Spring 3-20-2018

Different Languages, Different Questions: Language Versioning in Q&A

Andrew Vargo
vargo@kcg.ac.jp

Benjamin Tag
tagbenja@kmd.keio.ac.jp

Kai Kunze
kai.kunze@kmd.keio.ac.jp

Shigeo Matsubara
matsubara@i.kyoto-u.ac.jp

Different Languages, Different Questions: Language Versioning in Q&A

Andrew W. Vargo
The Kyoto College of Graduate Studies for Informatics
Email: vargo@kcg.ac.jp

Benjamin Tag
Graduate School of Media Design, Keio University
Email: tagbenja@kmd.keio.ac.jp

Kai Kunze
Graduate School of Media Design, Keio University
Email: kai.kunze@kmd.keio.ac.jp

Shigeo Matsubara
Graduate School of Informatics, Kyoto University
Email: matsubara@i.kyoto-u.ac.jp

Question and Answering (Q&A) communities have become effective forums for humans to collaborate and build accurate domain-specific archives of information. Stack Overflow is a prime example of a system which has effectively leveraged Q&A to build a strong archive of computer programming information. However, English is the dominant language in size and scope. To reach a wider audience, Stack Overflow has started language-specific sites. In this paper, we seek to understand how these language version sites are used, and whether they form unique Q&A structures or mirror the English version. The results indicate that each site is structured differently, and that users of different languages have different question asking patterns. The contributions from this work are useful in informing designers of systems attempting to conduct language versioning and provide an argument for developing sites within languages, rather than only providing translated versions.

Keywords: Collaborative Computing Systems; Language Versioning, CSCW

1.0 Introduction

Large distributed knowledge-sharing platforms, such as Wikipedia, often have a desire to expand their services and repositories into other languages (Bao et al., 2012). This desire can serve multiple complementary purposes including expanding the reach of the system to a diverse number of languages, creating more traffic to generate potential income, and supporting egalitarian aims, such as the spread of knowledge to information poor communities. However, language versioning is not an easy task to undertake. The auxiliary sites are often significantly smaller than the parent site and may suffer from lower participation rates and inferior or limited content (Bao et al., 2012). This is an important problem since individuals may rely on these distributed knowledge-sharing communities for personal and professional growth. Information aggregating and sharing systems must decide whether to expend energy on supporting full-fledged systems in different languages or rely on improving machine translation.

In this paper we look at a large scale technical Q&A system that is attempting to create language versions of its system to benefit non-English speakers. In specific, we study three language sites including Russian, Portuguese, and Japanese. These are operated by the highly successful English-base site Stack Overflow (SO). SO is a popular and important question-answer (Q&A) community for computer programming that was established in 2008. As of December 2017, SO averages 8.3 million visits per day and has a community of over 6.6 million users who have asked 13 million questions and have given 21 million answers. The community also successfully resolves 72% of all questions.

One issue for the SO community is that it has been a monolingual English site since its inception in 2008. There are significant challenges for non-native English speakers in using SO, such as issues with terminology and comfort levels with engaging in the community (C. Treude, Prolo, & Filho, 2015; Xu et al., 2016). Because of this, participation rates among countries in SO are affected by culture and English proficiency (Oliveira, Andrade, & Reinecke, 2016). SO started beta versions of language specific sites to alleviate these issues.

An important concern is how the sites create their archives in relation to each other. SO aims to create a highly accurate and complete archive of computer programming questions, and this archive is expected to be duplicated across all language sites (Hanlon, 2014). A stated concern from community members is that the different language versions could result with different language communities talking about different topics and technologies (Maciel, 2014). This is an issue for users who believe in having a universal authority for computer programming Q&A. Instead of having all question topics being centralized in English SO, there would instead be areas of conversation that are focused in different languages.

For a Q&A system Stack Overflow, is it realistic to assume that different languages result in similarly structured question archives? If the answer is “yes”, then translating the archives may make sense given the lack of scale for most languages. If the answer is “no”, then steps may be needed to make sure that basic concepts are covered in each language corpus. In this manuscript, we present an analysis of questions across four question sets taken from a two-year period. By investigating the questions through user-generated tags, we find that the question archives are all significantly different from each other. The research questions and contributions can be summed up as follows:

- **Do the language versions of Stack Overflow ask different questions?** The content of questions can be determined via Stack Overflow’s universal tagging system. After translating and normalizing tags, we used the log-likelihood (LL) to determine similarity between popular tags in the corpora. Adjusted alpha results from both group and pairwise tests show that the corpora have different tags from each other.

- **Do the language versions of Stack Overflow present different related tag structures for universally popular tags?** Even if the structure of a popular tag is different for each site, the structure for tags that are popular across all sites could be similar. We chose “Javascript”, a tag in the top 3 for each language version, and mapped the co-related tags. The results show that central tags are different for each version. This indicates that there is a unique question-asking paradigm for each language.

Our paper investigates the impact language versioning has on domain-specific Q&A. To do this, we present the framework for the quantitative study by presenting an overview of the language versioning process and its users before presenting the results. We also present a discussion of the importance of the results in informing translation efforts as well as providing a justification for language versioning efforts in collaborative information gathering communities.

2.0 Background

In this section, we provide an overview of previous studies on language versioning. We then provide background on Stack Overflow's reasoning and process behind conducting language versioning. We also provide a brief case study of the Japanese site that provides background on the motivations of the users of a language versioned site. Finally, we present our research questions based on the presentation of the background.

2.1 Language Versioning

We might first assume that languages share similar information across boundaries, but this is shown to be false in social networks, where there are vast differences in content and production (Hong, Convertino, & Chi, 2011). This makes intuitive sense if we consider that different cultures will have different norms and ways of communication. Peer production communities exhibit these differences, including differences in what would seem to be universal fact-based information, such as that found in an online encyclopaedia. Hecht (Hecht & Gergle, 2010) and Bao (Bao et al., 2012) found that the differences between language versions of Wikipedia are significant. Cultural viewpoints seem to greatly affect the way that information is conceived and reported. However, whether this applies to a Q&A site like Stack Overflow (SO) and Japanese Stack Overflow (JSO) is unknown. We do know from previous research that the vast number of questions in sites like SO are those which have concrete answers (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2012; Christoph Treude, Barzilay, & Storey, 2011). Would computer programmers, who are attempting to navigate real world programming issues (Christoph Treude et al., 2011), ask materially different questions? This question has not been studied in detail in previous research.

2.2 Stack Overflow's Process of Language Versioning

SO is a large popular Q&A community that is focused solely on computer programming. Much of its success has been attributed to its incentive system which allows users to vote on each other's content and thus give and take away reputation points (Mamykina, Manoim, Mittal, Hripcsak, & Hartmann, 2011). These reputation points have been shown to be a strong incentive for participation (Tausczik & Pennebaker, 2012) and have been exported to all of SO's related sites. Another reason for its success has been its ability to command and maintain a site with user curated material based on clear rules (Correa & Sureka, 2013; Li, Zhu, Lu, Ding, & Gu, 2015; Mamykina et al., 2011).

One of these rules is the mandate that all questions must be in English and must be clear and understandable. This, of course, creates a significant barrier for those who do not read and write English at a level which facilitates contributing. One of the founders of SO argued that programming essentially requires the use of English as a *lingua franca* (“The Ugly American Programmer”, 2009). However, as time progressed, there was a move towards language versioning.

In 2014, SO decided to develop other language versions in-house. An independently created version of SO had been developed in Russian in 2012. The site, finally integrated with Stack Exchange in 2015, boasted over 55,000 questions, 29,000 users, and 31,000 visits per day. This provided evidence that a non-English site-programming community could work.

The justification for language versioning is to reach more users. For instance, the introductory post for the Portuguese language site argues that 10% of the world’s programmers are in China, but only 4.8% of visits come from China, Japan, and Korea combined (Hanlon, 2014). These numbers, the SO administration argues, have been because of language constraints. The administrators stated when the first beta language version in Portuguese was launched they expected to maintain centralization due to the critical mass of the English site and they expected almost every question that was asked on the language version site to also be asked and answered on the English site (Hanlon, 2014).

Site	Questions	Users	Age in Months
English	14M	7.1M	106
Russian	147K	74K	50
Portuguese	70K	50K	32
Japanese	12K	12K	26

Table 1. List of Stack Overflow Language Versions at the Beginning of 2017

Japanese was considered a good candidate due to the strength of its programming community. Japanese Stack Overflow (JSO) went public in October 2014, and the site was announced on the SO official blog on December 15, 2014. The announcement was a topic of discussion on the SO blog; 100 comments were posted by 49 unique users, many of whom addressed the pros and cons of launching the site. In order to better understand the arguments of the community, the first and second authors conducted a Grounded Theory session (Glaser & Strauss, 2009). The open coding session found 13 codes which coalesced into 5 themes:

- **Splintering**: The community will be split, will cover different technologies, and will decrease the amount of participation from Japanese members.
- **English is a necessity for advanced programming**: Japanese usage will stunt the growth of novice programmers.
- **Reduction of Poor English Questions on SO**: Fewer limited English speakers will contribute bad questions.
- **Gateway to SO**: Users will use JSO and other language sites and move to the main SO site.
- **Will Broaden Knowledge**: JSO will discuss uncommon problems which will make their way to SO.

The first and fifth themes were influential in informing this study. First, they contradict the official narrative SO had put forth themselves with the launch of the language sites. These commenters do not assume the language versions will be mirror copies of the main site, but rather that specific topics and technologies may be discussed instead. In addition, the first theme contained a sub-theme which argued for translating the main SO site, rather than creating separate Q&A communities.

2.3 Motivation for Using a Language Specific Site: Case Study of Japanese Stack Overflow

Why would a user choose to participate in a language specific site rather than the larger and more active English site? For some users the answer is obvious; they lack the necessary proficiency to ask and answer questions on the English site. However, we can see from Japanese Stack Overflow (JSO), that many of the early contributing users were active contributors on English Stack Overflow (ESO). In the first 6 months of JSO, 265 users were active on ESO as well. This is in comparison to 1,078 users who were only active on JSO. In addition, users who were in the top 5% of all ESO users were also the most active in asking questions and providing answers. Table 2 shows the cumulative averages for the two categories of JSO contributors.

The data in Table 2 shows us that the formation of JSO is connected to the ESO community beyond name and site infrastructure. To understand the nature of this connections, we conducted interviews with JSO users.

Category	Users	Answers	Questions

JSO Only	938	M=1.8
		SD=6.51
			M=1.6
			SD=3.49

ESO Top 1%	39	M=7.3
		SD=11.6
			M=2
			SD=6.2

Table 2. Japanese Stack Overflow Users

We obtained the email addresses of all 74 users who listed their address in their public profiles, whom we then asked for interviews. We were able to gather ten participants. Interviews were conducted over email. The first and third authors conducted semi-structured interviews over a two-week period. The interviewers could ask more specific questions when they felt it was appropriate. In total, 186 questions and responses were logged. Eight of the interviewees had accounts on SO as well as JSO, while two only had accounts on JSO.

The interview questions had themes including: 1) Programming history and language abilities, 2) the motivation for participation on JSO, 3) the relationship between JSO and ESO, 4) the value of reputation on the sites. We used a Grounded Theory method (Glaser & Strauss, 2009) to analyse the data. Special care was taken to represent the Japanese responses as best as possible. Analysis of the interviews identified several major themes as shown in Table 3.

Overall, the results of the interview study indicate that the connected users between ESO and JSO see the new language site as a gateway for beginner programmers, and not as a replacement for ESO. From the interviews, we would expect that questions on

JSO would be more basic or fundamental computer programming questions compared to ESO. It is interesting, since it does indicate that while ESO's question archive was developed as a natural progression of supply and demand, JSO has a more directed path.

This case study is useful for framing how one site, JSO, came into existence. We can see how the early stages of the site were born out of a movement from the dominant site. It is important to consider that the origins of each site may affect how each community acts and what questions they ask. Portuguese Stack Overflow (PSO) has a very similar background to JSO, having been developed by ESO site users (Hanlon, 2014). Russian Stack Overflow (RSO) has a much different background from JSO and PSO, having been started outside the community.

These are important distinctions to make. Should we consider that PSO and JSO will be more similar due to their shared origins, or should we consider that RSO and ESO will be similar due to their independence of development? It is difficult to measure the impact of origin directly, but it is a variable that must be considered when discussing results.

Theme	Example
Supporting Japanese IT	"(I joined) *to support the framework for beginner programmers." (User5 in Japanese)* *"As a native Japanese speaking developer, I'd like to contribute to the Japanese IT industry." (User 7 in English)*
ESO is used for problem-solving	*"I do not use the Japanese version to resolve problems as much as the English version." (User 4 in Japanese)* *"I answer on both sites, but there is a difference in content. In the English version, I answered specific questions. In the Japanese version, I answer the more common questions." (User 3 in Japanese)*
Reputation points are not important on JSO	*"Reputation is more valuable on the main site because it is global." (User 1 in Japanese)* *"I don't care at all about the reputation points on the Japanese site" (User 6 in English)*

Table 3: Examples of Responses from Interviews

2.4 Discussion

On one hand, it seems likely that the different languages will ask similar questions to each other due to the limited domain. Unlike Wikipedia, where culture greatly determines what or who is important (Bao et al., 2012), computer programming Q&A is focused on solving the problems of people searching for real and practical solutions (Anderson et al., 2012; Christoph Treude et al., 2011). On the other hand, our interview study showed that there was significant influence from the primary language site, which may affect the questions asked.

Even if the ratio of question topics asked in each site is different (or similar), are the related topics similar? Very few questions have one user-generated tag to describe it. A universally popular tag could have similar related tags across languages. That would indicate that the questioning process across languages are similar. The opposite would indicate that the different language sets are asking different types of questions.

3. Question Data

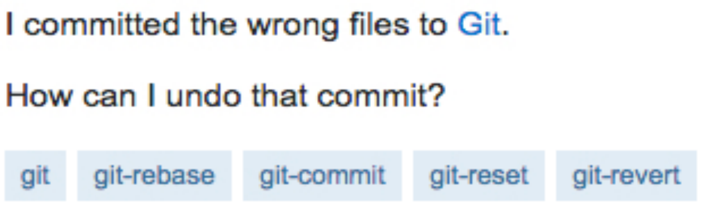
In this section, we describe the data that is used to analyse the question corpora of English Stack Overflow (ESO), Portuguese Stack Overflow (PSO), Russian Stack Overflow) RSO, and Japanese Stack Overflow (JSO). We then discuss the appropriateness of using tags as a data point to compare question corpora. Finally, we explain the process of normalization for the tags.

3.1 Question Tags

A challenge in comparing questions asked in different language corpora is determining the data for comparison. Ideally, entire questions and their content could be analysed and compared on exact terms. However, this is not only difficult given large data sets, but also may obfuscate similarities. That is, language used to describe a question may be different, but the content is not. To gain an overview of similarity between the question corpora, we use question tags which describe the main topics of the questions.

In all Stack Overflow sites there is a system for tagging questions. The purpose of this system is to describe the topic accurately, allow experts to quickly identify questions, and to allow for useful indexing of questions (“What are tags, and how should I use them?”). Users can choose up to five tags to describe their question, and they are required to use at least one tag. Tags are not ordered by importance, rather they must all be relevant to the question. They are, however, ordered by popularity (“What are tags, and how should I use them?”). Tags are then displayed at the bottom of the question body, as shown in Figure 1.

The example in Figure 1 shows the wide range of options that are available in tagging. Tags can be very specific or broad and allow users and the community to appropriately choose. As such, the number of tags depends on the size of the community. For example, in ESO there are 43,085 tags that have been used at least once.



I committed the wrong files to Git.

How can I undo that commit?

Figure 1. Example of a Question and Its Tags

Too many tags and variations would make taxonomy via tags unmanageable. In order to combat this, the administration started the process of creating a tag master list in which synonyms are merged together (“What are tag synonyms and merged tags?,”). For example, if a user were asking a question about PageRank Algorithm, it would be logical to tag the question either “Algorithm”, “Algorithms” or both. The tag master list avoids these redundancies by merging synonyms into the master tag. Figure 2 shows an example of a master and synonym relationship. This system has been replicated across all languages.

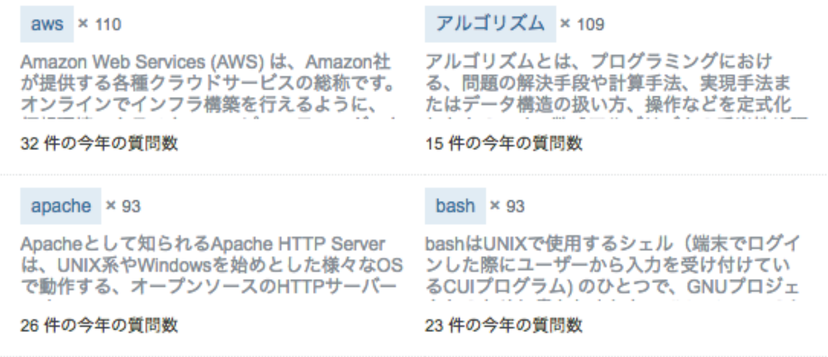
Master	←	Synonym	Creator	Renames
.htaccess × 57420		htaccess	Jul 26 '10 at 19:02	10410

Figure 2. Example of Master and Synonym Tags

A concern with using tags for analysis is a reliance on their accuracy. Can users be trusted to accurately tag their questions? Editing of questions by qualified members of the community helps assure that tags are accurate (Li et al., 2015). The communities are encouraged to search out inaccurate tags and change them to conform to community standards, and they do so actively (Correa & Sureka, 2014; Li et al., 2015; Vargo & Matsubara, 2016).

3.2 Tag Normalization

An obvious issue that hinders analysis is the difference in language between the four sites. As Figure 3 shows, most tags in each language site are in English. However, some tags are presented in the site language. In Figure 3, we see the tag “アルゴリズム” which translates directly into “Algorithm”. A corpora comparison without translation would be useless.



aws × 110

Amazon Web Services (AWS) は、Amazon社が提供する各種クラウドサービスの総称です。オンラインでインフラ構築を行えるように、

32 件の今年の質問数

アルゴリズム × 109

アルゴリズムとは、プログラミングにおける、問題の解決手段や計算手法、実現手法またはデータ構造の扱い方、操作などを定式化

15 件の今年の質問数

apache × 93

Apacheとして知られるApache HTTP Serverは、UNIX系やWindowsを始めとした様々なOSで動作する、オープンソースのHTTPサーバー

26 件の今年の質問数

bash × 93

bashはUNIXで使用するシェル（端末でログインした際にユーザーから入力を受け付けているCUIプログラム）のひとつで、GNUプロジェ

23 件の今年の質問数

Figure 3. Tags on Japanese Stack Overflow: The Top Right Tag is Japanese for Algorithm

Using the master synonym list, we translated tags in Portuguese, Russian, and Japanese into English. The results of translations were checked against machine translation to ensure accuracy.

4. Analysis of Question Corpora

4.1 Data Sets

We performed two analyses on four corpora from English Stack Overflow (ESO), Portuguese Stack Overflow (PSO), Russian Stack Overflow (RSO), and Japanese Stack Overflow (JSO). All questions asked in a 24-month period between the beginning of November 2014 and the end of October 2016 were obtained from the Stack Exchange data explorer. The number of questions and tags in each site is shown in Table 4.

The data sets are different in size, with ESO 40 times larger than the other languages combined. In addition, we can see that there are discrepancies between the ratio of tags per question, with ESO providing an average of 3 tags per question compared to 2 tags per question on JSO. Any test on the data sets require normalization in analysis.

Site	Questions	Tags	Tags per Question
English (ESO)	4,608,931	13,875,878	3.01
Portuguese (PSO)	42,164	108,033	2.56
Russian (RSO)	64,125	155,496	2.42
Japanese (JSO)	9,208	18,488	2.01

Table 4. Language Sites and Number of Tags and Questions

4.2 Do the language versions of Stack Overflow ask different questions?

The null hypothesis is that the languages are not significantly different from each other. Specifically, we assume that the most popular tags in each language version are similarly distributed across the various languages. To test this, we took the top 25 tags across each site and compared their normalized frequencies. The top 25 tags in each language version provides coverage of almost every question asked in the 24-month period. All four sites had at least 99% of all their questions containing at least one of the top 25 tags. In total, there are 43 tags that are in at least one of the top 25 tags for all four language sets.

To compare expected frequencies across the data set we used the Log-Likelihood (LL) test, which can compare keywords amongst corpora of different sizes (Rayson & Garside, 2000). LL is a contingency test with a statistic similar to Chi-Squared (Rayson, 2008). LL normalizes by comparing relative expected frequencies based on the entire population of the corpus (Rayson, 2008; Rayson & Garside, 2000), thus it is a useful tool for identifying structural differences between corpora of different sizes.

We conducted both groupwise and pairwise tests with a total of 301 comparisons. Therefore, we consider all tests to have an adjusted alpha $p<0.05$ to be $p<0.00016$ using Bonferroni Correction. For the groupwise test, we consider a LL score (G-test) of greater than 24 to be significant, and for the pairwise, a result greater than 15.5. In addition, we measure Bayes-Factor, which shows strength of difference. A score of greater than 10 indicates strong differences.

Tag	Log Likelihood	Bayes BIC	Tag	Log Likelihood	Bayes BIC
.net	843.4302618*	797.3254906	monaca	5568.539653*	5522.434882
ajax	665.3195587*	619.2147875	mysql	1752.798235*	1706.693464
algorithm	515.2795249*	469.1747538	node.js	640.6765277*	594.5717565
android	2075.276235*	2029.171463	objective-c	867.4350835*	821.3303124
angularjs	1085.705483*	1039.600712	onsen-ui	1431.304119*	1385.199348
arrays	440.0998629*	393.9950917	os-x	603.1148608*	557.0100897
asp.net	490.7934195*	444.6886484	php	8988.499952*	8942.395181
asp.net-mvc	1262.068759*	1215.963988	python	2761.685533*	2715.580762
c	221.6225221*	175.5177509	r	2291.777984*	2245.673213
c#	1473.765391*	1427.66062	regex	125.6873619*	79.58259075
c++	1078.043712*	1031.938941	ruby	1530.97083*	1484.866059
css	667.8303377*	621.7255666	ruby-on-rails	1612.662362*	1566.557591
database	849.4298113*	803.3250401	sql	493.448244*	447.3434729
html	1044.189648*	998.0848772	sql-server	338.1206603*	292.0158891
html5	783.4355541*	737.3307829	string	255.6131244*	209.5083532
ios	3427.559941*	3381.45517	swift	1573.295421*	1527.19065
java	1642.976333*	1596.871562	unity3d	444.7771328*	398.6723616
javascript	700.6201567*	654.5153855	windows	156.6513022*	110.546531
jquery	879.7062011*	833.6014299	wordpress	124.0019744*	77.89720326
json	103.3854668*	57.28069563	wpf	574.6894916*	528.5847205
linux	569.1973065*	523.0925354	xcode	821.1311931*	775.026422

Table 5. Groupwise Log-Liklihood Results *p<0.05

As Table 5 shows, all groupwise comparisons have significant results. This shows that no tag is distributed evenly among the four languages. What this indicates at first is that computer programming is not as limited in its domain as the system designers thought. However, the groupwise test does not conclusively prove that all the sites are different from each other, as one site could prove to be skewing most of the results. To account for this, we conducted six pairwise tests as shown in Table 6. The results clearly show that most pairs are significantly different.

None of the corpora have similar distributions of their most popular tags. We can reject the null hypothesis for R1 and conclude that the different language sites do choose different distributions of topics.

Comparison	Significant	Not Significant
ESO vs PSO	33	9
ESO vs RSO	39	4
ESO vs JSO	36	6
PSO vs RSO	33	9
PSO vs JSO	36	6
RSO vs JSO	33	9

Table 6. Summary of Pairwise Tests

None of the corpora have similar distributions of their most popular tags. We can reject the null hypothesis for R1 and conclude that the different language sites do choose different distributions of topics.

4.3 Do the language versions of Stack Overflow present different tag structures for universally popular tags?

The fact that the distribution of tags is different for each of the languages does not mean that there is a fundamentally different way of asking questions regarding similar tags. To examine this, we chose a universally popular tag “Javascript”, which ranked No. 1 on ESO, No. 2 on PSO, No. 3 on RSO, and No. 1 on JSO. We then sought to understand the related tags that are included when a “Javascript” tagged question is asked.

A potentially useful tool would be to use Normalized Mutual Information (NMI) to compare tag hierarchies (Tibély, Pollner, Vicsek, & Palla, 2013). The problem with NMI, however, is that there needs to be a hierarchy of tags. Popularity of the tag does not indicate the hierarchy of the question. In some cases, such as Figure 1, it is relatively easy to designate the node as being “Git” and the subsequent tags as modules. However, many questions are not the same. For example, a question tagged “iOS” and “Android” might be equally important.

To provide an analysis of shared information around the tag “Javascript”, we chose a modified version of Sankey Diagrams which are based on force-directed graphs. In these graphs, the left-most tags represent the most central tags, while size of a bar indicates the relevant frequency to all the questions that were also tagged with “Javascript”. Graphs were created with a normalized link-strength of 3.

There are some notable differences in the centrality of related tags. JSO, for one, includes a technology (Monaca and Onsen-UI) which fails to even make the other charts. It is central to many of the “Javascript” questions that are asked on the site. RSO is dominated by a centrality of “Jquery” which is a tag which is present in almost all of the “Javascript” questions. PSO has “Ajax” and “AngularJS” as its most central tags, and ESO has “HTML” as its most central tag.

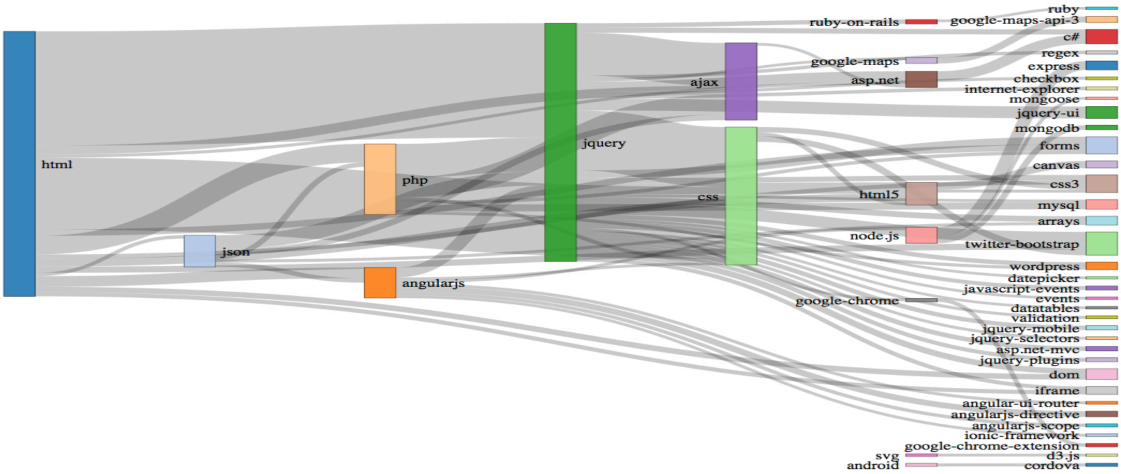


Figure 4. ESO Sankey Diagram

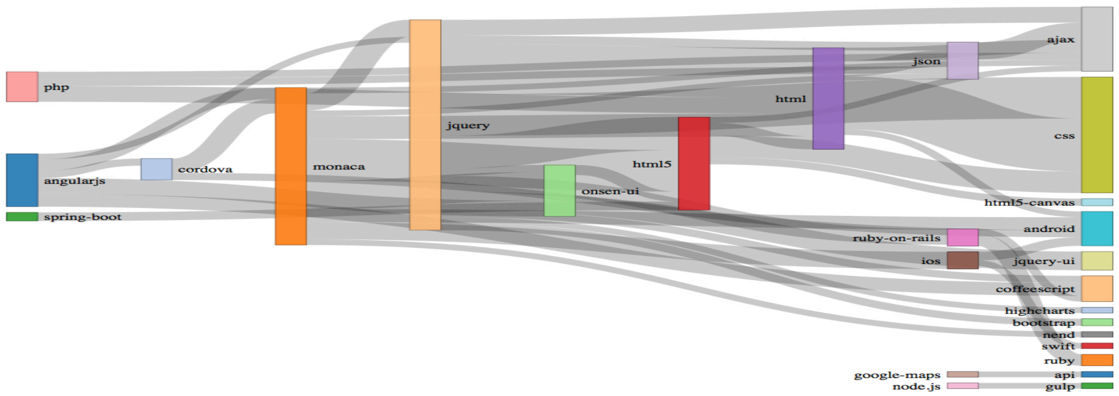


Figure 5. JSO Sankey Diagram

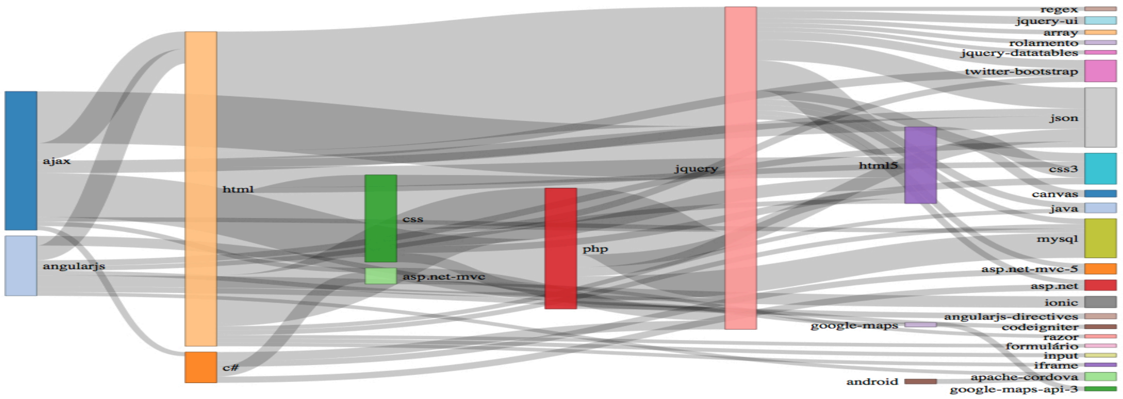


Figure 6. PSO Sankey Diagram

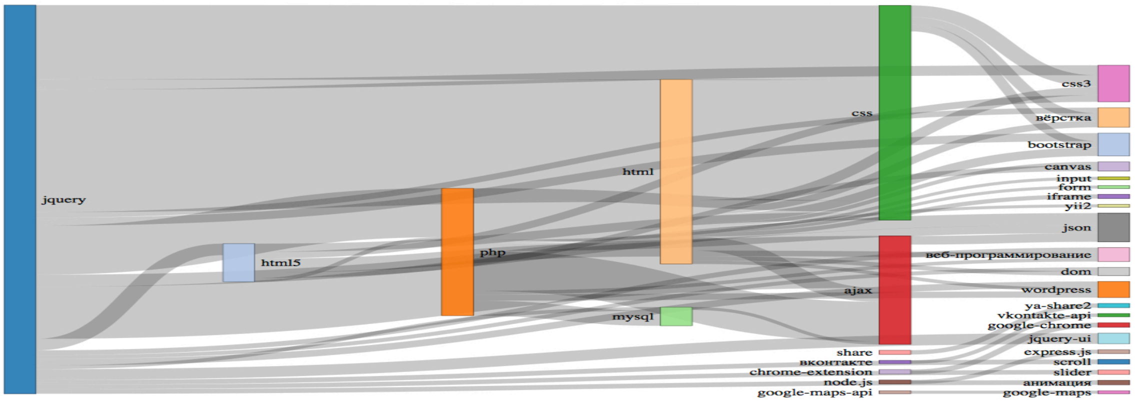


Figure 7. RSO Sankey Diagram

This does not mean that there are not similarities. “Jquery” is a large part of each question set. However, the results still show that the different language versions associate different topics around a popular tag like “Javascript”, and the result is a question set which has a different shape. This indicates that language or locality is important to how and what questions are asked. Even the shared topic areas are visually different.

5. Discussion

The results of the analysis show that the questions asked between the four corpora are significantly different. This adds to the research indicating that culture and language impacts the type of information that is produced and curated by a community (Bao et al., 2012). The results extend the previous work by showing that even in a closed technical domain this impact can be seen.

This result is important for three reasons. First, we can consider that only naïve designers will expect language versioning to result in similar corpora. We should expect that language versioning will result in splintering effect where content is not replicated across sites. This includes situations in which the initial contributors come from the parent site, like Japanese Stack Overflow (JSO).

The second reason these results are important is that they can inform directed translation efforts. Wikipedia is a venue where much effort has gone into creating effective methods for translation of material (Hautasaari, 2013; Ishida, 2011). Lessons from these efforts indicate that it is difficult to locate the gaps in knowledge *a priori*. In a situation like the Stack Overflow system (SO), we might assume that the most basic and common questions are asked first, because they are the most basic questions. However, in JSO we see that country specific questions, like “Monaca” tagged questions show up as a central tag. A possibility is that users are already translating basic questions by themselves instead of investing time into asking basic questions. This might explain the different shapes of the communities.

The differences can be used to inform translation efforts by showing site administrators where gaps in knowledge exist. If the goal of the community is to provide the most access to knowledge, then there is ample opportunity to find places where translation efforts will have the greatest effect.

Most significant, however, is that the results show the importance of not just translating versions of a site into another language. As our results show, it is incorrect to assume that the structure and content of the dominant language site will be replicated (even on a smaller scale). Instead, we see that sites in different languages have different approaches to topics and include technologies that are largely absent from the dominant site.

Without language-specific sites, these discussions might be lost to the all communities. If the sites were mere replicas of each other, then translation-only efforts would make sense. With these results, however, the lack of language specific sites would result in a loss of information. With this in mind, site administrators of dominant language communities might consider including translation versioned sites to their own.

6. Conclusion

This study aimed to determine whether language versions of a technical community result in similar or different question archives. We used a statistical method, Log-Likelihood, and a visual method, Sankey Diagrams, to explore the user-generated question tags as they describe the archives. We found that between four languages the tags were significantly different in distribution. We also found that the related tags around a universally popular tag were different in amount and centrality.

The results from this study are important for informing designers of collaborative information gathering communities interested in language versioning. In addition, we argue that the results are important for both informing translation efforts and justifying the existence of language versioned communities.

Acknowledgements

The authors would like to thank Ari Hautasaari at the Graduate School of Interdisciplinary Information Studies at the University of Tokyo for his encouragement and invaluable feedback on this manuscript. This research was partially supported by a Grant-in-Aid for Scientific Research (A) (17H00759, 2017-2020) from Japan Society for the Promotion of Science (JSPS).

References

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2012). Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 850–858). New York, NY, USA: ACM. https://doi.org/10.1145/2339530.2339665

Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., & Gergle, D. (2012). Omnipedia: Bridging the Wikipedia Language Gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1075–1084). New York, NY, USA: ACM. https://doi.org/10.1145/2207676.2208553

Correa, D., & Sureka, A. (2013). Fit or Unfit: Analysis and Prediction of “Closed Questions” on Stack Overflow. *arXiv:1307.7291 [cs]*. Retrieved from http://arxiv.org/abs/1307.7291

Correa, D., & Sureka, A. (2014). Chaff from the Wheat: Characterization and Modeling of Deleted Questions on Stack Overflow. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 631–642). New York, NY, USA: ACM. https://doi.org/10.1145/2566486.2568036

Glaser, B. G., & Strauss, A. L. (2009). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Transaction Publishers.

Hanlon, J. (2014, February 13). Can’t We All be Reasonable and Speak English? Retrieved November 16, 2017, from https://stackoverflow.blog/2014/02/13/cant-we-all-be-reasonable-and-speak-english/

Hautasaari, A. (2013). “Could Someone Please Translate This?”: Activity Analysis of Wikipedia Article Translation by Non-experts. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 945–954). New York, NY, USA: ACM. https://doi.org/10.1145/2441776.2441883

Hecht, B., & Gergle, D. (2010). The Tower of Babel Meets Web 2.0: User-generated Content and Its Applications in a Multilingual Context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 291–300). New York, NY, USA: ACM. https://doi.org/10.1145/1753326.1753370

Hong, L., Convertino, G., & Chi, E. (2011). Language Matters In Twitter: A Large Scale Study. Presented at the Fifth International AAAI Conference on Weblogs and Social Media. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2856

Ishida, T. (2011). *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*. Springer Science & Business Media.

Li, G., Zhu, H., Lu, T., Ding, X., & Gu, N. (2015). Is It Good to Be Like Wikipedia?: Exploring the Trade-offs of Introducing Collaborative Editing Model to Q&A Sites. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1080–1091). New York, NY, USA: ACM. https://doi.org/10.1145/2675133.2675155

Maciel, J. (2014, December 16). Stack Overflowへようこそ. Retrieved November 16, 2017, from https://stackoverflow.blog/2014/12/16/stack-overflow-in-japanese/

Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., & Hartmann, B. (2011). Design Lessons from the Fastest Q&a Site in the West. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2857–2866). New York, NY, USA: ACM. https://doi.org/10.1145/1978942.1979366

Oliveira, N., Andrade, N., & Reinecke, K. (2016). Participation Differences in Q&A Sites Across Countries: Opportunities for Cultural Adaptation. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (pp. 6:1–6:10). New York, NY, USA: ACM. https://doi.org/10.1145/2971485.2971520

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549. https://doi.org/10.1075/ijcl.13.4.06ray

Rayson, P., & Garside, R. (2000). Comparing Corpora Using Frequency Profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9* (pp. 1–6). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1117729.1117730

Tausczik, Y. R., & Pennebaker, J. W. (2012). Participation in an Online Mathematics Community: Differentiating Motivations to Add. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 207–216). New York, NY, USA: ACM. https://doi.org/10.1145/2145204.2145237

The Ugly American Programmer. (2009) Retrieved November 16, 2017, from https://blog.codinghorror.com/the-ugly-american-programmer/

Tibély, G., Pollner, P., Vicsek, T., & Palla, G. (2013). Extracting Tag Hierarchies. *PLOS ONE*, *8*(12), e84133. https://doi.org/10.1371/journal.pone.0084133

Treude, C., Barzilay, O., & Storey, M.-A. (2011). How Do Programmers Ask and Answer Questions on the Web? (NIER Track). In *Proceedings of the 33rd International Conference on Software Engineering* (pp. 804–807). New York, NY, USA: ACM. https://doi.org/10.1145/1985793.1985907

Treude, C., Prolo, C. A., & Filho, F. F. (2015). Challenges in Analyzing Software Documentation in Portuguese. In *2015 29th Brazilian Symposium on Software Engineering* (pp. 179–184). https://doi.org/10.1109/SBES.2015.27

Vargo, A. W., & Matsubara, S. (2016). Editing Unfit Questions in Q&A. In *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* (pp. 107–112). https://doi.org/10.1109/IIAI-AAI.2016.83

What are tags, and how should I use them? - Help Center - Stack Overflow. (n.d.). Retrieved November 23, 2017, from https://stackoverflow.com/help/tagging

What are tag synonyms and merged tags? How do they work? - Meta Stack Exchange. (n.d.). Retrieved November 23, 2017, from https://meta.stackexchange.com/questions/70710/what-are-tag-synonyms-and-merged-tags-how-do-they-work

Xu, B., Xing, Z., Xia, X., Lo, D., Wang, Q., & Li, S. (2016). Domain-specific Cross-language Relevant Question Retrieval. In *Proceedings of the 13th International Conference on Mining Software Repositories* (pp. 413–424). New York, NY, USA: ACM. https://doi.org/10.1145/2901739.2901746