A comparative study of programming languages used in Google Code Jam

Tiago Martins
up201305044@fc.up.pt

Computer Science Department,
Faculty of Sciences of University of Porto

June 20, 2017

Abstract

This article studies and analyzes the use of programming languages in the context of competitive programming, using Google Code Jam international competition as the case study. First, we'll calculate the most used languages in this contest and then we'll make a comparison with two other programming languages's rankings, to observe how the top programming languages varies between the different ranking contexts.

It will be also analyzed how the use of the top languages have changed over the years of competition, by observing the percentage of usage in each one of those years.

By using the data from two different years, we studied how the use of the top five languages changes throughout the competitions rounds.

We'll show which where the most represented countries that ever participated in this competition and make a small comparison with their population number. Finally, we'll present the most successful countries, based on the times they reached the finals and the number contestants that represented each country in said finals.

1 Introduction

The comparison of programming languages is a subject that has always aroused great interest. Which one is the most efficient? Which one is the easiest to use? Which one is the best overall? These are some questions that have already been studied in papers like [8] and [9].

In this project, programming languages are analyzed in the context of competitive programming. This competition is a sport where participants must write programs capable of solving given problems[2]. Programming competitions date back to the early seventies, to events like the ACM International Collegiate Programming Contest (ACM-ICPC)[1], and the interest in this sport has been growing over the years. Although there are several popular programming competitions, in this study we chose to analyze Google Code Jam (GCJ), since it is a well-know contest and there is a lot of data and statistics available about its tournaments.

GCJ is an international programming competition[5] organized by Google. The first tournament was made in 2003, and has been held every year since then, with the exception of 2007. The tournament itself is divided in seven rounds, where the competitors must solve several algorithmic problems in a limited amount of time, using any programming language. This last feature is another reason why GCJ was chosen.

The main objective of this report is to analyze the programming languages used in the *GCJ* competition. First, we'll do a ranking of which were the most used languages in the whole contest, and, using the top 5 languages, do a comparison with 2 other programming languages rankings.

The data that we'll analyze is available in the **go-hero.net** website[2]. This website provides the tournament information[4] and statistics from 2008 to 2016. However, the data from the year 2008 was excluded from this study, since the structure of the tournament rounds in that particular year was different from the following years. The rounds from the analyzed years, 2009 to 2016, are divided into **Qualification Round**, **Round 1A**, **Round 1B**, **Round 1C**, **Round 2**, **Round 3** and **Final Round**.

In this study we'll also investigate how the usage of the top programming languages has changed in each one of the analyzed years. In an identical way, we will try to find how the usage of these languages varies through the rounds of a specific competition year.

[1] https://icpc.baylor.edu

[2] https://www.go-hero.net/jam

2 Obtaining the data

2.1 Tools used

Several scripts were implemented to do the web scraping from the **go-hero.net** website. Web scraping is a technique used to extract data from websites[3], which can be saved in local files for later analysis.

All the scripts implemented were written in `Python` and are available in a *GitHub* repository[3] created for this study. In order to obtain the html from each page, it is used the *urllib*[4] package, that can retrieve it with the *request* module. After obtaining the html, it is used the *BeautifulSoup*[5] library to navigate through the html tree and get the desired data from it. All the information is stored in several *csv* files, which are then analyzed with the help of the *NumPy*[6] package to load and manipulate the data from them. To plot the graphs, it is used the *Pyplot* module from the *matplotlib*[7] library.

2.2 Data Obtained

The first script implemented had the objective of retrieving the data that is the basis of this study: all the different languages ever used in the competition. The script simply gets all the languages used in each year and groups them together in a list, removing the duplicates. In an identical script, instead of programming languages, it is obtained the list of all the countries that have participated in the competition.

The next data to be retrieved was the table that contained, for each language, the number of contestants that used it in each round, for a specific year. For each year, the correspondent table is stored in a file with the name **langs_year_XY.csv**, where **XY** represents the year. It was also obtained the data from a similar table, where in this case, it represents the number of contestants from each country in each one of the rounds. As for the programming languages, it is created a file for each year, with the name **users_per_year_XY.csv**, where **XY** represents the year.

Finally, it is obtained, for each year, the number of users that submitted solutions using at least 3 different programming languages and also the number of languages that the contestant with more languages submitted used.

[3] https://github.com/Tiaghoul/iic-GoogleJamStudy
[4] https://docs.python.org/3/library/urllib.html
[5] https://www.crummy.com/software/BeautifulSoup/bs4/doc

3 Analyzing the data

3.1 Languages used

Firstly, it was observed that throughout the 8 years of competition that this study focuses on, 166 different languages were used to correctly answer a problem. However, around 40 of those languages were used not more than 2 times and 120 weren't used more than 20 times.

Looking at the number of participants for each year that used more than 3 languages, there is no fixed pattern or evolution over the years. The peak was in the year 2013, where 166 contestants answered the problems using at least 3 different languages. The most different languages used in a year by a single contestant was 23, achieved both in 2015 and 2016.

3.2 Language utilization

To find which languages were used to solve more problems over the years, it was calculated the percentage of usage for each language, by measuring the mean of the usage percentage of said language in each year.

The 10 most used languages in GCJ and the respective percentage are represented in the Table 1. Looking at the percentages of usage, it's clear that `C++` language has a commanding lead over the remaining languages, with a percentage of usage almost 4 times bigger than the 2nd most used language, `Java`.

Table 1: Percentage of the 10 most used languages

Language	Usage (%)
C++	60.901
Java	16.965
Python	10.920
C#	3.243
C	2.597
Ruby	1.056
Haskell	0.753
Pascal	0.587
PHP	0.507
Perl	0.478

`Python` is featured in 3rd place, and is, in addition to `C++` and `Java`, the only language with more than 10% of overall use.

Although we're analyzing only 10 languages out of a group of 166, the percentage of usage of `C++` is so overwhelming that even the 4th language, `C#`, and 5th language, `C`, are representing a mere 3 and 2 percent.

[6] http://www.numpy.org/
[7] https://matplotlib.org/index.html

Table 2: Top 5 languages ranking in different contexts

Language	GCJ Rank (%)	TIOBE Rank (%)	RedMonk Rank
C++	1 (60.901)	3 (5.723)	5
Java	2 (16.965)	1 (14.493)	2
Python	3 (10.920)	4 (4.333)	3
C#	4 (3.243)	5 (3.530)	5
C	5 (2.597)	2 (6.848)	9

The last 5 languages from this rank are `Ruby`, `Haskell`, `Pascal`, `PHP` and `Perl`, respectively. The total percentage of these 5 languages gives a value around the percentage of `C#`.

When analyzing the top 5 most used languages for each one of the years, we observed that they are the same in all the years, in exactly the same order as the top 5 most used languages.

3.3 Comparison with other language rankings

In order to see if there was a difference between the top 5 most used languages in GCJ and other available programming language rankings, it was made a comparison with 2 other rankings, both of them focusing on different language contexts.

The two rankings used in this comparison were:

- *TIOBE Index*: is an indicator of the popularity of a programming language[7], where the ratings are based on the number of times a language is searched in popular search engines, such as *Google*, *Youtube* and *Wikipedia*. The search query used is `+"<language> programming"` and the ranking is updated once a month;

- *RedMonk Programming Language Rankings*: it's an index updated twice a year that extracts language rankings from *GitHub* and *Stack Overflow* and combines them in a list that reflects the correlation between the number of pull requests of a language(*GitHub*) and language discussion(*Stack Overflow*)[6]. Its objective is to extract insights of potential future adoption trends.

Results are displayed in the Table 2. The *TIOBE* ranking contains both the rank and overall percentage for each language, while the *RedMonk* ranking only contains the language's position.

An overview of Table 2 gives us the information that all the languages are ranked differently in each one of the contexts.

Analyzing the *TIOBE* ranking, we can observe that its top 5 languages are the same as GCJ 5 most used languages, although their rank order is different.

The `Java` language, which is the 2nd most used in GCJ, takes the 1st place in the `TIOBE Rank`, with a percentage of popularity 2 times bigger than the 2nd place `C`, that is in 5th place in GCJ ranking. Even though `C++` is the winner in the context of competitive programming, it only comes in 3rd place in the language popularity ranking.

However, when analyzing the *RedMonk* ranking, it's easy to observe that his top 5 languages aren't the same as the top 5 from GCJ. When the context represents the number of pull requests from *GitHub* and the amount of discussion in *Stack Overflow*, the `C++` language drops even further down to the 5th place (tied with `C#`), while both `Java` and `Python` surpass it, in 2nd and 3rd places respectively.

The *RedMonk* top 5 is completed with `Javascript` in 1st place and `PHP` in 4th. These 2 languages are represented in 13th and 9th in the GCJ index.

We can conclude that the dominance that `C++` has in the GCJ ranking is not translated in the other 2 rankings, and that `Java` is the language that stands out more, both in the popularity ranking and in the pull requests/discussion ranking.

3.4 Language evolution over years

In order to find how the usage of each language varied over the years, several graphs were plotted to illustrate the percentage of usage of a group of languages in each one of the 8 years of competition.

When plotting the graph for the top 5 languages, a decision was made to leave *C++* out of the graph, since his high values made the evolution of the other languages barely perceptible. His evolution was nearly linear over the years, having its peak in 2015, with around 68% of usage.

However, in addition to `C`, `C#`, `Java` and `Python`, it was also plotted the total percentage of the remaining languages used in each year, represented by `Other`.

Analyzing the graph in Figure 1, it's possible to observe that the usage of `C` and `C#` has been decreasing over the years, and `C` even dropped to only 1% in 2016. In the case of `Java`, although it had several increases over the years, the overall balance is negative, since it had 20% of use in 2009 and dropped to 15% in 2016.

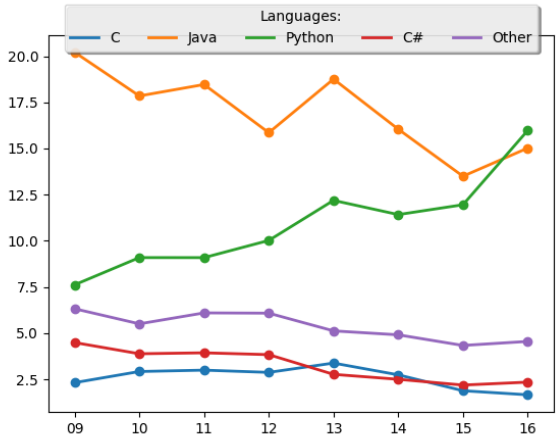

Figure 1: Evolution of top 5 languages over the years (without `C++`)

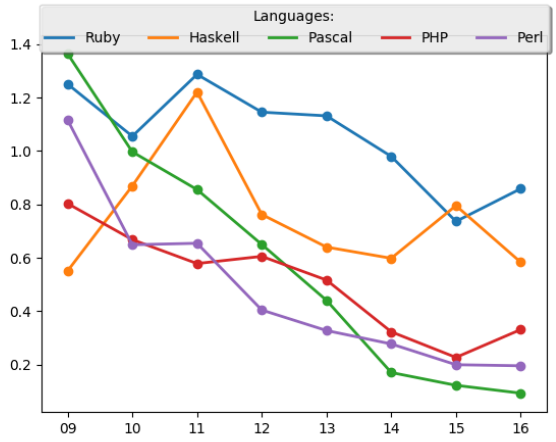

Figure 2: Evolution of 6th to 10th most used languages over the years

Out of the 4 languages, `Python` is the one that had the greatest growth, having more and more contestants using it over the years (with only a small exception in 2014). It passed from 7.5% in 2009 to around 16% in 2016, being that the only year that it surpassed `Java`.

Finally, looking at the values for `Other`, we can conclude that the total use of the remaining languages was almost similar every year, with only a small decrease in the more recent years.

The graph displayed in Figure 2 represents the evolution of usage of the 6th to 10th most used languages in GCJ over the years.

Although the values in this graph are representing some of the most used languages from GCJ, the biggest percentage of usage displayed is only 1.4%, in the year of 2009, by `Pascal`.

The first impression in this graph is that, with only the exception of `Haskell`, the percentage of use of all the other 4 languages has been decreasing every year. The most affected language was `Pascal`, which was the most used in 2009, but decreased each year, until it was the least used in 2016, with less than 0.1% of usage. The `Perl` language had a similar evolution to the one of `Pascal`.

The usage of `PHP` also decreased over time, but not as abruptly as the previous two languages.

Finally, `Haskell` was the only language that had a similar percentage of usage in 2009 and 2016, having its peak in 2012, with around 1.2%.

The decreasing values observed in Figure 2 are the result of the recent increase of `C++` and `Python` usage, as previously stated.

3.5 Language evolution over rounds

We then tried to find how the usage of the top 5 languages changed throughout the 7 rounds of 2 particular years of GCJ competition.

The graphs in Figures 3 and 4, representing years 2013 and 2016 respectively, are, again, excluding the `C++` language, since, as expected, his high values make the other language's evolution less discernible. These graphs are also representing the percentage of usage for the remaining languages used in 2013 and 2016

However, analyzing the excluded values of `C++` for both years, is possible to observe that the use of this language increases as each round passes, but it's in the round **R2** that it has the biggest growth, passing from around 50% in the first 4 rounds to more than 70% in the 3 final rounds, thus demonstrating the dominance previously mentioned.

The `C++` percentage increase in the final rounds is an expected result, since it is an advanced phase of the competition and the more experienced contestants tend to use their *go-to* language.

Looking at the Figures 3 and 4, we can see that `C`, `C#` and also the remaining languages, represented by `Other`, maintain almost the same values from **Qualification Round** to **R2**, where the values start decreasing, until they reach 0% in the **Final Round**. The only exception is `C` in 2013, which is used in the last round, unlike in 2016.

Both in 2013 and 2016, `Python` also started decreasing after round **1C**, although it had a small increase in the final of 2016.

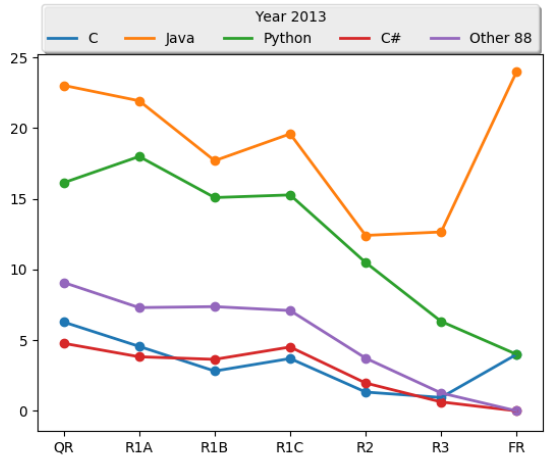

Figure 3: Evolution of top 5 languages in 2013 (without `C++`)

The evolution of `Java` in 2013 was a little bit unusual, since it had several increases and decreases over the rounds, and surprisingly, its peak was in the **Final Round**.

Focusing only in `Python` and `Java` in both graphs, we can observe that, in 2013, `Java` had a higher percentage of usage than `Python`, and that in 2016 the opposite happened. This confirms the previous statement about Figure 1, where it was said that `Python` only surpassed `Java` in 2016.

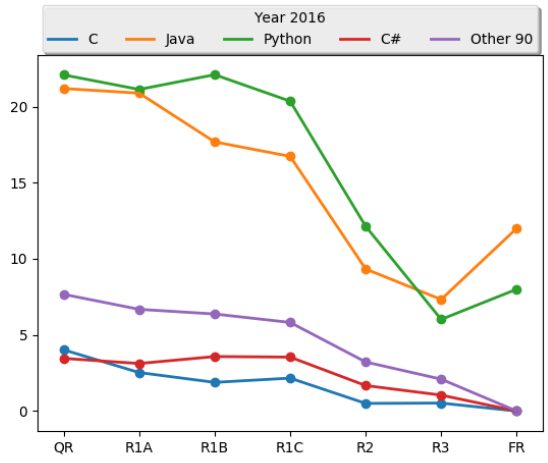

Figure 4: Evolution of top 5 languages in 2016 (without `C++`)

3.6 Represented countries

In this section we'll make a small study about the countries that participated in GCJ over the years.

Analyzing the list that contains all the countries that have ever participated in the competition, we can observe that it has 217 entries, because it is divided not only in countries, but also in insular areas, archipelagos, overseas collectivities, amongst others.

Table 3: Overall percentage of top 5 countries

Country	GCJ (%)	World (%)
China	15.940	18.5
Russia	11.360	1.9
India	9.059	17.9
United States	8.714	4.3
Japan	6.499	1.7

The Table 3 contains the 5 most represented countries in GCJ, and it also contains the percentage of each country in the world population[1].

With no surprise, China is the most represented country in GCJ, with a percentage almost similar to its world percentage. Russia is a case in point, since it represents 11% of GCJ contestants while having a much lower 1.9% of world population. Meanwhile, India, despite having a population 9 times bigger than Russia, it's represented by less contestants, with 9% in GCJ.

Closing the top 5 are the United States and Japan, respectively, both having a percentage in GCJ a couple times bigger than their respective world percentage.

3.7 Most successful countries

There are 28 different countries that ever reached the **Final Round**. However, we'll be only focusing in the 7 countries that were represented more times, since they're the only ones that went to the finals in at least 5 occasions.

Table 4: Top 7 most successful countries

Country	Times in final	Contestants
Belarus	6	10
China	7	22
Japan	8	20
Poland	8	18
Russia	8	52
Ukraine	6	13
United States	6	9

These countries are represented in Table 4, which also contains the number of contestants, of each country that participated in the **Final Round**.

It's possible to observe that Japan, Poland and Russia are the 3 countries that were represented 8 times in the **Final Round**, which means, in all the years that this study focuses on.

Although Russia is the 2nd most represented country in GCJ, only surpassed by China, it took more than the double of contestants to the final that China did, being thus, the most represented country in the finals of GCJ, with 52 contestants.

Finally, Belarus, Ukraine and the United States reached the final 6 times and were represented by a similar number of contestants. However, it's very unexpected to see Belarus in this top 7 group, since it is a much less represented country in GCJ than the other 6 countries.

It is also important to mention the absence of India in this group, since it is the 3rd most represented country in the competition.

4 Conclusion

In this study, we started by stating that 166 different programming languages were used to solve a problem in GCJ, although around 70% of those languages weren't used more than 20 times.

Then, we concluded that, in the context of competitive programming, `C++` is the language to beat, having a giant lead over the 2nd and 3rd most used languages, `Java` and `Python`, respectively. `C#` in 4th place, and `C` in 5th, have a similar percentage of usage, around 3%. The top 10 is closed by `Ruby`, `Haskell`, `Pascal`, `PHP` and `Perl`, being the total percentage of these 5 languages only 3% of the overall use.

When comparing the top 5 ranking that we obtained, with 2 other programming languages indexes, we saw that `C++` is no longer the leading language, dropping several spots in both rankings. `Java`, `Python` and `C#` maintained almost the same place in all the rankings, and Java even reached the 1st place in the *TIOBE* ranking, which represents the popularity of programming languages based on the number of searches.

It was also concluded that the use of `Python` has been growing over the years, and that it even exceeded `Java` in 2016. The widely used `C++` has had a small growth in recent years. Meanwhile, the values of the already little used `C`, `C#`, `Pascal`, `Perl`, `PHP` and `Ruby`, have been decreasing to even lower values.

Analyzing the evolution over the rounds we saw that the usage of `C++` is even bigger in the last 3 rounds of a competition and that all the other languages maintain almost the same values throughout the initial rounds, before starting to decrease (some of them to 0%) when the round **R2** is reached.

Finally, it was observed that a highly represented country doesn't necessarily correspond to a successful one. Even though Russia is a much less populated country than India, it has a higher representation in GCJ and it's even the country that took more contestants to the **Final Round**. We also concluded that Belarus was able to reach the final for 6 times, as many as United States and Ukraine, despite having many fewer contestants in the GCJ than those 2 countries.

It's also worth mentioning that, with the exception of the `R` language, we could not find any other language in which the percentage of usage grew year after year, consistently and uninterruptedly.

References

[1] Countries in the world by population (2017), 2017, http://www.worldometers.info/world-population/population-by-country/

[2] Competitive Programming, https://en.wikipedia.org/wiki/Competitive_programming

[3] What is Web Scraping?, https://www.webharvy.com/articles/what-is-web-scraping.html

[4] Code Jam Language Stats, https://www.go-hero.net/jam/

[5] What is Code Jam?, https://code.google.com/codejam/about

[6] O'Grady S., (2017). The RedMonk Programming Language Rankings: January 2017, http://redmonk.com/sogrady/2017/03/17/language-rankings-1-17

[7] TIOBE Index for June 2017, 2017, https://www.tiobe.com/tiobe-index/

[8] Bissyandé, T., et al., (2013). Popularity, Interoperability, and Impact of Programming Languages in 100,000 Open Source Projects. In Proceedings of the 37th Annual Computer Software and Applications Conference (COMPSAC '13), IEEE Computer Society Washington, DC, USA, 2013, 303-312.

[9] Nanz, S., Furia, C., (2015). A Comparative Study of Programming Languages in Rosetta Code,. In Proceedings of the 37th International Conference on Software Engineering - Volume 1 (ICSE '15), IEEE Press Piscataway, NJ, USA 2015, 778-788.