnish. Regarding writing and reading Portuguese skills, a significant percent of them have an excellent (native) level (94% in writing) and (96% in reading), and the remaining 6% in writing and 4% in reading have an advanced level. These results are coherent due to the official language in Brazil is the Portuguese. In the same way about writing and reading English skills. The level most frequently played by the participants is that of advanced (52% in writing) and (67% in reading), followed by intermediate (35% in writing) and (17% in reading), excellent (native) 13% and 6% respectively, and survival (level) 7% and 3% respectively. This fact indicates that an average of respondents uses English in their software development activities. Finally, a considerable part of the respondents use Portuguese in their homes (84%) and 87% at streets, banks, post offices, etc.; and Portuguese and English are used at work (62%) and school (65%). **Miguel** ►*In general terms, a large proportion of them uses their mother tongue because they feel more comfortable using it, but another group considers that they need to use both Portuguese and English due to they could be part of international software development companies and/or they could receive a bilingual education in their schools/universities*◄ asuumption (see table II).

	Portuguese		English	
	Writing	Reading	Writing	Reading
Survival level	-	-	7%	3%
Intermediate	-	-	35%	17%
Advanced	6%	4%	52%	67%
Excellent (Native)	94%	96%	6%	14%

Table II
LANGUAGE SKILLS IN SO-PT

2) Stack Overflow in Portuguese (SO-PT): The 38% of the respondents are users with between 1 and 2 years as members of SO-PT, followed by 33% with less than 1 year, and 29% with more than 3 years. This results have sense due to SO-PT was launched in 2013 and most of its users are from SO-EN.

Regarding contributions and frequency, the contribution types asked of the respondents were: creating questions, commenting on questions, answering questions, commenting on answers, editing questions and/or answers, voting up/down in questions and/or answers. The roles most frequently played by the participants are that of both never and almost never (about once a year) in all, followed by rarely (not more than once a month) and ocassionally (about once a week); and frequently (almost) was less chose. These results reported that SO-PT users do not use the Stack Overflow in Portuguese for these types of contributions (for more details see table III) **Miguel** ►*graph*◄ .

3) Stack Overflow in Portuguese (SO-EN): The majority of the respondents (40%) are users with between 4 and 6 years as members of SO-EN, followed by 29% with more than 6 years, and 13% with less than 1 year. The remaining 18% corresponds to some people did not answer all the questions (because of some conditional questions)

	Never	Almost Never	Rarely	Occasionally	Frequently
Creating questions	55%	26%	18%	1%	
Commenting on questions	35%	38%	20%	5%	2%
Answering questions	35%	36%	23%	4%	2%
Commenting on answers	41%	31%	23%	3%	1%
Editing questions and/or answers	54%	25%	16%	3%	1%
Voting up questions and/or answers	28%	24%	23%	18%	7%
Voting down questions and/or answers	39%	33%	17%	10%	1%

Table III
CONTRIBUTIONS IN SO-PT

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

[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-translat 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] G. B. Guðmundsdóttir, "The use ofict in south african classrooms and the double literacy trap," in *Educational Challenges in Multilingual Societies: LOITASA Phase Two Research*, Z. Desai, M. A. S. Qorro, and B. Brock-Utne, Eds. African Books Collective, 2010, pp. 147–172.

[9] 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

[10] 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

[11] 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

[12] 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.

[13] 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

[14] 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

[15] B. a. Kitchenham and S. L. Pfleeger, "Principles of Survey Research Part 2 : Designing a Survey Sample size Experimental designs," *Software Engineering Notes*, vol. 27, no. 1, pp. 18–20, 2002. [Online]. Available: http://portal.acm.org/citation.cfm?doid=566493.566495

[16] D. Ford, J. Smith, P. J. Guo, and C. Parnin, "Paradise unplugged: Identifying barriers for female participation on stack overflow," in *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, ser. FSE 2016. New York, NY, USA: ACM, 2016, pp. 846–857. [Online]. Available: http://doi.acm.org/10.1145/2950290.2950331

[17] Google, "Google Forms."

ACKNOWLEDGMENT

The authors would like to thank...