The linguistic dominance of English in computer science: A case study of Francophone programmers

Alex Graves
8 May 2019

FREN229: French in the world
Sophie Degât-Willis

I. Introduction

In order to be a programmer, no matter who you are, you must learn English. The field of computer science is a worldwide one, with programmers who come from a myriad of countries and who speak a wide variety of languages. However, a single language dominates programming languages, conceptual education, and research in the domain: English.

In fact, every frequently-used programming language — including Java, C++, Python, and JavaScript — is written in English. Even the programming language Caml, which was developed in France at the National Institute for Research in Computer Science and Automation (INRIA), uses English keywords because it was based on another language that had previously used English. Other examples are the popular language Python, which was created in the Netherlands, and Ruby, which was established in Japan. Both were developed in English. For non-Anglophones, this phenomenon creates an environment of constant contact between English and their mother tongue. This interaction, like every instance of languages in contact, results in code-switching, language transfer, and loan words. Many technical words have been translated into French, like *base de données* for *database* and *intelligence artificielle* for *artificial intelligence*, or Gallicized, like *déboguer* for *debug*. Certain others have been accepted as loan words, notably the word Internet.

This situation raises certain principal questions about computer science and languages: Why is English so dominant in the field of computer science? How does the power of English in the world of computer science influence French programmers? In order to explore these subjects, I will present a literature review of the history of English in computer science as well as the manifestations of the language's influence on the technical French lexicon.

In addition, in order to discuss the human experiences of Francophone programmers, I will recount an interview with Eric Fouh, a computer science professor at the University of Pennsylvania in the United States, who was born in France and studied there. This interview will present a comparison between computer science education methods in France and in the United States, as well as a discussion of the differences that arise for programmers in one country versus the other.

Finally, I will demonstrate the process and results of a data analysis of the website GitHub on the frequency of French and English in programs written by Francophones. Specifically, this study will examine the language differences in the comments that the authors

wrote in those programs. The differing percentages of French and English in the comments will effectively present an example of the dominance of English in the everyday work of computer programmers.

II. The history of computer science

The field of computer science is dominated by the English language. Today, it is clear that American businesses, such as Apple, Google, and Microsoft, are at the forefront of the field. This predominance has persisted through the entire history of the discipline. Both Charles Babbage and Ada Lovelace, who are considered the first pioneers of computer science, were born, studied, and worked in England in the middle of the 19th century. Another computer scientist who played an important role is Alan Turing, who was also British. Because of his work on algorithms and his theory of computation, he is often considered the father of computer science.[1]

In the 1960s and 1970s, the growth of Silicon Valley in the San Francisco Bay Area of California helped to establish the preeminence of the United States in the domain of computer science, and as a result, consolidated the strength of the English language. The new technologies developed in the region triggered the microcomputer revolution, which marked the emergence of the first personal computers. During the 1980s, the personal computer industry grew enormously, and models like the ZX Spectrum and the Commodore 64 became very popular. The former was produced by Sinclair Research, a British company, and the latter was created by Commodore International, an American business. These personal computers supported the programming language BASIC.

BASIC was, for many people, their first exposure to computer science. In this way, personal computers and this programming language served as the developmental base of the programming industry and of the popularization of computer science as a field and a career.[2] However, because of the origins of the largest businesses, the keywords of BASIC in addition to many learning materials were in English.

---

[1] S. Barry Cooper and Jan van Leeuwen, *Alan Turing: His Work and Impact* (Waltham: Elsevier Science, 2013), 481.
[2] Harry McCracken, "Fifty Years of BASIC, the Language That Made Computers Personal," *Time*, April 29, 2014.

This prevalence of English in programming languages persisted over time, and the ones most frequently used today continue to use English keywords. All of the key factors that constitute the history of the computer science discipline have contributed to a dominance of the English language.

III. The dominance of English in the domain of computer science

As the history shows, the English language has always had a significant place in the expansion of the field of computer science. This predominance manifests itself in the way that English serves as the common language of this discipline. I will examine this phenomenon of the lingua franca in addition to the position of English in computer science education and the effects on the French lexicon.

English as the lingua franca of computer science

English has long been the common language of science. Starting at the end of the 19th century, the United States established its place as an economically and politically dominant world power, a process that was accelerated by the two World Wars. As a result, English took a leading position in international communication and, similarly, in the domain of science.[3] This foundation facilitated the establishment of English as the lingua franca of computer science during the emergence of the field in the 1940s and 1950s.

From those years to today, the domain of computer science has evolved rapidly and thus its language has changed concurrently. Furthermore, there are many sub-disciplines (such as algorithms, artificial intelligence, and computer graphics) that grow separately with their own lexicons developed around the world. According to Jean-Bernard Koechlin, "the computer is a place of contact between the specialty's dominant languages, American English and French, which hold this role principally for economic reasons."[4] These factors create a technical language of computer science that is constantly evolving and in contact with different languages, notably English and French.

[3] Rainer Enrique Hamel, "The dominance of English in the international scientific periodical literature and the future of language use in science," *AILA Review* 20 (2007): 53-71, 56.
[4] J. B. Kœchlin, "Le français, l'anglais, l'ordinateur… et les gens," *Le français en contact avec l'anglais* 21 (1998): 159-171, 159.

The field of computer science is clearly very global, always a collaboration between researchers and programmers from a wide array of countries across the globe. In a field so cooperative and international, the existence of a common language is incredibly useful and English serves as this lingua franca. This phenomenon also exists in the natural sciences, where the English language has continued to grow its power. In 1980, 74.6% of natural science publicans were in English, and in 1996, this figure became 90.7%. Among the different disciplines for the latter year, mathematics and physics had the highest percentages (94.3% and 94.8%, respectively). These "pure" sciences are the most similar to computer science, which closely resembles math. French was used in just 2.3% of math publications in 1996. On the whole, by the end of the 20th century, English was used in at least three quarters of publications in any field.[5]

Moreover, English is specifically a significant part of the act of programming. Although there were attempts to create programming languages with keywords that were not in English, they did not garner widespread usage. In standard programming languages, style rules require programmers to write their comments and variable names, functions, and classes in English.[6] In a survey performed by Philip J. Guo about the barriers to learning computer science, 96% of respondents said that they read instructional materials in English. Of the people in the study who responded that they read them in more than one language, only one percent used both French and English.[7]

English is the sole dominant language in the world, and in global history there has never been a language so dominant.[8] It is clear that this dominance also exists in computer science, where English is clearly the common language.

The role of English in education

Knowledge of English is a huge advantage in the world of computer science. In a survey performed by the Association of French-Speaking Computer Scientists (AILF) in 1984, 53% of

[5] Hamel (2007), 57-60.
[6] Philip J. Guo, "Non-Native English Speakers Learning Computer Programming: Barriers, Desires, and Design Opportunities," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), 2.
[7] Guo (2018), 4.
[8] Hamel (2007), 54.

the 180 respondents had a "good or very good" knowledge of English. Furthermore, "in order to succeed in computer science," a majority "search to learn [English], the ideal being to do their studies in the US." However, 79% of the respondents characterized the linguistic influence of the United States as a "difficulty" or "barrier" and 90% expressed a desire for Gallicization in order to "better communicate" and "understand computer scientists."[9]

In the interview, Professor Eric Fouh said that in France, all of his computer sciences classes were taught in French. Despite this fact, however, English played a necessary role in his education. For computer science, Fouh said, "we had to take English classes." In the classroom, English also served an important purpose. "If the word is in English, we use the word in English," he explained. For example, in programming languages, "we would say *pour i partant de zéro à dix*" which meant *for i starting at zero to ten*. He added, "we write in English, but we explain it in French."

This teaching that necessitates bilingualism can have negative effects. The study about the barriers that non-Anglophones experience when learning computer science and programming found that the biggest problems were instructional materials that are only in English, technical communication in English, and difficulties in learning programming and English at the same time.[10] For certain respondents, the process of learning programming motivated them to improve their English.[11]

Furthermore, a study by Sayamindu Dasgupta and Benjamin Mako Hill found that novice users who learned programming in an environment that used the main language of their country demonstrated new concepts faster than those who used an English system. These conclusions align with a significant amount of research which implies that education in one's mother tongue has positive impacts on learning.[12]

Obviously, the predominance of English in the field of computer science plays a significant role in how the subject is taught. For certain beginning programmers, the necessity to learn English can create problems, whereas others find that its prevalence motivates them to further develop their English skills. Moreover, the presence and importance of the English

[9] Kœchlin (1988), 168.
[10] Guo (2018), 5.
[11] Guo (2018), 8.
[12] Sayamindu Dasgupta and Benjamin Mako Hill, "Learning to Code in Localized Programming Languages," *Proceedings of the Fourth ACM Conference on Learning@Scale* (2017): 33-39, 33.

language in computer science education further solidify its place as the lingua franca of the discipline.

The effects on the lexicon of the French language

The dominance of English and its constant presence in the domain of computer science create a situation where languages are always in contact. According to Jean-Bernard Kœchlin, "American English influences the orthography, lexicon, syntax, and style of the technical language of computer science, in a process similar to that which one can observe for business, advertising, music, and other technical languages."[13] By extension, English influences the orthography, lexicon, syntax, and style of the French language in general. This sentiment is reflected by Beatrice Bagola, who wrote: "The permanent contact with English, or more accurately American English, principal language of this new technology [Internet], influences the French language."[14]

In October 1986 in Paris, the Franterm symposium found that of 300 French technical terms, 20% were loan words, 60% were calques, and 20% were original French terms.[15] Of course, there are many new terms today, given all of the technical innovation during the past thirty years. An example that has existed for a long time is *langage orienté objet* (which designates a programming language that uses an object structure) which comes from the English term *object oriented language*. In certain cases, terms created by programmers or users of computers stay in standard French, despite efforts to change or Gallicize the words. For example, the initialism *PC*, which comes from the initials of *personal computer*, is also used in French. The attempt to replace the term with *OP* for *ordinateur personnel* was unsuccessful, and *PC* persists in the French lexicon.[16]

The development of technical terms is well illustrated by the case of those of the Internet. There are often several terms to describe the same thing, like for the English word *chat*: the official term in French is *causette*; in Quebec, it is *clavardage* or *bavardage*; and the word most

[13] Kœchlin (1988), 159.
[14] Beatrice Bagola, "L'américanisation de la langue française sur Internet ? Quelques aspects de la terminologie officielle et de l'usage des internautes," *Globe* 7, no. 2 (2004): 101-124, 111.
[15] Kœchlin (1988), 165.
[16] Kœchlin (1988), 164.

used by Francophones on the Internet is *chat* or *tchat*.[17] This phenomenon is clearly a result of languages in contact, but it is not solely because of this situation: It is also influenced by governmental attempts at language policy. According to Bagola, "One notices that the [French] lexicon used on the Internet is characterized by a competition not only between French and English, but also among the interior of France or of Francophone countries. There is no doubt that the language is enriched by legislative measures, but it is also done by the particular usage of its speakers."[18]

Similarly, this contact between languages exists in the lives of programmers. According to Élisabeth Eek in a study of different publications, "in effect, these demonstrate a remarkable linguistic wealth [in the French language] despite a marked influence from the English language on certain concepts, those, naturally, which were discovered in the United States."[19] She also discussed terms used for databases, like *CREATE INDEX* or *EXPLAIN*: "The fate of these terms in regards to their integration into the French lexicon will depend on the level of linguistic culture of the users and of the health of computer science research in France, capable in the long term of engendering more words of French appearance."[20] Obviously, the lexicon of the French language is influenced by new technologies, and when those technologies are developed by Anglophones, the relevant terms are borrowed from English. Eek's argument implies that the dominance of a certain language's term is determined in large part by the technology's country of origin. This phenomenon can also be seen in calques used for technologies that were developed in the United States, like *sécurité de la couche de transport* (for *transport layer security*, a specification used for securing communication and exchanges on the Internet) or *autorité de certification* (for *certificate authority*, an entity that uses digital certificates to ensure security on the Internet).

For some people, this interaction between the two languages is seen as a detriment. In Bagola's study, she cited a publication from the Treasury Board Secretariat of Quebec that said: "The usage of English in computer science technologies impoverishes the French language and culture. Code is mainly conceived, developed, and marketed in the English language and, by that

[17] Bagola (2004), 121.
[18] Bagola (2004), 123.
[19] Élisabeth Eek, "La langue française de l'informatique envisage depuis une perspective américaine." *Meta* 43, no. 3 (1998): 455-462, 456.
[20] Eek (1998), 457.

process, English has become the language of data and text processing."[21] Eek has a more positive point: "We don't see, as a result, the danger of a pseudo 'death' of the French language taking shape on the horizon, by the way of an 'invasion' of American English."[22] The two authors noted that the government plays an important role in the influence on the French language through language policy. Bagola cited again the publication that declared that "the government must play a decisive role regarding the promotion and utilization of French in computing technologies. In effect, in adopting a policy in the domain, the government thereby recognizes all the economic, social, and cultural significance of French through computing technologies."[23] According to Eek, "It is naturally the responsibility of terminology commissions charged with converting American English terms into Gallicisms to study the semantic question of the concepts behind the imported words, instead of inventing lexical items of a nature that is incomprehensible to a French speaker."[24]

Although the English language maintains a strong influence on French in the domain of computer science, the latter continues to change, like all languages do when introduced to new situations of linguistic contact. The evolution it is undergoing is completely natural, and the French language is becoming richer as a result. According to Eek, "for as long as the French lexicon will be capable of producing a response of whatever form (neologism, calque, derivative, etc.) to the linguistic challenge of American English and that this form will don the aspect of a Gallicism such as *puce* [for *chip*] or *bogue* (for *bug*), the French language will be far from withering."[25]

IV. An analysis of French on the database GitHub

In order to more deeply explore French and computer science, I performed an analysis of comments on GitHub, which is a website to which millions of computer programmers upload their code. The site hosts a diverse array of programming languages and programmers of different origins. I will explain the questions that I studied, my sources and methods, and the results of the data analysis.

[21] Bagola (2004), 119.
[22] Eek (1998), 458.
[23] Bagola (2004), 119.
[24] Eek (1998), 458.
[25] Eek (1998), 458.

Questions to examine

It is obvious that English holds an authoritative position in the field of computer science, and I wanted to examine how this dominance manifests itself in the work of Francophone computer scientists. Programming languages have certain keywords that are always in English, and as a result lines of actual code in computer programs are unlikely to vary much based on the languages that their authors speak. Despite that, programmers often write comments to explain their code, and they can be written in any language. In an interview with Eric Fouh, he said that "in France, I would probably write comments in French." However, in the situation where a programmer wishes for a program to be used internationally, Fouh said that he would write "comments in English so that everyone can read." Thus, for my analysis, I posed the question: What percentage of comments from Francophone programmers are written in French? What percentage in English?

Languages differ considerably among the Francophone regions. There are certain countries where French is the only official language and it is spoken by almost everyone, like in France, but there are also regions where French exists alongside other languages, such as in Quebec with English. Therefore, I asked: How does the usage of languages vary between the different Francophone regions?

My sources and analytical methods

I used the website GitHub as a database, since there are many code files authored by programmers from around the world. In order to find users' locations, I started with a dataset called GHTorrent. From this dataset, I found the percentage of programmers who specified their location:

```
SELECT SUM(country_code!='\\N')/COUNT(*) FROM users;
```

I chose to examine five Francophone regions: France, Switzerland, Belgium, Quebec, and Senegal. These regions represent a diverse selection of linguistic situations in the French-speaking world. For each, I performed the following queries:

```
SELECT login FROM users WHERE country_code='fr';
SELECT SUM(country_code='fr')/SUM(country_code!='\\N') FROM users;
SELECT SUM(country_code='fr') FROM users;
```

The first found the username for each programmer in the specified country, given by a two-letter country code. I used these results later in order to connect the countries and the

comments from the code files. The two other queries found the percentage and total number of users from each country, respectively.

Next, I needed to connect the users with the code files to which they contributed. For this step, I used the service Google BigQuery, which allowed me to quickly analyze large datasets. First, I ran a query on a dataset called GitHub Archive, which allowed me to specify the dates. I used the following query in order to find the repositories (the locations where programmers save their code) to which the Francophone programmers I had found in the previous step had contributed code in April 2019. The word for contributing code to a repository is "push" (used in *PushEvent*):

```
SELECT repo.name, country_code FROM [githubarchive:month.201904] a
JOIN ( SELECT login, country_code FROM [ghtorrent.users] ) b
ON a.actor.login=b.login
WHERE type='PushEvent' GROUP BY repo.name, country_code;
```

After having found the repositories to which programmers from the given country had contributed, I needed to retrieve the files from the repositories and their contents. For each step, I used another dataset on BigQuery called GitHub Repos. I connected the repository names from the GitHub Archive dataset with those of this dataset, and then I selected the files and their contents that were in those repositories. In order to limit my dataset, I chose only the files that used the programming languages JavaScript, Python, or Java. These languages are all extremely popular and are used in a variety of contexts, like software engineering, web development, and data science.

```
SELECT b.path AS file, c.content AS content, a.country_code AS
       country_code FROM `ghtorrent.repos_countries` a
JOIN ( SELECT id, repo_name, path FROM `github_repos.files` WHERE path
       LIKE '%.js' OR path LIKE '%.py' OR path LIKE '%.java' ) b
ON a.repo_name=b.repo_name
JOIN ( SELECT id, content FROM `github_repos.contents` ) c
ON b.id=c.id;
```

After running all of the above queries, I had a dataset that contained the file name (and thus the programming language in which it was written), the text within the file, and the country of the programmer who contributed to the repository to which the file belonged. The dataset on BigQuery was immense, and processing the 2.5 terabytes of data took 52.8 seconds to run, which is an extremely long time.

The final step of my data collection was to extract the comments from the code files. I wrote a program (in the language Python) that read the contents of my dataset. Comments in Python are designated by a line that starts with #, and in JavaScript or Java, they start with //, /*, or *. For each file, my program checked each line for the appropriate comment symbols, and if they were present, it added the comment to a final separate dataset. This dataset contained the programming language used, the country the author listed, and the actual content of the comment.

Finally, I needed to analyze the languages of the comments. I again used the language Python and a module called "langdetect" in order to evaluate the language of each comment. I chose to break down the results by the countries of the programmers who had written the comments and the programming languages from which they came.

Discussion of the results of the analysis

First, I will present the basic statistics of my dataset. In order to better understand the results of the data analysis, it is important to recognize any potential problems with the initial data.

In the GHTorrent dataset, I found that only 7.58% of users actually specified their location. Although this percentage is rather low, it corresponds to 2,405,811 unique programmers on GitHub.

Of these users, 2.99% had their location set as France. This percentage corresponds to 71,852 people. For Switzerland, the percentage is 0.77% (18,553 people). Belgium is 0.61% (14,630 people), Quebec is 0.02% (494 people), and Senegal is 0.01% (354 people). A complication with Quebec is that it is a province, and therefore the users must specify their state in addition to their country. As a result, the percentage of Quebecois programmers is probably lower than in reality.

After my connections between the other datasets, I had 11,982,844 comments in total. For the countries: France has 8,571,475 of them, Switzerland has 2,196,470, Belgium has 1,036,290, Quebec has 174,151, and Senegal has 4,458. For the programming languages, there are 6,562,640 comments in Java, 3,728,116 in JavaScript, and 1,692,088 in Python. Unfortunately, I could not limit the comments only to those that were written by Francophones since in order to connect the datasets I had to use repositories, to which other programmers (who

would not necessarily be Francophones) could have contributed. Nevertheless, I thought that the comments I had gathered constituted a representative sample of those of Francophone computer scientists.

To start, I broke down the results by the programming languages of the programs in which the comments were found. The following table shows the top five languages that appeared in the comments:

Java	JavaScript	Python
English (80.26%)	English (73.88%)	English (78.90%)
None (2.52%)	French (3.43%)	French (3.37%)
German (2.08%)	Catalan (2.61%)	Catalan (2.58%)
Romanian (1.90%)	Italian (2.23%)	Romanian (1.65%)
Italian (1.69%)	Romanian (2.19%)	Italian (1.58%)

It is immediately obvious that English dominates the results, but also that there are significant differences in the percentages between the programming languages. In programs written in Java, there are often automatically generated comments that explain the program's functions, which could explain the high percentage of English in the Java comments. Similarly, for Python, many programmers place the program's license at the beginning of the file. The licenses are normally copied from a general format that is in English, which could increase the percentage.

For JavaScript and Python, French is the second percentage, and the languages that follow it are ones with many similarities to French. Rather than there being a (relatively) large percentage of Catalan programmers, it is more likely that the module incorrectly classified the language.[26] As a result, it is possible that the percentage of French comments is in fact higher than the results indicate.

For Java, the second percentage is "none." In the case where the module could not detect the language (for example, if the comment was a URL), I marked the language as "none." The reason that this percentage is so high is possibly because Java's automatically-generated

[26] The module's project description says the following: "Language detection algorithm is non-deterministic, which means that if you try to run it on a text which is either too short or too ambiguous, you might get different results every time you run it."

comments often include URLs for the program's documentation websites. German is the third percentage, which is also probably explained by incorrect results from the module because of generated comments. The lack of French in the first five languages surprised me, though I found that French was the sixth most frequently-appearing language, composing 1.65% of the comments.

Next, I compared the languages of the comments by region. The following table shows the top five languages that appeared in the comments:

France	**Switzerland**	**Belgium**	**Quebec**	**Senegal**
English (78.31%)	English (78.12%)	English (77.59%)	English (69.30%)	English (79.53%)
French (2.49%)	Catalan (2.73%)	None (2.59%)	German (9.24%)	French (4.62%)
Romanian (1.99%)	French (2.31%)	French (2.36%)	French (2.66%)	Danish (2.24%)
None (1.83%)	Italian (2.01%)	Catalan (1.99%)	None (2.24%)	Catalan (2.02%)
Italian (1.81%)	Romanian (1.86%)	Italian (1.94%)	Catalan (2.06%)	Italian (1.91%)

It is again obvious that English constitutes the great majority of comments. The presences of Catalan and "none" as the second percentage for Switzerland and Belgium are probably explained by the above reasons. However, I found the case of Quebec fascinating: English constituted just 69.3% of the comments and the second highest percentage was German, with 9.24% of comments. I noted that German was also higher than I had expected for Java, and therefore I further examined the distribution of comments. Of the data from Quebec, 167,976 comments came from programs that were written in Java, while there were just 4,170 for Python and 2,005 for JavaScript. Obviously, the problems created by Java had an enormous effect on the results from Quebec.

In order to evaluate the results without the influence of Java, I redid the analysis, excluding that programming language, for Quebec and for all the regions:

Quebec (without Java)	**All regions (without Java)**
English (76.31%)	English (75.45%)
French (4.36%)	French (3.42%)
Catalan (4.18%)	Catalan (2.60%)
Romanian (2.35%)	Italian (2.03%)
Italian (1.64%)	Romanian (2.02%)

With these adjustments, it is readily apparent that the French language constitutes 3.42% of Francophone programmers' comments in JavaScript and Python. It seems that the actual percentage should be at least a little higher because the frequencies of Catalan, Italian, and Romanian are partially exaggerated as a consequence of the module's errors of language classification. However, despite all of these facts, English is clearly the sole dominant language of the comments written by Francophone programmers.

V. Conclusion

The lexicon of computer science will continue to be dominated by English, and as a result the French language in the domain of computer science will always be evolving. The English language, which attained this dominance on account of the role of English and the United States in the history of the field, serves as the common language of computer science. This position is reinforced by the education of the discipline, which for the most part necessitates knowledge of English. Obviously, these situations create constant contact between English and French. Thus, the French computer science lexicon contains many calques as a consequence of the speed of the evolution of the technology and, by extension, the language.

For Francophone programmers, the power of the English language affects their daily lives. In order to succeed in the domain, it is more and more necessary to learn English and this extra prerequisite in a field that is already difficult to enter can create barriers for Francophone learners. On the other land, the necessity of the English language in the world of computer science contributes to bilingualism and multilingualism in non-Anglophone communities, since the desire to master computer science can motivate learners to improve their English. In coding, the data analysis implies that, although English dominates the comments of Francophone programmers, French still has a place.

As long as English holds a dominant position in the field of computer science, the lexicon of the French language will continue to borrow words from the former language, which will contribute to the Anglicization of French. These changes, however, are not necessarily negative: they are a natural phenomenon of the evolution of a language.