Is Stack Overflow in Portuguese attractive for Brazilian Users?

Miguel Botto-Tobar*
Eindhoven University of Technology
The Netherlands
m.a.botto.tobar@tue.nl

Weslley Torres
Eindhoven University of Technology
The Netherlands
w.silva.torres@tue.nl

Angela Lozano
HealthConnect
Belgium
angela.lozano.rodriguez@gmail.com

Mark G.J. van den Brand
Eindhoven University of Technology
The Netherlands
m.g.j.v.d.brand@tue.nl

Bogdan Vasilescu
Carnegie Mellon University
USA
vasilescu@cmu.edu

Alexander Serebrenik
Eindhoven University of Technology
The Netherlands
a.serebrenik@tue.nl

ABSTRACT

Stack Overflow (SO) is the reference for asking and answering programming-related questions. In early 2014 Stack Overflow em Português (SO-PT) was announced with the goal to reach developers that are not sufficiently proficient in the English language to fully participate in SO. Almost four years later we study how the simultaneous availability of SO and SO-PT impacted Brazilian software developers. A priori, the impact could have been either empowering or impeding. To address this question, we combine interviews, analysis of trace data from SO and SO-PT and a survey of 229 Brazilian software developers. Our results indicate that the developers recognize availability of the information, response speed and accessibility as strong points of SO, and lower barrier to entry and presence of Brazilian-specific information as strong points of SO-PT. In large, SO remains more popular than SO-PT, and SO-PT is not perceived as a viable alternative to SO.

CCS CONCEPTS

• **Social and professional topics** → **Geographic characteristics**; • **Software and its engineering** → **Collaboration in software development**; • **Human-centered computing** → *Empirical studies in collaborative and social computing*;

KEYWORDS

Interview; Qualitative Methods; Survey; Social Media/Online Communities

ACM Reference format:
Miguel Botto-Tobar, Weslley Torres, Angela Lozano, Mark G.J. van den Brand, Bogdan Vasilescu, and Alexander Serebrenik. 2018. Is Stack Overflow in Portuguese attractive for Brazilian Users?. In *Proceedings of ICGSE '18: 13th IEEE/ACM International Conference on Global Software Engineering , Gothenburg, Sweden, May 27–29, 2018 (ICGSE '18),* 9 pages.
https://doi.org/10.1145/3196369.3196377

*Also with University of Guayaquil, Ecuador.

ICGSE '18, May 27–29, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5717-3/18/05…$15.00
https://doi.org/10.1145/3196369.3196377

1 INTRODUCTION

Nowadays, developers are using online forums widely as a way to ask questions and/or answer others about different issues related to software development. However, many such online communities base their content on a unique language, English, and the choice for English as the community language might limit the access to knowledge for developers who are not fluent enough.

A priori, one might however argue that professional software developers should be able to express themselves in English since English is the language of software [8]. Indeed, whereas for artifacts and platforms targeting end users, such as user interfaces, user manuals, and support platforms, the need for translation and, broader, localization of software elements has been commonly recognized [1, 16], this is less obvious for artifacts and platforms targeting software developers. For instance, the "Java for Consumers" web page[1] exists in languages such as Dutch and French, while no such counterparts exist for the "Java for Developers" web page.[2] Still, the Java 8 documentation has been translated, *e.g.*, into Japanese,[3] the documentation of PostgreSQL 9.5 into Russian,[4] and there are companies dedicated to the translation of technical books, *e.g.*, Novatec is a Brazilian company specialized in translating O'Reilly books into Portuguese. Moreover, online developer communities exist, *e.g.*, in Spanish[5] and French,[6] and Stack Overflow (SO),[7] the largest Questions & Answers (Q&A) site targeting software developers in addition to English supports equivalent Q&A platforms in Portuguese, Russian, Spanish, and Japanese.

The question hence arises of the function of those non-English original or translated information sources in the different online developer communities. Do they empower developers by providing them with access to technological documentation? Do they impair developers' 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, and outdated? Are those information sources still relevant in 2018 despite the progress made in automatic translation?

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

The goal of our study is, therefore, to understand the motivations behind the usage of the English and non-English platforms supporting communities of software developers. To reduce the impact of the differences between the platforms pertaining to their organization or user interface, we focus on SO and Stack Overflow em Português (SO-PT).[8] SO-PT is the oldest non-English clone of SO and shares with SO its basic infrastructure, gamification mechanisms, and has partly overlapping contributors' community. Furthermore, we solely consider Brazilian software developers as they constitute the largest group of lusophone software developers in the world and the largest country-based group in SO-PT.

The research goal that guides our study is understanding what is the impact of creating SO in Portuguese on the Brazilian developer community. Since this research goal is broad, we start by posing **RQ1**: Do the Brazilian developers experience SO-PT as beneficial, detrimental, or neither? To answer RQ1, we divided it into four sub-questions:

- **SQ1.** Do Brazilian developers use SO-EN, SO-PT, or both?
- **SQ2.** Do they perceive SO-PT resources as scarce, erroneous, or out-dated?
- **SQ3.** How do they usually use each website?
- **SQ4.** What are the motivations to use SO?

Furthermore, recent advancement made in automatic translation made us wonder whether such localized sites as SO-PT are relevant in 2018. We pose thus **RQ2**: is SO-PT still relevant in 2018 despite the progress made in automatic translation?

2 RELATED WORK

Brazil's IT industry is large: according to the consultancy company A.T. Kearney it employes 1.7 million people [24]. Several recent studies have been dedicated to business opportunities for Brazilian software companies [3, 9], off-shoring and other global software development strategies they employ [23, 24], as well as the double role of the domestic market as a catalyst and an inhibitor of the industrial success [7]. Takhteyev [27] has studied how foreign forums are used in Brazil. One of his findings is that while many Brazilian IT developers can read the more technical texts with ease, this is not necessarily the case for more conceptual texts. Moreover, "it appears that many interviewees have difficulty expressing themselves in English (even in writing)", a barrier already recognized in the past [22, 29]. Indeed, Brazil is labeled as having "low proficiency" in English by the most recent edition of the EF English Proficiency Index.[9] Takhteyev also suggests that the lack of proficiency can encourage the use of texts already available online instead of creating new texts of their own; this would imply that Brazilian developers are more likely to passively explore such websites as SO and SO-PT instead of actively contributing to them. Finally, Takhteyev reports prejudice against the Portuguese sources present among Brazilian interviewees. This would suggest more negative attitude towards SO-PT as opposed to SO.

Observations that Brazilian software developers have difficulty expressing themselves in English suggest that the insufficient mastery of the English language might be a barrier to participation of Brazilian software developers in SO. In a closely related study of barriers to SO participation Ford et al. [12] identified through interviews and survey such barriers as "Nothing Left to Answer" and "Fear of Negative Feedback".

Importance of English as a foreign language when training future software developers has been stressed by Bakanova [5]. She argued that English has influenced modern programming languages far beyond the choice of specific words as keywords: the entire idea that a program consists of "unchangeable" units-words is influenced by the fact that Modern English is an analytic language that primarily conveys relations between words through their order or helper words such as prepositions; this is in sharp contrast with such fusional languages as Russian or Portuguese that use word modifications, i.e., inflections. Moreover, she has argued that the need to conceptualize the ideas in a foreign language furthers away the conceptualization from the reality and, hence, contributes to the development of abstraction skills.

This discussion of SO vs. SO-PT is also related to the question of the role of English as a neutral *lingua franca* or as a mechanism of domination [2, 11, 28]. 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 to alleviate scarcity of the non-English resources or their tardiness. If, however, English is seen as a domination mechanism, dominated developers need support to overcome the "invisibility" of their contributions and challenges. Carmel [8] discusses why English is the dominant language in software and lists such reasons as the "first-mover advantage", *i.e.*, the birth of software industry in the United States resulting in English becoming the lingua franca. Similarly, House [14] discusses challenges related to English as lingua franca (ELF) in Germany. To the best of our knowledge, the only work discussing ELF in the context of software development is the one of Lutz [18]. Lutz highlighted that move from the native language to ELF is not easy and is associated with "loss of power"; moreover, at least in the corporate context ELF is combined with company-specific terminology often derived from the native language [18].

3 METHODOLOGY

We use a mixed-methods approach as advocated by Easterbrook et al. [10]: We started by conducting a series of five interviews, four with Brazilian software developers and the fifth with the Brazilian SO community manager. Next, we performed a large scale data analysis based on the data dumps of the SO and SO-PT websites.[10] Finally, to get more profound insights in the differences of behavior on SO and on SO-PT, we surveyed software developers registered on both platforms and we additionally interviewed an user who has a high reputation with more than 1000 answers in SO-PT . These techniques allowed us to gather information about the Brazilian developer community, and then answer our research questions.

3.1 Interviews

In order to understand how Brazilians use SO-PT, we started by conducting 4 semi-structured interviews [25]. To cover the diversity schools of thought that might exist in Brazil, we interviewed

[8] http://pt.stackoverflow.com/

[9] https://www.ef.edu/epi/regions/latin-america/brazil/

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

Table 1: Information about the interviewees.

Interviewed	Age	Education level	# Years working in the industry	English Level	Interview duration	Manner
A	26	System Analysis and Development degree	7 years	Advanced *	40min	Skype
G	36	Master in Computer Science	7 years	B2 (CEFR) / 6.5 (IELTs)	20min	Face to face
K	25	System Analysis and Development degree	3 years	Basic*	50min	Skype
M	34	Computer Science degree	10 years	Intermediate*	30min	Skype
UserX	22	Studing Computer Science	5 years	Intermediate*	70min	Skype

* Level attributed by them, since they have never been formally evaluated. Detailed information about the interviewees can be found at https://goo.gl/NKHHvd.

four developers from different regions of Brazil: Brasilia, Pernambuco, Santa Catarina, and Sao Paulo, *i.e.*, center-west, northeast, south, and southeast of Brazil, respectively. Table 1 presents more information about the interviews and how each interview was conducted. Developers have been recruited through social media. One of these interviewees never used SO-PT, but we decided to interview her/him in order to understand the reason why s/he never used it. To obtain a complementary perspective we conducted two more interviews: one with the Brazilian community manager at Stack Overflow, and the another one with an user who accesses SO-PT regularly and has a high reputation. According to the statistics provided by SO-PT, this user (we are going to calling him UserX) has helped more them 400.000 people.

We followed the guidelines proposed by Seaman [25]. At the beginning of the interview, we clarified that there are no wrong or right answers. We kept the interview as 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. Figure 1 presents the interview guide.

We recorded the audio of our first interview. Since the next interviewee has indicated being more comfortable using instant messaging software, the remaining interviews have been conducted using instant messaging software. The use of instant messaging to conduct interviews has been discussed and found advantageous in the social science research [13, 21]. This view, applied to software engineering, is also shared by Steinmacher [26]. Steinmacher, however, points out that reliability of the responses obtained using instant messaging software might be threatened by the respondents being distracted during the interview due to simultaneous engagement in several activities. To reduce the impact of this threat we ensured that the respondents have rapidly answered the questions.

All interviews have been conducted in Portuguese and then translated to English. Both the Portuguese and English versions of the interviews can be found at https://goo.gl/hf58FE.

3.2 Data Analysis

A complementary perspective on the research questions we plan to answer can be obtained through an exploratory data analysis from SO and SO-PT available in the official Stack Exchange (SE)[11] data dump, in order to identify users that were overlapping (or not) with

(1) How do you use SO/SO-PT? Have you ever asked a question SO/SO-PT?
(2) Why do you use SO-PT instead of SO?
(3) On which subjects is it easier, and on which harder, to find answers on SO-PT compared to SO?
    - (if the respondent is not satisfied with the SO-PT content) What do you think about creating new content yourself, *e.g.*, translating from English to Portuguese?
(4) Did you encounter an unanswered question you could answer? Did you answer it?
    - If yes, was your answer accepted? If it was not accepted, why do you think it was not accepted?
    - If not, why not?
(5) Do you feel motivated to help other people? If not, what could be done to motivate it?
(6) Do you know other people using SO-PT?
    - If no, why do you think you are the only one among your peers using it?
    - Do you think that your peers that do not speak English using SO with the help of an on-line translation tool?
(7) In your opinion what is the importance of SO having the Portuguese version?

Figure 1: Interview guide.

Table 2: Data extraction outcomes.

	SO	SO-PT	Brazilian on SO-PT	Brazilian overlapping
Users	8,123,754	70,980	1,493	12,549
Posts	28,198,565	196,571	4,567	92,833
# Questions	10,663,884	88,310	2,838	27,640
# Answers	17,534,681	108,261	1,729	65,193
Tags	50,812	2,822		

the SO website, and also to analyze the contribution of bilingual and monolingual users.

We cleaned the SO/SO-PT data by eliminating 184 (183 in SO and 1 in SO-PT) anonymous users. None of these users had AccountId (*i.e.*, user identifier for all Stack Exchange websites), LastAccessDate, WebsiteUrl, Location, UpVotes, DownVotes or Age; all of them had the same display name (*i.e.*, "a25bedc5-3d09-41b8-82fb-ea6c353d75ae"), and whenever they had a ProfileImageUrl, it was the same.[12] These accounts were created at different times, from Nov 2015 to Oct 2017. We could not come identify why these anonymous users might have the same display name but no other data.

We identified 42,450 users that have accounts both on SO and SO-PT. We cleaned the free-text location profile data to extract countries,[13], identifying the location of 16,730 SO-PT users (or 24%; the remaining users did not fill in any location), including 15,150 having accounts both on SO and on SO-PT. By far the largest group among the users that have accounts both on SO-PT and on SO, and among those that only have an account on SO-PT, listed

[11] https://archive.org/details/stackexchange, as of December 1, 2017

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

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

Brazil as their location: 12,549 and 1,493, respectively (see Table 2). Consequently we focus only on Brazilian software developers in the remainder of the paper.

3.3 Survey

We aim to better understand the motivations of Brazilian software developers contributing to Stack Overflow (SO) and Stack Overflow in Portuguese (SO-PT), and their participation (or lack thereof) in the two platforms. To this end, we carried out a survey designed following the recommendations of Kitchenham and Pfleeger [17]. The survey[14] had two versions (Portuguese and English) and consisted of 34 questions, 12 of them open-ended question.

Firstly, we collected 3,225 email addresses of developers that had public information listed on their SO or SO-PT profile pages and had indicated Brazil as their location, and we then deployed the survey by inviting 1,050 of them personally by email (27 emails failed to deliver). After two weeks we received 216 answers (97 on the English survey and 118 on the Portuguese) and closed the survey. Analysis of the open-ended survey responses consisted of card sorting [19], a widely used technique for open coding [4].

4 RESULTS AND DISCUSSION

In this section, we address our two research questions by presenting the analysis and results of the quantitative and qualitative data gathered in our study.

4.1 RQ1: Brazilian developers' experience with SO and SO-PT

Recall that to answer **RQ1**, we divided it into **SQ1**–**SQ4**.

4.1.1 Which SO version is used? (SQ1). Stack Overflow is used by Brazilian developers on both versions English and its counterpart in Portuguese. This gives rise to the need for knowing which version of Stack Overflow they use English, Portuguese or both? 53 respondents of the English survey preferred SO, 12 preferred SO-PT and 5 do not have a clear preference for either of the platforms. Among the respondents of the Portuguese survey, 48 prefer SO, 5 prefer SO-PT and 11 do not have a clear preference for either of the platforms. Summarizing, respondents of both surveys have a clear preference for SO. This is particularly striking for the respondents of the Portuguese survey who are less proficient in English.

When asked about the reasons for their preference for SO, respondents indicate high quality of its content (82 on the English survey and 94 on the Portuguese software) and popularity (38 and 42, respectively). The latter point can be confirmed by observing the traffic popularity as reported by Quantcast.com: in January 2018 SO has received 12.9M views from Brazil as opposed to 1.7M views from Brazil received by SO-PT. We also observe that some respondents that do not have a conscious preference for one of the platforms might still be more actively engaged in SO than in SO-PT as the answers from SO might be ranked higher by the search engines: "I don't prefer SO-PT over SO. It's just a matter of in which one I find (the answer) first..." (R39). Furthermore, even though Brazilian developers are not necessarily proficient in English, they learn "almost automatically, do the searches in English" (I3) and subsequently the search engines refer them to SO.

[14] Available online in English at https://goo.gl/bMBSj1 and in Portuguese at https://goo.gl/N2oVXJ.

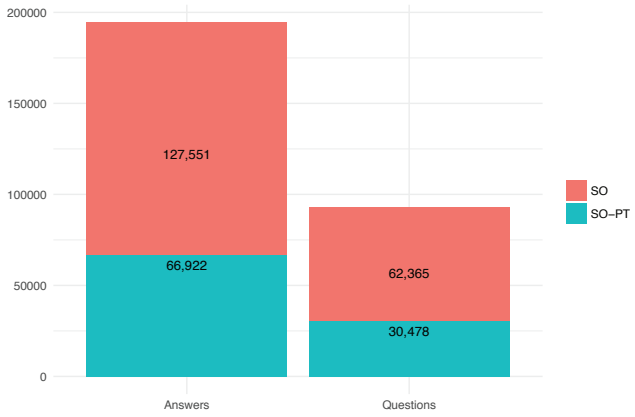

Figure 2: Brazilian Developers usage.

According to the data analysis SO is also more popular with Brazilian developers. It has more posts (questions and answers) than its counterpart in Portuguese (Figure 2). This fact is due to the population size; 38,939 (SO) against 14,042 (SO-PT) of Brazilian users registered, when they post a question the average of getting an answer in SO (1,8) is slightly higher to SO-PT (1,5), and acceptance rate of their answers is also high (31% vs. 29%).

4.1.2 SO-PT resources (SQ2). We asked the Brazilian developers through interviews and the survey, what is the status of SO-PT resources and whether there are topics for which it is easier to find answers on one platform compared to the other. Among the survey participants, 12 out of 97 respondents (EN) noted differences in quality of content between the two platforms, *e.g.*, answer rates and quality of answers:

> Respondent 13 (Survey EN): " ...there is a lower rate of answers and fewer frameworks/languages available [on SO-PT]".
>
> Respondent 8 (Survey EN) : "SO-PT has less quality than other Stack Overflow websites. I think the community is made by more people which have difficulty in the English version of SO thus they do not know SO's culture, acting "immature". Some poor questions and answers".
>
> Respondent 21 (Survey EN): "I've largely abandoned SO-PT shortly after its release for much less quality overall".

Concern about (actual or perceived) lower quality of the SO-PT content concurs with the earlier observation of Takhteyev [27] that Brazilian software developers are biased against Portuguese sources of information. Moreover, the impression that the SO-PT content quality is not as high as the English version was also mentioned during the interviews: differently from the survey, the quality aspects mentioned during the interviews were related to how easy it is to find an answer for their questions, rather than whether the answer was well written or not. According to the interviews, the possible reason that it is easier to find answers in SO than in SO-PT

is that there are more users in SO, and also that typical software error messages they get are in English so when they search on Google the first pages are from SO.

The community manager of SO-PT defended the quality of SO-PT content:

> "...There are brilliant people in the Portuguese version of StackOverflow, the quality of their answers is really good. Of course not every answer is excellent, but for sure you can find excellent answers..."

This perception is in agreement with the UserX. According to him, there are several high quality questions and answers:

> "...The quality is great. There are some really good questions/answers, for instance, related to: mathematics, logic, theory of computation, complexity of algorithms..."

This suggests that those developers who answered the survey might use SO-PT in a superficial way, since when they find an error in the software they use search engines to find the solution as soon as possible, usually landing on SO. We discuss in more detail how they use each web-site in Section 4.1.3.

4.1.3 Brazilian usage of SO and SO-PT (SQ3). Thus far is known the Brazilian developers are the largest community in SO-PT and they are also overlapping on SO, nevertheless, it is not known exactly how is their participation in each community (SQ3). In order to answer this question, we performed a single Wilcoxon rank-sum test on their data trace, we found out a similar answer to SQ1, which means they tend to ask/answer to a larger extent in the English version of Stack Overflow.

We also carried out the two-tailed Wilcoxon rank-sum test on the rating given to each contribution by survey respondents to compare each population. Since we compare different aspects of the same population we apply the Benjamini-Hochberg correction [6] to adjust the p-values for multiple comparisons.

Table 3 shows results related to SO-PT and SO contributions. All adjusted p-values not exceeding .05 have been typeset in boldface. We used 5 to indicate "Frequently (almost daily)" and 1 for "Never". The columns labeled EN and PT indicate the mode/median for English and Portuguese respectively. The column labeled ES indicates the effect size which was calculated by taking the absolute value of cliff's delta effect size. The last columns indicate the Likert distribution for English (EN Likert) Portuguese (PT Likert), and they start from "Never" (1) to "Frequently (almost daily)" (5).

We observe that while there are no statistically significant differences among the behavior frequency as reported in the Portuguese version of the survey, respondents that preferred to answer in English are more frequently asking questions and express encouragement (upvotes) on SO as opposed to SO-PT.

On the other hand, we also found out factors that could affect the way to the choice of which version the website of Stack Overflow (Portuguese/English) to contribute, these are detailed below:

- Content differences SO vs SO-PT because of the quality of interactions.

Table 3: Contributions on SO and SO-PT.

Contr.	p-value	EN	PT	ES	EN Likert	PT Likert
			Survey in English			
CrQ	**.021**	1/1	1/2	.2		
CoQ	.064	2/2	1/2	.2		
AQ	.127	2/2	1/2	.1		
CA	.064	1/2	1/2	.2		
EQA	.127	1/1	1/2	.1		
VUQA	**.021**	1/2	3/3	.2		
VDQA	.064	1/2	1/2	.5		
			Survey in Portuguese			
CrQ	.105	3/2	2/2	.5		
CoQ	.618	3/3	3/3	.6		
AQ	.154	3/3	3/2.5	.6		
CoA	.312	3/3	3/3	.6		
EQA	.365	3/3	3/2	.4		
VUQA	.105	5/4	4/4	.8		
VDQA	.882	3/3	3/3	.6		

Type of contributions (Contr): Creating Questions (CrQ), Commenting Questions (CoQ), Answering Questions (AQ), Commenting Answers (CoA), Editing Questions/Answers (EQA), Voting Up Questions/Answers (VUQA), and Voting Down Questions/Answers (VDQA).

- Language dominance also played an important role because most of respondents indicated the searchable content is available only in English, and hostility to users with less language proficiency.

According to the information gathered through the interviews, we noticed that they basically use both websites as a lurker, which means that they are not active in the communities, they neither answer nor ask questions. The reasons behind this behavior are mainly due to lack of available time to help the community answering questions, and another reason why they do not ask questions is that they do not need to. Usually, they find answers that can help them, without the need to create a question.

Although we have not found in the interviews any barrier, besides the lack of time, that could make developers not contribute in SO and SO-PT. We found nine survey respondents (eight in the survey in Portuguese and one in the survey in English) that do not feel confident/comfortable to make contributions (asking or answering questions) in SO. And three respondents said SO community is hostile, as described bellow:

> Respondent 98 (Survey PT) : "... I know several Brazilian developers who do not interact in the SO because they are not fluent in English. I am fluent in English, but still I was harassed just because I asked a not well formulated question. They immediately closed the question, they did not even give me the opportunity to improve it. So you can imagine how is it for those who do not speak English properly..."

The same hostile behavior was mentioned by UserX. He said that in one of the first time he tried to ask a question in SO. His question was closed. According to him, nobody helped him to improve the question in order to clarify it. Although he felt really bad about that, it did not stop him to use SO. He finalized his answer with the following:

Table 4: Kendall τ_b correlation between the survey questions related to the their English level and how often they contribute to SO/SO-PT.

Metrics	Survey in English		Survey in Portuguese	
	Correlation	p-value	Correlation	p-value
Answer Questions (SO) x English Reading Skills	0.309	2.56^{-3}	0.222	1.23^{-2}
Answer Questions (SO) x English Writing Skills	0.246	1.39^{-2}	0.256	4.19^{-3}
Answer Questions (SO-PT) x Comment Answers (SO-PT)	0.847	*	0.813	*
Ask Questions (SO) x Answer Questions (SO)	0.445	5.97^{-6}	0.504	8.38^{-9}

* The p-value is too small to be represented or computed exactly due to imprecision of the floating point calculations. The complete data can be found in the website https://goo.gl/kFEpAa.

> "... In SO-PT, when there is an unclear question, we talk to the author for several minutes, trying to understand what the author needs and we help him/her to edit the question..."

4.1.4 Why do they use English version? (SQ4). According to the results presented in the previous sections, the Brazilian community has the tendency to use the English version of Stack Overflow, however, this reason is still unclear. In order to answer this question (SQ4), we looked through some answers from the survey about why they prefer the version in English instead of the Portuguese, and we quoted some of them:

> Respondent 1 (Survey EN): "A much larger user base, and a higher chance of finding better answers/questions"

> Respondent 73 (Survey EN): "There's a lot more content (in SO) than SO-PT, I've been a member there for much longer and have invested time into writing good answers and questions, I have more privileges there"

We also found positive Kendall τ_b correlation between the English skills and SO contributions, see Table 4. For instance, according to the English version of the survey, the correlation between "answer questions (SO)" and "English writing skills level", and "answer questions (SO)" and "English reading skills level" are 0.246 (p-value 0.013) and 0.309 (p-value 0.002) respectively. Although these correlations are weak, they might indicate those users could contribute more often if their English level were higher.

The interviewed suggested that one of the reasons developers do not use SO-PT is just because they do not know there is a version in Portuguese. When the version in Portuguese was created, the English one was already popular.

We asked the SO-PT community manager the reason SO is more popular in Brazil than the Portuguese version. He answered the following:

> "... the English version is much older, it has 10 millions questions. And the Portuguese version has only 3 years and much less questions, and we don't advertise it too well... so when the Portuguese version was created the English version was already popular, with millions of accesses per day..."

One of the interviewed prefers to be more active in SO-PT because it is newer and needs more contributions. We found the same behavior in the survey. Which six respondents of the survey (one in the survey in English and 5 in the survey in Portugues) said that there are more opportunities to contribute in SO-PT (with answers) than SO.

However, another interviewee would prefer to be more active in the English version because (according to him) English is the official language for software development. Thus, in that sense, he could get an answer faster than in the Portuguese version. Which is almost in line with what SO-PT community manager said:

> "...any Brazilian programmer that speaks English knows that if he makes a question in the English version of StackOverflow, more people will be able to read it and answer it..."

Table 5: Percentage of Developers who use or not ETT.

	Survey in EN	Survey in PT
Use ETT	8.64%	15.46%
Do not use ETT	91.35%	84.54%

4.2 Information sources still relevant in 2018

In order to answer RQ2, we asked developers if they use electronic translation tools (ETT) to translate the content present in SO. Table 5 presents the results. Only 8.64% and 15.46% of developers who answered the survey in English, and Portuguese (respectively), admitted to use some ETT. Although the amount of developers might be low, it was what we expected from developers who answered the survey in English, since less than 10% have an intermediate level of reading in English. However, we expected a higher percentage from those who answered the Portuguese version of the survey, because 28,68% of them declared to have a survival to intermediate English reading level.

Figure 3 presents the most used ETTs. The most used ETT is "Translation Websites", which can indicate that developers were only interested on the translation of that specific word, instead of the definition and details.

In the interview phase, we asked what is the importance of SO-PT. One of the interviewees answered that it is important for those who do not speak English. However, they think that in the future, SO-PT is not going to be necessary anymore, because, according to them, knowing English is essential for those who work with technology and want to continue in this area. Other interviewees answered they are not fluent in English but it is not a problem since they use an ETT to translate those words they do not know. One of the interviewees gave 9 out of 10 as a score to those tools, in their opinion the ETT is good enough.

We asked the same question to the Stack Overflow community manager. For him, the importance of having a Portuguese version of SO is :

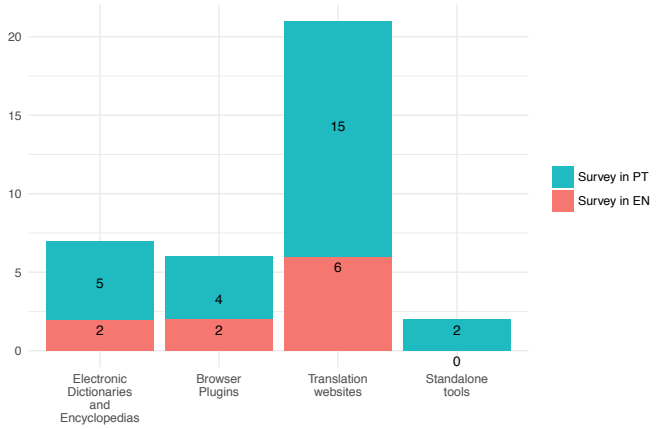

Figure 3: Use of Electronic Translation tools.

> "...to spread the knowledge to the ones who, for some reasons, didn't have opportunities to learn English, and because of that they couldn't have access to that kind of knowledge before, and now they have this opportunity..."

He also said that it is not only a matter of know the language itself, but it is more a matter of being part of the community. For him, it is not because one knows how to speak English that s/he will feel comfortable doing it or wants to do it. The same sense applies to be part of the community. It is not because one knows how to speak English that this one will be able to be part of the community.

In that sense, those information sources will be relevant, because there are barriers that ETTs cannot easily break.

4.3 Discussion

Our study has shown that Brazilian software developers prefer SO as opposed to SO-PT. The developers are aware of the challenges imposed by them not being able to express themselves fluently in English. However, they still prefer SO due to larger coverage of relevant topics, better answer quality, and a lager community of experts that might answer their questions. Not surprisingly, preference for SO is more present among the respondents that are more fluent in English, the latter being reflected in their choice for the survey questionnaire in English. An important subgroup of Brazilian software developers uses electronic translation tools.

Our observations concur and complement earlier results. Similarly to the findings of previous studies [22, 27, 29] some of our survey respondents recognize the English communication as the barrier to participation in SO. Moreover, bias against the Portuguese language sources reported by Takhteyev [27] might be (partly) responsible for the preference for SO as opposed to SO-PT. Furthermore, his conjecture that the lack of English language proficiency can encourage the use of texts already available online instead of creating new texts has been confirmed by the observed "lurking" behavior of Brazilian software developers on SO. They contribute to the community by asking or answering questions less frequently than they passively consume the answers already present on SO. Our results complement the existing ones by focusing on the specific knowledge sharing platform (SO vs. SO-PT) as well as by considering the use of electronic translation tools.

Our results have several implications for SO and SO-PT community managers as well as for future research. First, the success of localized platforms such as SO-PT might be jeopardized by developers mistrusting resources presented in their own language. Therefore, we argue that SO-PT should invest in creating a substantial number of high quality resources in the local languages, reflecting local technological preferences, in addition to creating the welcoming community that might sustain the website. Furthermore, researchers should investigate techniques for automatic support of non-English speaking developers, *e.g.*, by combining machine translation and chat bots [20]. Indeed, while the quality of machine translation is continuously improving, there are some barriers that electronic translation tools cannot easily overcome: developers proficient enough in English to understand the answer might still be more comfortable with formulating their questions in Portuguese.

5 THREATS TO VALIDITY

As any empirical study our work is subject to several threats to validity.

Construct validity pertains to our operationalisation of the construct of "attractiveness" of the website. Indeed, we juxtapose SO-PT and SO-EN and discuss *relative* attractiveness in terms of differences in the content, its perceived quality and availability, as well as differences between the communities surrounding each one of the websites. We have explicitly excluded from consideration features such as UI or gamification strategies that are shared by both websites. As such our conclusions are not necessarily adequately represent the idealized construct of "attractiveness". However, SO-EN is clearly the most popular programming Q&A website worldwide; therefore it can be seen as a valid benchmark to measure attractiveness of competing Q&A websites such as SO-PT.

Internal validity pertains to validity of the analysis machinery employed and inferences made. We have combined three different kinds of analysis: interviews, surveys and analysis of the data dumps. When it comes to the *interview*, one might argue that our choice for the instant messaging software could affect on the quality of the data. However, the use of instant messaging to conduct interviews has been discussed and found advantageous both in the social science research [13, 21] and in software engineering [26]. Hence, we believe the impact of this threat to be negligible. To ensure that the survey respondents understood the questions we have offered them the possibility to answer them in English and in Portuguese; furthermore, by selecting respondents among SO/SO-PT users we ensure that the survey participants have a basic familiarity with the concept of Q&A and SO/SO-PT. Moreover, the interviews have been conducted and the Portuguese survey data has been analyzed by the native speaker of Brazilian Portuguese eliminating threats that might have been introduced if assistance of an interpreter or translator might have been required [15]. The main issue that might affect the validity of our study is related to the sampling of participants. We did not carry out probabilistic sampling for the selection of respondents. Our recruitment strategy

could have incurred a possible selection bias (for example, a high probability of profile similarity among the respondents in SO/SO-PT). Furthermore, assessment of proficiency in English reading and writing is based on self-reporting and as such reflects perception of the proficiency rather than the actual proficiency.

External validity pertains to concerns related to generalization beyond the sample studied. While we expect that some of our findings might be transferable to other Q&A platforms such as SO-ES and other groups of software developers we consider this to be a topic of a follow-up study.

6 CONCLUSION

We identified that English is the language of software, and there are content and on-line communities with resources for non-English speaking developers, however, these are not enough. For instance, we studied one of non-English on-line communities, Stack Overflow in Portuguese. The content offered in the SO-PT has been rated in some of the cases as focused on software tools/approaches used in Brazil, it is therefore, could be seen as a limitation for other Portuguese speaking SO-PT users. Furthermore, the quality in on-line communities also plays a very important role. Brazilian developers responded the resources presented on SO-PT are lower quality or obsolete, which prevents many of them having access to a good solutions.

The use of Stack Overflow in its two versions (EN/PT) was one of the issues analyzed with Brazilian developers, whom in turn are overlapping on both SO and SO-PT. The results indicated they use more the English version instead. Indeed, some of them stated this finding, however, it would turn out an existential problem of SO-PT, since its the largest user community do not prefer to participate on it or they do in fewer amount, and whether this trend continuous its content could not be increased in the next years.

Even though the Electronic Translations Tools have made progress, resources for non-English speaking developers are still relevant in 2018: those resources can be beneficial when addressing the specifics of the local context as well as offer knowledge to developers who did not have opportunity to learn English.

Our results have several implications for SO and SO-PT community managers as well as for the researchers. We observe that success of localized platforms such as SO-PT might be jeopardized by developers mistrusting resources presented in their own language. Creation of such high quality resources and creation of the community that might make those resources sustainable is a necessary prerequisite to creation of successful Q&A platform. Furthermore, researchers should investigate techniques for automatic support of non-English speaking developers, *e.g.*, by combining ETT and chat bots.

ACKNOWLEDGMENTS

We are grateful to Brazilian software developers, interview and survey participants, for sharing with us their knowledge and opinions. This research was supported by the SENESCYT-Ecuador (scholarship program 2013-2).

A SURVEY RESULTS

The survey results are available at: http://www.win.tue.nl/~mbottoto/resources/icgse2018/survey.html

REFERENCES

[1] Sameer Abufardeh and Kenneth Magel. 2010. The impact of global software cultural and linguistic aspects on Global Software Development process (GSD): Issues and challenges. In *4th International Conference on New Trends in Information Science and Service Science*. IEEE, Piscataway, NJ, USA, 133–138.

[2] Ulrich Ammon (Ed.). 2001. *The Dominance of English as a Language of Science: Effects on Other Languages and Language Communities*. Number 84 in Contributions to the Sociology of Language. De Gruyter, Berlin. https://books.google.be/books?id=-qkUlGnAs0kC

[3] Ashish Arora and Alfonso Gambardella. 2005. The Globalization of the Software Industry: Perspectives and Opportunities for Developed and Developing Countries. *Innovation Policy and the Economy* 5 (2005), 1–32. https://doi.org/10.1086/ipe.5.25056169

[4] Alberto Bacchelli and Christian Bird. 2013. Expectations, Outcomes, and Challenges of Modern Code Review. In *Proceedings of the 2013 International Conference on Software Engineering (ICSE '13)*. IEEE Press, Piscataway, NJ, USA, 712–721. http://dl.acm.org/citation.cfm?id=2486788.2486882

[5] M. V. Bakanova. 2011. On the necessity of learning the English language by students-future programmers. *Izv. Penz. gos. pedagog. univ. im. V. G. Belinskogo* 24 (2011), 540–543. (in Russian).

[6] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300. http://www.jstor.org/stable/2346101

[7] Antonio J. Junqueira Botelho, Giancarlo Stefaunto, and Francisco Veloso. 2006. *The Brazilian Software Industry*. Oxford University Press.

[8] Erran Carmel. 1997. American Hegemony in Packaged Software Trade and the "Culture of Software". *The Information Society* 13, 1 (1997), 125–142. https://doi.org/10.1080/019722497129322 arXiv:http://dx.doi.org/10.1080/019722497129322

[9] Sandro LuÃŋs Diesel Cortezia and Yeda Swirski de Souza. 2011. An analysis of the internationalization of small Brazilian software companies. *Brazilian Business Review* 8, 4 (2011), 23–43.

[10] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. 2008. Selecting Empirical Methods for Software Engineering Research. In *Guide to Advanced Empirical Software Engineering*, Forrest Shull, Janice Singer, and Dag I. K. Sjøberg (Eds.). Springer London, London, 285–311. https://doi.org/10.1007/978-1-84800-044-5_11

[11] G. Fewer. 1997. Beyond the language barrier. *Nature* 385, 6619 (27 2 1997), 764–764. http://dx.doi.org/10.1038/385764a0

[12] Denae Ford, Justin Smith, Philip J. Guo, and Chris Parnin. 2016. 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 (FSE 2016)*. ACM, New York, NY, USA, 846–857. https://doi.org/10.1145/2950290.2950331

[13] Vanessa Hinchcliffe and Helen Gavin. 2009. Social and Virtual Networks: Evaluating Synchronous Online Interviewing Using Instant Messenger. *The Qualitative Report* 14, 2 (2009), 318–340. http://www.nova.edu/ssss/QR/QR14-2/hinchcliffe.pdf

[14] Juliane House. 2014. English as a global lingua franca: A threat to multilingual communication and translation? *Language Teaching* 47, 3 (001 007 2014), 363–376. https://doi.org/10.1017/S0261444812000043

[15] Inez Kapborg and Carina Berterö. 2002. Using an interpreter in qualitative interviews: does it threaten validity? *Nursing Inquiry* 9, 1 (2002), 52–56. https://doi.org/10.1046/j.1440-1800.2002.00127.x

[16] Gregory E. Kersten, Mik A. Kersten, and Wojciech M. Rakowski. 2002. Software and Culture: Beyond the Internationalization of the Interface. *Journal of Global Information Management* 10, 4 (2002), 86–101.

[17] Barbara a Kitchenham and Shari Lawrence Pfleeger. 2002. Principles of Survey Research Part 2 : Designing a Survey Sample size Experimental designs. *Software Engineering Notes* 27, 1 (2002), 18–20. https://doi.org/10.1145/566493.566495

[18] Benedikt Lutz. 2009. Linguistic Challenges in Global Software Development: Lessons Learned in an International SW Development Division. In *IEEE International Conference on Global Software Engineering*. IEEE, Piscataway, NJ, USA, 249–253. https://doi.org/10.1109/ICGSE.2009.33

[19] T. Menzies, L. Williams, and T. Zimmermann. 2016. Perspectives on data science for software engineering. In *Perspectives on Data Science for Software Engineering*, Tim Menzies, Laurie Williams, and Thomas Zimmermann (Eds.). Morgan Kaufmann, Boston, 3 – 6. https://doi.org/10.1016/B978-0-12-804206-9.00001-5

[20] Alessandro Murgia, Daan Janssens, Serge Demeyer, and Bogdan Vasilescu. 2016. Among the Machines: Human-Bot Interaction on Social Q&A Websites. In *CHI Conference on Human Factors in Computing Systems (CHI Extended Abstracts)*. ACM, 1272–1279. https://doi.org/10.1145/2851581.2892311

[21] Raymond Opdenakker. 2006. Advantages and Disadvantages of Four Interview Techniques in Qualitative Research. *Forum Qualitative Sozialforschung / Forum:*

Qualitative Social Research 7, 4, Article 11 (2006), 13 pages. https://doi.org/10.17169/fqs-7.4.175

[22] James B. Pick, Martha Garcia Murillo, and Carlos J. Navarrete. 2007. Information technology research in Latin America: Editorial introduction to the special issue. *Information Technology for Development* 13, 3 (2007), 207–216. https://doi.org/10.1002/itdj.20070

[23] Rafael Prikladnicki, Jorge Luis N. Audy, Danela Damian, and Toacy C. de Oliveira. 2007. Distributed Software Development: Practices and challenges in different business strategies of offshoring and onshoring. In *International Conference on Global Software Engineering (ICGSE 2007)*. 262–274. https://doi.org/10.1109/ICGSE.2007.19

[24] Rafael Prikladnicki and Erran Carmel. 2013. Is time-zone proximity an advantage for software development? The case of the Brazilian IT industry. In *35th International Conference on Software Engineering (ICSE)*. 973–981. https://doi.org/10.1109/ICSE.2013.6606647

[25] Carolyn B. Seaman. 1999. Qualitative Methods in Empirical Studies of Software Engineering. *IEEE Trans. Softw. Eng.* 25, 4 (July 1999), 557–572. https://doi.org/10.1109/32.799955

[26] Igor Fábio Steinmacher. 2015. *Supporting newcomers to overcome the barriers to contribute to open source software projects.* Ph.D. Dissertation. University of São Paulo.

[27] Y. Takhteyev. 2007. Using Foreign Forums. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*. 79–79. https://doi.org/10.1109/HICSS.2007.594

[28] Christine Tardy. 2004. The role of English in scientific communication: lingua franca or Tyrannosaurus rex? *Journal of English for Academic Purposes* 3, 3 (2004), 247–269. https://doi.org/10.1016/j.jeap.2003.10.001

[29] Kival C. Weber, Roberto A. R. Almeida, Danilo Scalet, and Vanderlei V. Ortêncio. 1999. Software standardization process in Brazil. In *Software Engineering Standards, 1999. Proceedings. Fourth IEEE International Symposium and Forum on*. 9–15. https://doi.org/10.1109/SESS.1999.766573