La dominance linguistique de l'anglais dans l'informatique

Alex Graves
8 mai 2019

FREN229 : Le français dans le monde
Sophie Degât-Willis

I. Introduction

Pour être programmeur, peu importe qui vous êtes, il faut que vous appreniez l'anglais. Le champ de l'informatique est mondial avec les programmeurs qui viennent de plusieurs pays et qui parlent une grande variété de langues. Cependant, une seule langue domine les langages de programmation, l'enseignement des concepts, et la recherche dans le domaine : l'anglais.

En fait, tous les langages de programmation qui sont fréquemment utilisés — comme le Java, le C++, le Python, et le JavaScript — sont écrits en anglais. Même le langage informatique Caml, qui a été développé en France à l'Institut national de recherche en informatique et en automatique (INRIA), emploie les mots-clés en anglais parce qu'il était basé sur un autre langage qui a utilisé l'anglais. D'autres exemples sont le langage très populaire Python, qui a été créé aux Pays-Bas et le Ruby, qui a été établi au Japon. Les deux étaient développés en anglais. Pour les non-anglophones, ce phénomène crée un environnement de contact constant entre l'anglais et leur langue maternelle. Cette interaction, comme tous les contextes des langues en contact, crée des mélanges, des interférences, et des emprunts. Plusieurs mots ont été traduits en français, comme *base de données* pour *database* et *intelligence artificielle* pour *artificial intelligence*, ou francisés, comme *déboguer* pour *debug*. Certains autres sont acceptés comme emprunts, tel que le mot Internet.

Cette situation soulève certaines questions principales au sujet de l'informatique et des langues : Pourquoi l'anglais est-il si dominant dans le domaine de l'informatique ? Comment la puissance de l'anglais dans le monde de l'informatique influence-t-elle les programmeurs francophones ? Pour explorer ces sujets, je présenterai une analyse de la documentation au sujet de l'histoire de l'anglais dans l'informatique et aussi des manifestations de l'influence de l'anglais sur le lexique technique français.

De plus, pour discuter les expériences humaines de programmeurs francophones, je raconterai une interview avec Eric Fouh, professeur d'informatique à l'Université de Pennsylvanie aux États-Unis, qui vient de la France et y a étudié. Cette interview présentera une comparaison entre l'enseignement de l'informatique en France et aux États-Unis, ainsi qu'une discussion des différences entre être programmeur dans les deux pays.

Finalement, je montrerai le processus et les résultats d'une analyse de données du site GitHub sur la fréquence du français et de l'anglais dans les fichiers de code qui étaient écrits par les francophones. Spécifiquement, cette étude examinera les différences entre les commentaires

dans les logiciels qui sont écrits par des francophones. Les pourcentages du français et de l'anglais dans les commentaires montrent la manifestation de la dominance de l'anglais dans le travail des programmeurs.

II. L'histoire de l'informatique

Le champ de l'informatique est dominé par la langue anglaise. Aujourd'hui, il est évident que les entreprises américaines sont à l'avant-garde du domaine, comme Apple, Google, et Microsoft. Cette dominance a persisté durant toute l'histoire de la discipline. Les deux personnes qui sont considérés les premiers pionniers de l'informatique, Charles Babbage et Ada Lovelace, sont nés, ont étudié, et ont travaillé en Angleterre au milieu du 19e siècle. Un autre informaticien qui a joué un rôle très important est Alan Turing, qui était aussi britannique. À cause de son travail sur les algorithmes et sa théorie de la calculabilité, il est souvent considéré le père de l'informatique.[1]

Dans les années 1960 et 1970, la croissance de Silicon Valley dans la région de la baie de San Francisco en Californie a aidé à établir la dominance des États-Unis dans le domaine de l'informatique, et par conséquent, a consolidé la puissance de l'anglais. Les nouvelles technologies qui y étaient développées ont déclenché la révolution des micro-ordinateurs, les premiers ordinateurs personnels. Dans les années 1980, l'industrie des ordinateurs personnels s'est énormément développée, et les modèles comme le ZX Spectrum et le Commodore 64 sont devenus très populaires. Le premier était produit par Sinclair Research, une entreprise britannique, et le dernier était créé par Commodore International, une entreprise américaine. Ces ordinateurs personnels ont supporté le langage de programmation BASIC.

BASIC, pour beaucoup de gens, était leur première exposition à l'informatique. Ainsi, les ordinateurs personnels et ce langage informatique ont servi comme la base du développement de l'industrie de programmation et de la popularisation de l'informatique comme un domaine et une carrière.[2] Cependant, à cause des origines des entreprises les plus fortes, les mots-clés du BASIC en plus de beaucoup de matériaux didactiques étaient en anglais.

[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.

Cette puissance de l'anglais dans les langages de programmation a persisté, et les langages de programmation les plus fréquemment utilisés aujourd'hui ont leurs mots-clés en anglais. Tous les facteurs qui constituent l'histoire du domaine ont contribué à une dominance de la langue anglaise.

III. La dominance de l'anglais dans le domaine de l'informatique

Comme l'histoire le montre, la langue anglaise a toujours eu une place important dans l'expansion du domaine de l'informatique. Cette dominance se manifeste dans le fait que l'anglais sert comme langue véhiculaire de l'informatique. J'examinerai ce phénomène du *lingua franca* en plus de la position de l'anglais dans l'enseignement de la discipline et les effets sur le lexique du français.

L'anglais comme lingua franca de l'informatique

L'anglais a longtemps été la langue véhiculaire de la science. Depuis la fin du 19e siècle, les États-Unis sont devenus dominants économiquement et politiquement dans le monde, un processus qui a été accéléré par les deux Guerres mondiales. Par conséquent, l'anglais a pris une position dominante dans la communication internationale et, similairement, dans le domaine de la science.[3] Ce fondement a facilité d'établir l'anglais comme le *lingua franca* de l'informatique pendant l'émergence du champ dans les années 1940 et 1950.

Pendant ce temps et aujourd'hui, le domaine de l'informatique a évolué rapidement et donc sa langue a changé avec la même vitesse. En plus, il y a plusieurs sous-disciplines (comme les algorithmes, l'intelligence artificielle, et l'infographie) qui grandissent séparément avec leurs propres lexiques développés autour le monde entier. Selon Jean-Bernard Kœchlin, « l'ordinateur est un lieu de contact des langues dominantes de la spécialité, l'anglo-américain et le français, qui tiennent ce rôle d'abord pour des raisons économiques »[4]. Ces facteurs créent une langue technique informatique qui est constamment en évolution et en contact avec les langues différentes, notamment l'anglais et le français.

[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.

Le champ de l'informatique est clairement très mondial, toujours une collaboration entre les chercheurs et les programmeurs de plusieurs pays autour le monde entier. Dans un champ si collaboratif et mondial, l'existence d'une langue véhiculaire est très utile et l'anglais sert comme ce *lingua franca*. Ce phénomène existe aussi dans les sciences naturelles, où l'anglais a continué à grandir sa puissance. En 1980, 74,6 % des publications des sciences naturelles étaient en anglais, et en 1996, ce chiffre est devenu 90,7 %. Parmi les disciplines différentes pour l'année dernière, les mathématiques et la physique ont des pourcentages les plus élevés (94,3 % et 94,8 %, respectivement). Ces sciences « pures » sont les plus similaires à l'informatique, qui ressemble beaucoup aux mathématiques. Le français commande juste 2,3 % des publications mathématiques en 1996. Dans l'ensemble, à la fin du 20e siècle, l'anglais était utilisé dans au moins de trois quarts de publications.[5]

De plus, l'anglais est spécifiquement une grande partie de la programmation. Quoiqu'il y ait eu les tentatives de créer les langages de programmation avec des mots-clés qui ne sont pas en anglais, il n'y a pas d'usage répandu. Dans les langages informatiques standards, les règles de codage demandent aux programmeurs d'écrire leurs commentaires et noms de variables, fonctions, et classes en anglais.[6] Dans une enquête par Philip J. Guo au sujet des barrières de l'apprentissage de l'informatique, 96 % des sondés ont dit qu'ils lisent des matériaux didactiques en anglais. Des personnes dans l'étude qui ont répondu qu'elles les lisent en plus d'une langue, seulement 1 % utilisent le français et l'anglais.[7]

L'anglais est la seule langue dominante dans le monde, et dans l'histoire mondiale, il n'y a jamais été une langue si dominante.[8] Il est évident que cette dominance existe aussi dans l'informatique, où l'anglais est clairement la langue véhiculaire.

Le rôle de l'anglais dans l'enseignement

La connaissance de l'anglais est un grand avantage dans le monde de l'informatique. Dans une enquête de l'Association des Informaticiens de Langue française (AILF) en 1984,

[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.

53 % des 180 sondés avait une « bonne ou très bonne » connaissance de l'anglais. De plus, une majorité « pour accéder à l'informatique, cherche à apprendre [l'anglais] l'idéal étant de faire des études aux E.-U. ». Cependant, 79 % des sondés ont caractérisé l'influence linguistique des États-Unis comme « une gêne » ou « une barrière » et 90 % avaient la demande de francisation pour « mieux communiquer » et « comprendre les informaticiens »[9].

Dans l'interview, Eric Fouh dit qu'en France, toutes ses classes de l'informatique étaient enseignées en français. Malgré ce fait, cependant, l'anglais a joué un rôle nécessaire dans son éducation. Pour l'informatique, Fouh dit que « on devait prendre des cours d'anglais. » Dans la salle de classe, l'anglais a aussi servi un but important. « Si le mot est en anglais, on utilise le mot en anglais, » explique-t-il. Par exemple, dans les langages de programmation, « on va dire *pour i partant de zéro à dix* » ce qui a dit *for i starting at zero to ten*. Il ajoute : « on écrit en anglais, mais on explique ça en français. »

Cet enseignement qui nécessite le bilinguisme peut avoir les effets négatifs. L'étude au sujet des barrières que les non-anglophones traversent en apprenant l'informatique et la programmation a trouvé que les plus grands problèmes étaient le matériel didactique qui est seulement en anglais, la communication technique en anglais, et les difficultés d'apprendre la programmation et l'anglais au même temps.[10] Pour certains sondés, le processus d'apprendre la programmation les a motivé à améliorer leur anglais.[11]

De plus, une étude par Sayamindu Dasgupta et Benjamin Mako Hill a trouvé que les novices qui ont appris la programmation d'un milieu qui était localisé dans la langue principale de leur pays ont montré les nouveaux concepts plus rapide que ceux qui ont utilisé le système en anglais. Ces conclusions s'alignent sur beaucoup de recherche qui implique que l'enseignement dans la langue maternelle aident l'apprentissage.[12]

Évidemment, la dominance de l'anglais dans le domaine de l'informatique joue un rôle important dans l'enseignement du sujet. Pour certaines personnes, la nécessité d'apprendre l'anglais peut créer des problèmes tandis que certaines autres trouvent que sa dominance les motive à développer leurs compétences de l'anglais. De plus, la présence et l'importance de

[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.

l'anglais dans l'enseignement de l'informatique consolident plus sa place comme langue véhiculaire du champ.

Les effets sur le lexique de la langue française

La dominance de l'anglais et sa présence constante dans le domaine de l'informatique créent une situation des langues toujours en contact. Selon Jean-Bernard Kœchlin, « L'anglo-américain influence l'orthographe, le lexique, la syntaxe et le style de la langue technique informatique, en un processus semblable à celui que l'on peut observer pour le commerce, la publicité, la musique et d'autres langues techniques. »[13] Par extension, l'anglais influence l'orthographe, le lexique, la syntaxe, et le style de la langue française en général. Ce sentiment est reflété par Beatrice Bagola, qui a écrit : « Le contact permanent avec l'anglais, ou plus exactement l'anglo-américain, langue principale de cette nouvelle technologie [Internet], influence la langue française. »[14]

En octobre 1986 à Paris, le colloque Franterm a trouvé sur 300 termes techniques d'aspect français, 20 % étaient des emprunts, 60 % des calques, et 20 % des termes autochtones.[15] Bien sûr, il y a beaucoup de nouveaux termes aujourd'hui, étant donné toutes les innovations techniques pendant les trente années passées. Un exemple qui a existé longtemps est *langage orienté objet* (pour désigner un langage de programmation qui utilise le structure d'objets) qui vient du terme anglais *object oriented language*. Dans certains cas, des termes crées par les utilisateurs restent dans le français courant, malgré les efforts de changer ou franciser les mots. Par exemple, le sigle *PC*, qui vient de l'anglais (des initiales de *personal computer*), est utilisé aussi en français. La tentative de remplacer ce terme avec *OP* pour *ordinateur personnel* était infructueuse, et *PC* subsiste dans le lexique français.[16]

Le développement des termes techniques est bien illustré par les cas de ceux d'Internet. Il y a souvent plusieurs termes utilisés pour décrire le même chose, comme pour le mot anglais *chat* : le terme officiel en France est *causette* ; en Québec, il est *clavardage* ou *bavardage* ; et le

[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.

mot le plus utilisé par les francophones sur Internet est *chat* ou *tchat*.[17] Ce phénomène est clairement un résultat des langues en contact, mais il n'est pas seulement à cause de cette situation : il est aussi influencé par les tentatives de l'aménagement linguistique. Selon Bagola, « On constate que le lexique utilisé sur Internet est caractérisé par une concurrence non seulement du français avec l'anglais, mais aussi à l'intérieur de la France ou de la francophonie. Il ne fait pas de doute que la langue s'enrichit grâce aux mesures législatives, mais elle le fait aussi par l'usage particulier de ses locuteurs. »[18]

Similairement, ce contact entre les langues existe dans les vies des programmeurs. Selon Élisabeth Eek dans une étude des différents ouvrages, « en effet, ceux-ci font preuve d'une richesse linguistique remarquable en dépit d'une influence marquée de la langue anglaise sur certains concepts, ceux, naturellement, qui ont été découverts aux États-Unis. »[19] Elle a discuté les termes utilisé pour les bases de données, comme *CREATE INDEX* ou *EXPLAIN* : « Le sort de ces termes quant à leur intégration dans le stock lexical français dépendra du niveau de culture linguistique des utilisateurs et de la bonne santé de la recherche informatique en France, capable à long terme d'engendrer plus de mots d'allure française. »[20] Évidemment, le lexique de la langue française est influencé par les nouvelles technologies, et quand elles sont développées par les anglophones, les termes sont empruntés de l'anglais. Son argument implique que la dominance d'un terme d'une certaine langue est déterminée en grande partie par le pays d'origine de la technologie. Ce phénomène peut aussi être vu dans les calques comme *sécurité de la couche de transport* (pour *transport layer security*, une spécification pour sécuriser des échanges sur Internet) ou *autorité de certification* (pour *certificate authority*, une entité qui utilise des certificats pour assurer la sécurité sur Internet) pour les technologies qui ont été développées aux États-Unis.

Pour certaines personnes, cette interaction entre les deux langues est vue comme préjudice. Dans l'étude de Bagola, elle a cité une publication du Secrétariat du Conseil du trésor du Québec qui a dit : « L'utilisation de l'anglais dans les technologies de l'information appauvrit la langue et la culture françaises. Le langage informatique est surtout conçu, développé et

[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.

commercialisé en langue anglaise et, par le fait même, l'anglais devient la langue de traitement des données et des textes. »[21] Eek avait un point plus positif : « Nous ne voyons point, par conséquent, le danger d'une pseudo « mort » de la langue française se profilant à l'horizon, par le biais d'une « invasion » de la langue américaine. »[22] Les deux auteures ont noté que le gouvernement joue un rôle important dans l'influence sur la langue française par l'aménagement linguistique. Bagola a encore cité la publication qui a déclaré que « le gouvernement doit jouer un rôle déterminant en matière de promotion et d'utilisation du français dans les technologies de l'information. En effet, en se dotant d'une politique dans le domaine, le gouvernement reconnaît ainsi toute la portée économique, sociale et culturelle du français à travers les technologies de l'information. »[23] Selon Eek, « Il appartient naturellement aux commissions de terminologie chargées de convertir les termes anglo-américains en gallicismes de se pencher sur la question sémantique des concepts derrière les mots importés, au lieu d'inventer des lexies de nature hermétique pour un locuteur français. »[24]

Quoique la langue anglaise ait une influence forte sur le français dans le domaine de l'informatique, le dernier continue à changer, comme toutes les langues dans les nouvelles situations de contact. Cette évolution est complètement naturelle, et la langue française devient plus riche par conséquent. Selon Eek, « tant que le stock lexical français sera capable de fournir une réponse de quelque forme que ce soit (néologisme, calque, dérivé, etc.) au défi linguistique anglo-américain et que cette forme revêtira l'aspect d'un gallicisme tel que la *puce* ou le *bogue* (pour *bug*), la langue française sera loin de s'étioler. »[25]

IV. Une analyse du français sur la base de données GitHub

Pour explorer le français et l'informatique plus profondément, j'ai fait une analyse des commentaires sur GitHub, qui est un site web auquel plusieurs programmeurs téléchargent leur code. Il existe sur le site une grande diversité des langages informatiques et des programmeurs des origines différentes. J'expliquerai les questions que j'ai étudiées, mes sources et mes méthodes, et les résultats de l'analyse.

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

Les questions à examiner

Il est évident que l'anglais est dominant dans le champ de l'informatique, et j'ai voulu examiner comment cette dominance se manifeste dans le travail des programmeurs francophones. Les langages de programmation ont certains mots-clés qui sont toujours en anglais, et par conséquent les lignes de code dans les logiciels ne devraient pas varier beaucoup entre les langues. Cependant, les programmeurs écrivent souvent les commentaires pour expliquer leur code, et ils peuvent être écrits dans n'importe quelle langue. Dans l'interview avec Eric Fouh, il a dit que « en France, je vais probablement écrire les commentaires en français. » Mais, si un programmeur veut que le programme soit international, il a dit qu'il écrit « les commentaires en anglais pour que tout le monde lise. » Donc, pour mon analyse, j'ai posé la question : Quel pourcentage de commentaires des programmeurs francophones sont écrits en français ? Quel pourcentage en anglais ?

Les langues diffèrent considérablement entre les régions francophones. Il y a certains pays où le français est la seule langue officielle et il est parlé par presque tout le monde, comme en France, mais aussi les régions où le français existe à côté d'autres langues, comme au Québec avec l'anglais. Donc, j'ai demandé : Comment l'usage des langues varie-t-il entre les différentes régions francophones ?

Mes sources et mes méthodes d'analyse

J'ai utilisé le site GitHub comme une base de données, puisqu'il y a beaucoup de fichiers de code des programmeurs du monde entier. Pour trouver les endroits des utilisateurs, j'ai commencé avec un ensemble de données qui s'appelle GHTorrent. De cet ensemble, j'ai trouvé le pourcentage d'utilisateurs qui spécifient leur endroit :

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

J'ai choisi d'examiner cinq régions francophones : la France, la Suisse, la Belgique, le Québec, et le Sénégal. Ces régions représentent une sélection diverse de situations linguistiques dans la francophonie. Pour chacune, j'ai fait les requêtes suivantes :

```
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;
```

La première a trouvé les noms d'utilisateur pour chaque programmeur dans le pays spécifié. J'ai utilisé ces résultats plus tard pour connecter les pays et les commentaires dans les

fichiers de code. Les deux autres requêtes trouvent le pourcentage et le total des utilisateurs de ce pays, respectivement.

Ensuite, j'ai eu besoin de connecter les utilisateurs et leurs fichiers de code. Pour cette étape, j'ai utilisé le service Google BigQuery, qui m'a permis d'analyser rapidement les grands ensembles de données. Premièrement, j'ai fait une requête à un ensemble qui s'appelle GitHub Archive, qui m'a permis de préciser les dates. J'ai utilisé la requête suivante pour trouver les référentiels (les lieux où les programmeurs gardent leur code) auxquels les programmeurs francophones ont contribué du code en avril 2019. Le mot pour contribuer à un référentiel est « push » (utilisé dans *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;
```

Après avoir trouvé les référentiels, j'ai eu besoin de sélectionner les fichiers dans ces référentiels et leurs contenus. Pour cette étape, j'ai utilisé un autre ensemble sur BigQuery qui s'appelle GitHub Repos. J'ai connecté les noms des référentiels de l'ensemble GitHub Archive avec ceux de cet ensemble, et puis j'ai sélectionné les fichiers et leurs contenus qui sont dans ces référentiels. Pour limiter mon ensemble de données, j'ai choisi seulement les fichiers qui utilisent les langages de programmation JavaScript, Python, ou Java. Ces langages sont tous très populaires et sont utilisés dans une variété de contextes, comme le génie logiciel, la conception de sites web, et la science de données.

```
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;
```

Après toutes les requêtes ci-dessus, j'ai eu un ensemble de données qui a contenu le nom du fichier (et donc son langage de programmation), le texte dans le fichier, et le pays du programmeur qui a contribué au référentiel auquel le fichier appartient. L'ensemble sur BigQuery était immense, et cette requête a pris 52,8 secondes et elle a traité 2,5 téraoctets de données.

L'étape finale de la collecte de mes données était d'extraire les commentaires des fichiers. J'ai écrit un logiciel (dans le langage Python) qui a lu les contenus de mon ensemble de données. Les commentaires en Python sont marqués par une ligne qui commence avec #, et en JavaScript ou Java, ils commencent avec //, /∗, ou ∗. Pour chaque fichier, j'ai inspecté chaque ligne pour les symboles appropriés de commentaires, et s'il y en avait, j'ai ajouté le commentaire à un ensemble de données final. Cet ensemble a contenu le langage de programmation, le pays, et le contenu du commentaire.

Enfin, j'ai eu besoin d'analyser les langues des commentaires. J'ai utilisé encore le langage Python et un module qui s'appelle « langdetect » pour évaluer la langue de chaque commentaire. J'ai choisi de décomposer les résultats par le pays des programmeurs qui ont écrit les commentaires et les langages de programmation desquels ils sont venus.

Discussion des résultats de l'analyse

Premièrement, je présenterai les statistiques sur mes ensembles de données. Pour mieux comprendre les résultats, il est important de reconnaître les problèmes potentiels avec les données initiales.

De l'ensemble de données GHTorrent, j'ai trouvé que seulement 7,58 % des utilisateurs ont spécifié leur endroit. Quoique ce pourcentage soit assez bas, il correspond à 2 405 811 de programmeurs de GitHub.

De ces utilisateurs, 2,99 % ont leur endroit comme la France. Ce pourcentage correspond à 71 852 personnes. Pour la Suisse, le pourcentage est 0,77 % (18 553 personnes). La Belgique est 0,61 % (14 630 personnes), le Québec est 0,02 % (494 personnes), et le Sénégal est 0,01 % (354 personnes). Une complication avec le Québec est qu'il est une province, et donc les utilisateurs doivent spécifier leur état en plus de leur pays. Par conséquent, le pourcentage des programmeurs québécois est probablement plus bas qu'en réalité.

Après mes connexions entre les autres ensembles de données, j'ai eu 11 982 844 de commentaires au total. Pour les pays : la France en a 8 571 475, la Suisse en a 2 196 470, la Belgique en a 1 036 290, le Québec en a 174 151, et le Sénégal en a 4 458. Pour les langages de programmation, il y a 6 562 640 de commentaires en Java, il y en a 3 728 116 en JavaScript, et il y en a 1 692 088 en Python. Malheureusement, je ne pourrais pas limiter les commentaires seulement aux ceux qui étaient écrits par les francophones puisque la connexion a dû utiliser les

référentiels, auxquels d'autres programmeurs (peut-être les non-francophones) auraient pu contribuer. Néanmoins, j'ai pensé que les commentaires constituent un échantillon représentatif de ceux des francophones.

Pour commencer, je décomposerai les résultats par les langages de programmation des programmes dans lesquels les commentaires étaient trouvés. Le tableau suivant montre les cinq premières langues qui ont paru dans les commentaires :

Java	**JavaScript**	**Python**
anglais (80,26 %)	anglais (73,88 %)	anglais (78,90 %)
rien (2,52 %)	français (3,43 %)	français (3,37 %)
allemand (2,08 %)	catalan (2,61 %)	catalan (2,58 %)
roumain (1,90 %)	italien (2,23 %)	roumain (1,65 %)
italien (1,69 %)	roumain (2,19 %)	italien (1,58 %)

Il est immédiatement évident que l'anglais domine les résultats, mais aussi qu'il y a des différences significatives entres les pourcentages dans les langages de programmation. Dans les programmes écrits en Java, il y a souvent des commentaires qui sont automatiquement générés qui expliquent les fonctions, ce qui pourrait expliquer le pourcentage haut de l'anglais dans les commentaires du Java. Similairement, pour Python, beaucoup de programmeurs mettent la licence du programme au début du fichier. Les licences sont normalement copiées d'un format général qui est en anglais, ce qui pourrait augmenter le pourcentage.

Pour JavaScript et Python, le français est le prochain pourcentage, et les langues qui le suivent sont des langues avec beaucoup de similarités au français. Plutôt qu'il y a un (relativement) grand pourcentage des programmeurs catalans, il est plus probable que le module a faussement classé la langue. Par conséquent, il est possible que le pourcentage de commentaires français est en fait plus élevé que les résultats.

Pour Java, le deuxième pourcentage est « rien ». Dans les cas où le module ne pourrait pas détecter la langue (par exemple, si le commentaire était une URL), j'ai marqué la langue comme « rien ». La raison que ce pourcentage est si haut est possiblement que les commentaires qui sont automatiquement générés incluent souvent les URL pour la documentation du programme. L'allemand est le troisième pourcentage, ce qui est aussi probablement expliqué par les résultats incorrects à cause des commentaires générés. Le manque du français dans les cinq

premières langues m'a étonné, et j'ai trouvé que le français est la sixième langue avec 1,65 % de commentaires.

Ensuite, je comparerai les langues des commentaires par les régions. Le tableau suivant montre les cinq premières langues qui ont paru dans les commentaires :

La France	La Suisse	La Belgique	Le Québec	Le Sénégal
anglais (78,31 %)	anglais (78,12 %)	anglais (77,59 %)	anglais (69,30 %)	anglais (79,53 %)
français (2,49 %)	catalan (2,73 %)	rien (2,59 %)	allemand (9,24 %)	français (4,62 %)
roumain (1,99 %)	français (2,31 %)	français (2,36 %)	français (2,66 %)	danois (2,24 %)
rien (1,83 %)	italien (2,01 %)	catalan (1,99 %)	rien (2,24 %)	catalan (2,02 %)
italien (1,81 %)	roumain (1,86 %)	italien (1,94 %)	catalan (2,06 %)	italien (1,91 %)

Il est encore évident que l'anglais constitue la grande majorité des commentaires. Les présences du catalan et de « rien » comme le deuxième pourcentage pour la Suisse et la Belgique sont probablement expliquées par les raisons ci-dessus. Cependant, j'ai trouvé que le cas du Québec est fascinant : l'anglais constitue seulement 69,3 % des commentaires et le deuxième pourcentage est l'allemand, avec 9,24 %. J'ai noté que l'allemand est aussi plus élevé que j'avais attendu pour Java, et donc j'ai examiné plus la distribution des commentaires. Des données du Québec, 167 976 commentaires sont venus des programmes qui étaient écrits en Java, tandis qu'il y en a juste 4 170 du Python et 2 005 du JavaScript. Évidemment, les problèmes créés par le Java ont eu un effet énorme sur les résultats du Québec.

Pour évaluer les résultats sans l'influence du Java, j'ai refait l'analyse, en excluant ce langage de programmation, pour le Québec et toutes les régions :

Le Québec (sans Java)	Toutes les régions (sans Java)
anglais (76,31 %)	anglais (75,45 %)
français (4,36 %)	français (3,42 %)
catalan (4,18 %)	catalan (2,60 %)
roumain (2,35 %)	italien (2,03 %)
italien (1,64 %)	roumain (2,02 %)

Avec ces ajustements, il est bien évident que la langue française constitue environ 3,42 % des commentaires des programmeurs francophones. Il semble que le pourcentage soit au moins

un peu plus élevé parce que les fréquences du catalan, de l'italien, et du roumain sont partiellement exagérées en conséquence des erreurs du module. Cependant, malgré tous ces faits, l'anglais est clairement la seule langue dominante des commentaires des programmeurs francophones.

V. Conclusion

Le lexique de l'informatique continuera à être dominé par l'anglais, et par conséquent la langue française dans le domaine de l'informatique évoluera toujours. La langue anglaise, qui a atteint cette dominance en raison du rôle de l'Angleterre et des États-Unis dans l'histoire du champ, sert comme langue véhiculaire de l'informatique. Cette position est renforcée par l'enseignement de la discipline, qui nécessite pour la plupart la connaissance de l'anglais. Évidemment, ces situations créent un contact constant entre l'anglais et le français. Donc, le lexique français de l'informatique contient beaucoup de calques en conséquence de la vitesse de l'évolution de la technologie et, par extension, de la langue.

Pour les programmeurs francophones, la puissance de la langue anglaise affecte leurs vies quotidiennes. Pour réussir dans le domaine, il est de plus en plus nécessaire d'apprendre l'anglais et cet autre prérequis d'un champ qui est déjà difficile d'entrer peut créer des barrières pour les apprenants francophones. D'un autre côté, la nécessité de la langue anglaise dans le monde de l'informatique contribue au bilinguisme et au plurilinguisme dans les communautés non-anglophones, puisque le désir de maîtriser l'informatique peut motiver les apprenants à améliorer leur anglais. Dans le codage, l'analyse de données implique que, quoique l'anglais domine les commentaires des programmeurs francophones, le français a toujours une place.

Tant que l'anglais a cette position dominante dans le champ de l'informatique, le lexique de la langue française continuera à emprunter les mots de la première langue, ce qui contribuera à l'anglicisation du français. Ces changements, cependant, ne sont pas nécessairement négatifs : ils sont un phénomène naturel de l'évolution de la langue.