h e g

Helve’tweet : exploration d’un million de tweets géolocalisés en Suisse, février-août 2017

Mémoire de recherche réalisé par :

Agnes A. MOTISI-NAGY

Tania ZUBER-DUTOIT

Sous la direction de :

Arnaud GAUDINAT, professeur HES

Genève, le 17 janvier 2018

Master en Sciences de l’information

Haute École de Gestion de Genève (HEG-GE)

Hes·so GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

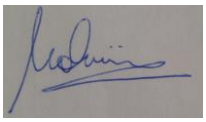
Déclaration

Ce mémoire de recherche est réalisé dans le cadre du Master en Sciences de l'information de la Haute école de gestion de Genève. L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans ce travail, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur/des auteurs, ni celle de l'encadrant.

«Nous attestons avoir réalisé le présent travail sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 17 janvier 2018

Agnes-Anna Motisi-Nagy



Tania Zuber-Dutoit

Remerciements

Ce travail n'aurait pas pu voir le jour sans les personnes suivantes, que nous tenons à remercier pour leur aide :

Arnaud Gaudinat pour avoir mandaté ce projet

Bastien Berger pour la récolte des tweets, la relance des serveurs et les réponses à nos questions

Claire Wuillemin, pour le semestre passé sur le projet et la douloureuse rédaction du *DMP*

Jean-Philippe Grimaître pour l'impression du poster

Jens Ingensand, professeur à la HEIG-VD pour l'utilisation de QGIS

Barbara Signori, responsable du service e-Helvetica de la Bibliothèque nationale suisse

Nos relectrices : Marie-Claire Dutoit, Christine Guinchard, Claire Wuillemin

Nos familles pour nous avoir supportées durant ce projet, particulièrement les derniers jours.

Résumé

Réseau social utilisé activement par 8% de la population suisse[1], Twitter permet à ses utilisateurs de géolocaliser leurs messages. Cette étude exploratoire quantitative, basée sur des messages géolocalisés en Suisse écrits entre le 18 février et le 31 août 2017, fait suite au projet GEoTweet consacré aux tweets genevois en 2014-2015. Elle se propose de répondre à trois questions de recherche pour évaluer les possibilités et les limites de l'utilisation des données fournies par l'API de Twitter lors des recherches sur la Suisse, dans les domaines de la sociologie des données et des sciences de l'information. Le focus est porté plus spécifiquement sur l'exploitation des données de géolocalisation, sur la problématique de l'identification des langues et sur les critères définissant un tweet suisse dans une perspective d'archivage.

Après l'introduction et la revue de littérature, le rapport présente la méthodologie utilisée, les biais identifiés et les outils créés pour les mesurer, les éviter ou du moins les minimiser. Une concordance a ainsi été créée entre les *place.id* de Twitter et la liste officielle des communes suisses pour pallier au caractère non vérifié (en partie obsolètes, en partie erronées) des données géographiques fournies par Twitter. Trois séries de tests ont également été menés pour vérifier la fiabilité de l'algorithme de reconnaissance de langue de Twitter pour l'échantillon. Ils montrent une marge d'erreur de 4,25% sur les grandes langues européennes, mais qui peut monter jusqu'à 92% pour une langue « exotique » comme l'indonésien.

Les analyses des tweets et des twittos ont permis de dégager des résultats importants. D'une part, elles montrent les fortes variations de leur nombre et de leur diversité linguistique à travers l'espace et le temps (p.ex. plus de comptes actifs en Suisse alémanique, mais plus de tweets en français dans l'ensemble ; plus de tweets pendant les périodes de vacances, mais baisse de la proportion des tweets et des twittos en langues nationales et en anglais). D'autre part la durée et l'étendue géographique de leur activité sont très variables (p.ex. 82% des comptes avec moins de 10 tweets, 68% actifs pendant un seul mois et 71% dans un seul canton). Des hypothèses ont été formulées et vérifiées pour expliquer ces résultats qui relèvent de la propension élevée des germanophones à twitter en anglais et de l'effet positif des loisirs sur l'envie et l'opportunité de twitter avec géolocalisation.

Dans la dernière partie, l'étude propose des pistes afin d'établir des critères pour reconnaître un tweet suisse, en se basant sur les analyses menées préalablement ainsi que sur les expériences menées dans d'autres pays du monde. Le contexte international et suisse de l'archivage des tweets est abordé, sans prétention de vouloir proposer une méthode, au vu de la complexité des enjeux sociologiques, techniques et légaux.

Mots-clefs : Twitter – étude exploratoire – Suisse – géolocalisation – langues

[1](Latzer et al., 2017)

Table des matières

Liste des tableaux

Liste des figures

Glossaire

Des hyperliens renvoient aux entrées du glossaire à la première occurrence des termes dans chaque chapitre. Ils se distinguent par l'utilisation d'une majuscule au début du mot.

Bounding.box : rectangle défini par des coordonnées géographiques, latitude et longitude qui entoure la localité*. Celles-ci sont consignées dans le champ *bounding.box.coordinates*.

Champs '*place'* : Les champs 'place' font partie des métadonnées des tweets géolocalisés : place.id, place.name, place.full.name, place.country.code etc. Ces métadonnées correspondent aux champs de la base de données géographique utilisée par Twitter pour proposer des localités* pour la *Géolocalisation manuelle*, ainsi que pour attribuer une localité aux coordonnées géographiques exactes d'une *Géolocalisation fine*. Le public n'a pas accès direct à cette base de données et les méthodes de sa constitution ainsi que de ces éventuelles mises-à-jour sont inconnues. Chaque localité* contenue dans la base contient une place.id et un place.country.code (code du pays dans lequel la localité se trouve). À une place.id correspondent en revanche plusieurs place.name et full.place.name, car le nom de la localité est traduit dans plusieurs langues.

Compte Twitter privé : les comptes privés doivent valider les demandes de suivi des autres twittos. Leurs tweets sont privés également : ils n'apparaissent pas dans les recherche ni dans les API, et ne peuvent pas être retweetés.

Compte Twitter public : les comptes sont publics par défaut, ainsi que les tweets associés. Un twittos peut suivre un compte public sans autorisation spécifique, voire même sans être inscrit à Twitter.

DMP (Data Management Plan) : plan de gestion des données.

Durée d'activité : voir *période d'activité.*

Géolocalisation : procédé manuel ou automatique qui permet d'ajouter une localité* à un tweet. Les données de géolocalisation alimentent les *Champs* '*place'* et/ou le champ *coordinates.coordinates*. À ne pas confondre avec localisation, procédé qui utilise d'autres méthodes pour détecter le lieu d'émission du tweet ou du twittos*.

Géolocalisation fine : procédé de géolocalisation automatique par les coordonnées GPS de l'appareil émetteur. Cette méthode alimente directement le champ de métadonnées *coordinates.coordinates* avec les coordonnées géographiques exactes du lieu d'émission. A partir de ce champ, Twitter calcule la localité* la plus proche qu'il consignera dans le champ *place.id* ainsi que dans les autres *Champs* '*place'*.

Géolocalisation manuelle : procédé de géolocalisation où l'utilisateur choisit une localité* dans une liste déroulante proposée par l'application quand il rédige son tweet. La liste est fournie par la base de données géographiques de Twitter. Cette méthode

alimente directement le champ de métadonnées *place.id* ainsi que tous les *Champs* '*place'*.

Hashtag : sujet sur Twitter, précédé d'un dièse, par exemple #switzerland. Ils sont utilisés pour regrouper tous les messages sur le même sujet, en cliquant sur le mot.

Langue productrice : voir source.lang.

Langue de compte : voir *user.lang.*

Localité Twitter : entité munie d'une *place.id*, faisant partie de la base de données géographiques de Twitter. Pour la Suisse les localités Twitter incluent les communes (city), les cantons (admin) et le pays (country).

Période ou durée d'activité : notion liée uniquement à ce projet. Nombre des mois du calendrier pendant lesquels un utilisateur twitte au moins une fois. Les périodes d'activités vont de 1 à 7 mois.

Source.lang ou langue productrice : notion liée uniquement à ce projet. Langue dans laquelle le tweet est rédigé, détectée automatiquement par Twitter. Elle est encodée selon la norme ISO 639-1 à deux lettres dans le champ *lang.keyword*.

Trending topic : sujets tendance, utilisés par de nombreuses personnes durant un temps donné, affichables par ville, pays ou monde entier.

Tweet : message de 140 caractères (280 dès novembre 2017) envoyé sur le réseau social Twitter.

Tweet privé : voir *Compte Twitter privé.*

Tweet public : voir *Compte Twitter public.*

Twitter API (Application Programming Interface) : interface qui permet de se connecter aux serveurs de Twitter pour récupérer des tweets de manière automatisée. L'API Streaming est gratuite mais ne fournit que 1% des tweets mondiaux, Firehose coûte cher mais fournit l'intégralité des tweets mondiaux. Entre les deux, d'autres API payantes fournissent des tweets en réponse à des recherches précises (utilisateur, hashtag, …).

Twittos : personne qui possède un compte Twitter. Dans notre étude, synonyme de compte*.

User.lang ou langue de compte: langue du compte choisie par le *twittos*. Celui-ci peut la changer librement et à tout moment. Elle est encodée par Twitter selon la norme 639-1 à deux lettres ou ISO 639-2 à trois lettres dans le champ *user.lang*.

Utilisateur actif : dans la littérature professionnelle ce terme peut avoir deux sens. Soit une personne qui possède un compte Twitter et qui se connecte régulièrement pour consulter ou pour rédiger des messages, soit uniquement celles qui rédigent un ou plusieurs messages. Dans cette étude, nous utiliserons le terme uniquement dans ce deuxième sens.

Politique de confidentialité de Twitter : extrait

> Ce que vous communiquez sur Twitter est visible partout dans le monde instantanément. Vous êtes ce que vous tweetez !

« **Tweets, abonnements, listes, profil et Autres informations publiques°:** Twitter est avant tout conçu pour vous aider à partager des informations avec le monde entier. La plupart des informations que vous nous fournissez par le biais de Twitter sont des informations que vous souhaitez rendre publiques. Vous pouvez nous fournir des informations sur votre profil, par exemple une courte biographie, votre localisation, votre site Web, votre date de naissance ou une photo. En outre, vos informations publiques comprennent les messages que vous tweetez, les métadonnées fournies avec les Tweets, telles que la date de vos Tweets et l'application client utilisée pour vos Tweets, des informations sur votre compte telles que la date de création, la langue, le pays et le fuseau horaire, ainsi que les listes créées, les personnes que vous suivez, les Tweets que vous marquez comme « J'aime » ou que vous retweetez, et des vidéos Periscope sur lesquelles vous cliquez ou avec lesquelles vous interagissez autrement (par exemple, en les commentant ou en les aimant) sur Twitter. Twitter diffuse instantanément et largement vos informations publiques à un large éventail d'utilisateurs, de clients et de services, y compris les moteurs de recherche, les développeurs et les éditeurs qui intègrent les contenus Twitter dans leurs services, ainsi que des entités telles que des universités, des organismes de santé publique et des entreprises d'études de marché qui analysent les informations pour en tirer des tendances et des comportements. Quand vous partagez des informations ou des contenus tels que des photos, des vidéos et des liens via les Services, vous devez réfléchir sérieusement à ce que vous rendez public. » [2]

[2] https://twitter.com/privacy?lang=fr

1. Introduction

Avec 300 milliards de tweets envoyés depuis le 21 mars 2006, plus de 500 millions chaque jour ; 330 millions d'utilisateurs mensuels au troisième trimestre 2017, dont 14% chaque jour (Twitter Inc., 2017b), Twitter est un réseau social vivant, qui intéresse de plus en plus les chercheurs dans différents domaines : sociologie, médecine, géographie, marketing, ..., notamment car il donne largement accès aux messages et autres informations transmises par son intermédiaire.

Après GEoTweet[3] (Jeanneret, 2015 ; Banfi, Béguelin, 2016), première étude sur les tweets genevois menée en 2015 dans le cadre du Master en Sciences de l'information, le professeur Arnaud Gaudinat souhaitait élargir la recherche au niveau suisse et disposer de données statistiquement éprouvées. A la différence de Geotweet, notre étude prendra en compte non seulement les tweets géolocalisés par la machine, mais également ceux qui sont attribués à un lieu par leur auteur. Certaines hypothèses genevoises seront reprises et étendues au contexte suisse, tandis que d'autres seront propres à notre projet.

Cette recherche scientifique, fondamentale, quantitative et exploratoire s'inscrit dans la continuité d'une série de projets effectués dans le domaine de la sociologie des données dirigés ou supervisés par le professeur A. Gaudinat. Elle a une double visée : obtenir des résultats statistiquement représentatifs sur des données non biaisées de tweets géolocalisés en Suisse et ouvrir des pistes de réflexion sur ce qui peut être considéré comme un "tweet suisse" dans le contexte de la problématique de l'archivage de ces messages.

Les questions de recherche auxquelles nous voulons apporter des réponses sont les suivantes :

1. Comment peut-on exploiter la géolocalisation de Twitter dans le contexte suisse ?
2. Dans quelle mesure le système d'identification automatique des langues par Twitter permet-il d'obtenir une image réelle de la diversité linguistique de la Suisse ?
3. Est-il possible de définir un tweet suisse à partir des données de notre échantillon ?

Les objectifs spécifiques permettant de répondre à ces trois questions sont les suivants :

1. **Mener une recherche exploratoire quantitative sur les tweets géolocalisés *publics* émis en Suisse en portant une attention particulière à leur géolocalisation et leurs langues :**

    1.1 Analyser statistiquement 1 million de tweets publics géolocalisés en Suisse, récoltés entre février et août 2017 et en visualiser la répartition spatiale, linguistique et temporelle.

    1.2 Vérifier sur l'échantillon les hypothèses suivantes du projet GEoTweet (Banfi, Béguelin, 2016) en les adaptant à l'échelle Suisse et proposer de nouvelles hypothèses pour expliquer la cause des résultats :

        1.2.a « Genève est un canton francophone, mais la pratique virtuelle linguistique via Twitter s'approche plutôt de celle d'une région bilingue (franco-anglaise) associée à la présence d'une importante variété linguistique » (H 1.1a GEoTweet)
        Hypothèse modifiée pour la Suisse : La Suisse est un pays avec trois langues officielles reconnues par Twitter, mais la pratique virtuelle

[3] http://geotweet.hesge.ch

linguistique montre quatre langues principales, avec l'anglais comme deuxième langue dans tous les cantons et première langue au niveau de la Suisse.

1.2.b « La quantité de tweets varie en fonction des heures de la journée et des jours de la semaine. La variation des communautés linguistiques virtuelles sur Twitter évolue différemment selon les heures de la journée, les jours de la semaine et les dates » (H 1.1b GEoTweet).
Hypothèse modifiée pour la Suisse : La quantité des tweets ainsi que les communautés linguistiques virtuelles sur Twitter évoluent différemment selon les mois.

2 **Mener une réflexion sur la notion et la définition de "tweet suisse" dans le cadre de l'archivage des tweets :**
- 2.1 Répertorier les projets d'archivage de tweets au niveau international
- 2.2 Proposer des pistes pour une définition de ce qu'est un "tweet suisse"

Bien que la recherche porte sur deux populations distinctes, les tweets publics géolocalisés en Suisse d'une part et les tweets suisses d'autre part, elle se base sur le même échantillon : 1'039'819 tweets publics géolocalisés en Suisse, émis et collectés entre le 18 février 01:00.00 GMT et le 31 août 2017 23:59.59 GMT.

La collecte des données s'est déroulée hors du périmètre de la recherche et a été effectuée par Bastien Berger, assistant à la Haute école de Gestion de Genève. La recherche proprement dite a été menée entre le 15 mars 2017 et le 10 janvier 2018 sous la supervision du Prof. Arnaud Gaudinat. La première partie de la phase conceptuelle (prise en main de Twitter et de Kibana, élaboration du *DMP*) a été menée conjointement par Agnes A. Motisi-Nagy, Tania Zuber-Dutoit et Claire Wuillemin. L'élaboration des outils d'analyse, les analyses quantitatives et qualitatives des données, la synthèse ainsi que la rédaction du rapport sont le fruit du travail d'Agnes A. Motisi-Nagy et de Tania Zuber Dutoit.

Notre rapport est structuré de la manière suivante : un glossaire définit les termes propres à Twitter ou les notions spécifiques abordées dans ce travail, puis une note tirée des conditions générales de Twitter avertit le lecteur sur l'utilisation des messages de cette plateforme. Suivent une revue de littérature explorant différents thèmes et qui nous a servi pour déterminer les questions de recherche, puis les aspects méthodologiques, les analyses menées sur les données ainsi que leurs visualisations et enfin les propositions pour définir un tweet suisse. Le rapport se termine sur les principaux résultats et des recommandations pour des recherches futures.

2. Revue de littérature

Twitter suscite un intérêt grandissant parmi les scientifiques, ce qui se traduit par un nombre croissant d'articles ou autres communications sur le sujet[4]. La revue de littérature qui suit a été établie tout au long du projet Helve'tweet, car les articles étaient écrits (et donc mis à disposition) en même temps que nous effectuions nos recherches.

Les bases de données et moteurs de recherche suivants ont été utilisés : ACM, IEEE, Google, Google Scholar, Qwant et Web of Science.

2.1 Twitter et ses utilisateurs

A notre connaissance, le nombre total d'inscrits sur Twitter n'est pas public ; des chercheurs l'estiment à près d'un milliard (Bruns, Weller, 2016). La répartition par pays n'est pas publique[5], mais sert au marketing avec d'autres informations captées par la plateforme : « *adresse IP, des données de localisation GPS précises ou des informations sur les réseaux sans fil ou les antennes relais- se trouvant à proximité de votre appareil mobile. Nous utilisons les informations relatives à votre localisation pour déterminer le paramètre de pays »*[6] (voir aussi Annexe 1).

2.1.1 En Suisse

Sans chiffres officiels, plusieurs personnes ou sites ont tenté d'estimer la Twittosphère suisse ou romande.

Pegasus Data (2013) : malheureusement abandonné, ce projet proposait plusieurs méthodes pour compter les *Twittos* romands : de l'estimation générale (115'000 utilisateurs), en passant par le comptage manuel (20'000 utilisateurs), pour arriver à l'évaluation comparative avec la France et la Belgique (10-34'000 utilisateurs), le site terminait ainsi : « *À la manière d'un sondage, il ne nous reste plus qu'à conclure qu'il y a sur Twitter quelques dizaines de milliers de Romandes et Romands, plus ou moins quelques milliers…* »

PME-Web – Mathieu Corthésy estimait en janvier 2015 la population suisse utilisant Twitter à 625'000 personnes (Corthésy, 2015). En se basant sur les 330 millions d'utilisateurs signalés par Twitter pour le troisième trimestre 2017, cela donnerait près de 757'000 *Utilisateur actif*s.

Un billet de blog récent (Plewnia, 2017) nous apprend que Twitter serait en quatrième position parmi les réseaux sociaux en Suisse, qu'il y aurait 702'000 usagers mensuels et une audience totale de 2,7 millions de personnes.

D'après une étude bisannuelle de l'Université de Zurich sur l'utilisation d'internet en Suisse, les applications de médias sociaux sont très répandues en Suisse (employées par 62% des personnes interrogées). 16% utiliseraient Twitter, dont 8% d'entre eux écriraient activement des messages (Latzer et al., 2017).

[4] 2'625 articles en 2016, 1'802 en 2017 en date du 12 janvier 2018 sur Web of Science

[5] Alors que le champ « Pays » existe dans le compte utilisateur, celui-ci n'est pas recherchable dans l'API Streaming de Twitter.

[6] https://help.twitter.com/fr/managing-your-account/how-to-change-country-settings

2.1.2 En France

Le Blog du modérateur publie régulièrement des informations sur Twitter et ses utilisateurs. Selon les derniers chiffres Médiamétrie (juillet 2016), Twitter comptabilise 21,8 millions d'utilisateurs mensuel (MAU) et 4,27 millions d'utilisateurs par jour (DAU) (Coëffé, 2017).

Depuis 5 ans, Harris Interactive réalise des études sur l'utilisation des réseaux sociaux par des personnes âgées de plus de 16 ans. L'étude (2017) présente Twitter comme le 3e réseau social le plus utilisé : 21% des sondés l'avaient utilisé au cours des 30 derniers jours, et 9% en ont un usage quotidien. 59% des utilisateurs sont des *millenials*, soit âgés entre 15 et 35 ans.

2.1.3 En Allemagne

Andreas Rickmann, journaliste spécialisé en réseaux sociaux, estime à 880'000 comptes actifs pour une audience totale de 12 millions de personnes (2017).

ExtraDigital (2016), agence spécialisée dans le marketing digital, faisait remarquer la difficulté d'utiliser Twitter en allemand, à cause de la longueur des mots de cette langue, en contradiction avec la concision requise par les 140 caractères de Twitter. Cela peut expliquer une utilisation plus 'passive', ou une tendance à écrire en anglais.

2.1.4 En Italie

Twitter est utilisé surtout depuis 2011 en Italie, suite à l'adoption par des célébrités de la télévision italienne, mais surtout l'explosion des dispositifs mobiles d'accès à internet. Les utilisateurs actifs étaient alors estimés à 4,5 millions (Mari, 2013), puis 6,4 millions en 2016 (Mosca, 2016).

2.1.5 Robots et cyborgs

Les robots (*automated agents*) sont utilisés pour différentes raisons : diffusion d'informations, transmissions urgentes en cas de crise, marketing, infiltration politique ou diffusion de contenus malveillants (Gilani et al., 2017a). Comme les robots essaient de se fondre parmi les autres comptes d'utilisateurs, il est difficile de savoir combien ils sont. Dans un autre article, Gilani donne une définition très claire de ce qu'il a utilisé pour distinguer les humains des robots : « *We define a 'bot' as any account that consistently involves automation over the observed period, e.g. use of the Twitter API or other third party tools, performing actions such as automated likes, tweets, retweets, etc.* » (Gilani et al., 2017b, p. 4).

Twitter estimait dans un rapport à l'« US Securities and Exchange Commission » en 2014 la part de faux comptes ou spammeurs à moins de 5%. Dans le même rapport, Twitter estimait que « *only up to approximately 8.5% of all active users used third party applications that may have automatically contacted our servers for regular updates without any discernable additional user-initiated action*°»[7] Au vu des 271 millions de comptes actifs lors de la parution de ce rapport, les premiers représenteraient 13,5 millions de comptes et les seconds 23 millions.

D'autres études estiment qu'il y a encore plus de comptes 'non-humains' ou 'partiellement non-humains'. En étudiant une soixantaine de critères sur la typologie du contenu et les métadonnées accessibles, Varol et al. (2017) estime la part de robots entre 9 et 15 %, mais

[7] https://www.sec.gov/Archives/edgar/data/1418091/000156459014003474/twtr-10q_20140630.htm

signale que des robots sophistiqués ont pu être mal catégorisés et que les '*cyborgs*[8]', sont très difficiles à repérer et catégoriser. Au final, il est difficile de distinguer « *a bot-assisted human or human-assisted bot* », certains humains réussissant même à être reconnu comme des robots (Moon, 2017).

Le manque de contenu original ou intelligent, l'envoi en masse de mises à jour, l'abondance d'URL malicieuses ou sans lien avec le sujet du tweet, ainsi que l'agressivité (s'abonne ou se désabonne massivement de comptes) sont les principales marques de reconnaissance des robots pour Chu et al. (2012). Cela leur a permis de proposer dès 2010 une catégorisation de ces machines en quatre parties : « *1) an entropy-based component, 2) a spam detection component, 3) an account properties component, and 4) a decision maker* » *(p. 811)*

Une compétition organisée par DARPA[9] entre 6 équipes en 2015 a permis de tester et confirmer plusieurs manières de les repérer : les robots ont généralement plus d'amis (souvent d'autres robots) que de followers, ils twittent tous les jours, beaucoup, ont toujours les mêmes données GPS. Leurs tweets contiennent beaucoup de liens et sont publiés dans des différentes langues (Subrahmanian et al., 2016).

Une autre manière de séparer les tweets écrits 'par des humains' ou 'par des robots ou cyborgs' est d'examiner le champ de métadonnées « Source » du message, par exemple « Twitter for Iphone » ou « dlvr.it ». Par ce moyen, Tsou et al. (2017) ont repéré entre 29,42% (Comté de San Diego) et 53,47% (ville de Columbus) de robots ou programmes automatisés, qu'ils ont catégorisé : Jobs, Advertisement, Traffic, News and Weather. Les plateformes les plus utilisées dans leurs tweets sont Instagram, qui géolocalise les tweets par défaut, et TweetMyJOB. Les auteurs terminent leur article en listant les étapes à mener pour éviter les biais dus au bruit (spam et robots), utilisateurs (un petit nombre de twittos écrivent beaucoup) et aux erreurs du système (superposition de *Bounding.box*).

L'équipe australienne TrISMA[10] (Moon, 2017 ; Bruns et al., 2017, voir aussi 2.3.1.2.1) applique plusieurs des critères susmentionnés à une collection de 4 millions de comptes australiens et les tweets associés : la recherche dans les noms (bot, ebook, …) a permis de repérer seulement 908 comptes sur les 4 millions de comptes archivés. D'autres recherches sont en cours actuellement : repérage de similarité dans les comptes (dates de création, noms, descriptions, images de profil similaires, url dupliquées …)

Des chercheurs américains ont développé une application pour tester les comptes, soit de manière individuelle, soit groupée par API, qu'ils ont ouvert au public dès mai 2014. Il serait intéressant de vérifier si les comptes identifiés manuellement comme robot sont reconnus comme tels aussi par *BotOrNot*[11] (Davis et al., 2016).

[8] Les '*cyborgs*' peuvent être de différents types : un humain qui utilise un logiciel tel Hootsuite, pour automatiser des envois à heure fixe ou pour twitter automatiquement quand ils ajoutent du contenu sur leur blog, ou des machines qui tendent à s'humaniser.

[9] US Defense Advanced Research Projects Agency

[10] Tracking Infrastructure for Social Media Analysis, www.trisma.org

[11] Devenu Botometer : https://botometer.iuni.iu.edu/#!

2.2 Aspects méthodologiques

2.2.1 Confiance et représentativité des *API*

Morstatter et al. (2013 ; 2017) ont comparé les tweets issus de l'API Streaming de Twitter et Firehose afin de vérifier si le pourcent de tweets récupérables gratuitement pouvaient être considérés comme représentatifs de la totalité des tweets mondiaux. Ils se sont intéressés aux sujets des messages (voir également Tromble et al., 2017), mais également à leur localisation. Pour ces chercheurs, leur conclusion est que les tweets récupérés dans une *Bounding.box* représentent presque totalement les tweets géolocalisés de cet endroit (90,1% des tweets dans leur étude). Cette surreprésentation des tweets géolocalisés ne veut cependant pas dire qu'il n'y pas de biais. Un moyen de s'en assurer est de comparer les tweets récoltés avec d'autres collections, comme ceux fournis par l'API Sample[12] ou en multipliant les critères de recherche afin de rester au-dessous de la barre des 1% des tweets mondiaux. Les auteurs ont également comparé les recherches depuis deux pays, Etats-Unis et Autriche, afin de s'assurer que les résultats fournis étaient identiques. De son côté, Graham (2014) pense que les *Twittos* géolocalisés ne sont pas un échantillon représentatif des utilisateurs Twitter car ils sont « *almost certainly biased by factors such as socioeconomic status, location, and education*°». Sloan et Morgan (2015) confirme cette hypothèse en analysant les aspects sociodémographiques des tweets (voir aussi sous 2.2.2.3)

Potting (2016, p. 24) rappelle que l'exhaustivité des données est à mettre en perspective avec les questions de recherche : est-ce que tous les tweets, ou une majorité, ou sur un sujet spécifique, sont attendus ? Il prévient aussi les chercheurs travaillant sur Twitter que les messages ou les comptes analysés ne représentent qu'une toute petite partie des messages ou des comptes de Twitter, et que ceux-ci sont une tranche spécifique de la population.

2.2.2 Géolocalisation

2.2.2.1 Techniques de localisation

Les tweets sont géolocalisables depuis fin 2009, date à laquelle Twitter rachète le logiciel GeoAPI de l'entreprise Mixer Labs[13].

La géolocalisation géographique sur Twitter se fait de deux manières différentes : soit l'utilisateur active le GPS de son appareil, soit il choisit manuellement un lieu parmi ceux proposés. Ce lieu est ensuite encodé selon les coordonnées géographiques fournis par Foursquare[14].

L'action de localiser ses tweets est de la responsabilité de la personne qui twitte, généralement appelée « twittos ». Il peut choisir de le faire pour toutes ses publications, ou seulement pour certaines[15]. Pour des informations plus détaillées, voir l'Annexe 1.

[12] Echantillon aléatoire d'1% des tweets°: https://developer.twitter.com/en/products/tweets/sample

[13] https://blog.twitter.com/official/en_us/a/2009/location-location-location.html et https://www.tomsguide.fr/actualite/Tweets,31392.html

[14] https://help.twitter.com/fr/safety-and-security/tweet-location-settings

[15] De plus, cela peut dépendre de l'appareil utilisé : un même twittos peut être localisé sur un smartphone, mais pas depuis un ordinateur.

2.2.2.2 Fiabilité et autres méthodes de localisation

Il est très difficile d'estimer la part de tweets géolocalisés par rapport à la totalité des tweets produits quotidiennement, et donc leur représentativité. En Suisse, à notre connaissance, aucune étude n'a permis d'établir le taux de tweets géolocalisés, leur représentativité et les différences éventuelles entre les différentes parties linguistiques du pays.

L'estimation qu'1 à 2% des tweets seraient géolocalisés est répandue largement, Twitter lui-même indique en octobre 2017 dans un courriel à l'Agence France Presse « *que seuls 2% des tweets sont géolocalisés* » (Agence France Presse, 2017). Mais il est impossible de savoir les variations, par exemple entre pays ou par type d'utilisateur. De nombreuses études ont donc tenté de préciser ce chiffre.

Sur un échantillon de 19,6 millions de tweets récoltés en 2011, seuls 0,7% contenaient des données géographiques structurées. Les chercheurs attiraient l'attention sur la non-représentativité des tweets géolocalisés et relevaient des biais socioéconomiques, géographiques et d'éducation (Graham et al., 2014). Ils ont comparé différents champs d'information géographiques :

- la localisation dans le profil : les auteurs ont fait reconnaître la localisation déclarée sur 4'000 comptes utilisateurs avec les 4 localités de leur recherche (Le Caire, Montréal, San Diego et Tokyo). 16% des comptes n'avaient pas d'indication, la moitié a pu être vérifiée via Google ou Yahoo dans les villes correspondantes. Pour les autres localisations déclarées, un examen manuel a permis de trouver que 35,6% étaient effectivement hors des 4 villes, ce qui suggère que les twittos ne modifient pas souvent leurs données de localisation ; 24,1% avaient des données non géographiques (Neverland) ou 21,2% trop génériques (Japon, Californie, …). Les autres comptes étaient bien situés dans les zones des villes, mais avec des noms abrégés (5,8%), plusieurs localisations (4,2%) ou avec des données géographiques textuelles non reconnues par Yahoo et Google (9,1%)
- les fuseaux horaires : les comptes reliés à leurs 4 zones de recherche avaient indiqué l'horaire correspondant à la ville retenue dans moins de 70% des cas – le plus souvent en indiquant le fuseau, mais pas forcément la bonne ville, par exemple UTC+5 correspond à Montréal ou Quito (Equateur). Les auteurs se demandent si ce champ et ses indications pourraient être faussés par les applications utilisées pour twitter.

Leurs conclusions sont que ces champs sont utiles, mais les données doivent être nettoyées avant l'analyse. En les cumulant, elles peuvent être utiles à une interprétation géographique des tweets.

L'article le plus cité dans la géolocalisation de Twitter, « *Mapping the global Twitter heartbeat: The geography of Twitter* » calcule que les tweets géolocalisés représentent entre 1 et 3% de tous les tweets, mais qu'en utilisant les données figurant dans le profil ce taux pourrait être augmenté à près de 34% (Leetaru et al., 2013). Weidemann et Swift (2013), en exploitant les métadonnées, notamment la localisation, le fuseau horaire et la langue du profil, arrivent à plus de 20% de tweets localisés au niveau de la ville ou même de la rue. Minot et al (2015) ont quant à eux analysé le contenu des messages et les interactions entre les utilisateurs ; leur méthode atteint une précision de 77% dans les 10 km.

De leur côté, Kumar et al. (2017) ont étudié 2,5 millions de tweets indiens récoltés du 8 novembre 2016 au 15 janvier 2017 et comparé les données géographiques indiquées dans

les « p*lace name* » et les données GPS. Ils ont trouvé que les données géographiques transmises par les tweets issus des applications tierces (Instagram notamment) n'étaient pas exactes dans 12% des tweets indiens étudiés.

2.2.2.3 Utilisateurs

Les tweets ne sont donc pas localisés par défaut, le *Twittos* doit actionner ce paramètre. Pour quelles raisons le fait-il ? Nous n'avons pas trouvé d'études sur le sujet, mais quelques explications possibles :

- Les flux d'actualités (*timeline*) par endroit sont actifs depuis 2 ans en collaboration avec Foursquare. Lors d'événements par exemple, les twittos peuvent donc être poussés à localiser leurs messages afin de créer une communauté ou rendre leurs tweets plus visibles (Nguessan, 2016) ;
- La méconnaissance : des twittos ont été surpris[16] d'apprendre que certains de leurs tweets étaient géolocalisés. Afin de prévenir cela, Chris Weidemann, suite à l'étude citée plus haut, a créé un site[17] pour tester des comptes Twitter et voir si les 200 derniers messages contiennent des données de localisation (Gates, 2013) ;
- Pour aider en cas d'urgence : en twittant avec des informations localisées, les twittos peuvent aider à sauver des vies, selon Rajabifard et al. (2016)

D'après une étude anglaise de Sloan et Morgan (2015), les hommes géolocaliseraient légèrement plus leurs tweets que les femmes, les personnes plus âgées légèrement plus que les jeunes, et les différences seraient significatives selon la langue des comptes des twittos : les plus géolocalisés sont en turc (8,8%), indonésien (7%), portugais (5,9%), thaïlandais (5,6%) et espagnol (4,4%), tandis que les moins géolocalisés sont en coréen (0,4%), japonais (0,8%), arabe (0,9%), allemand (2%) et polonais (2,2%).

2.2.3 Identification des langues

Les utilisateurs doivent définir une langue pour leur compte, mais chaque tweet a une langue qui lui est attribuée par l'algorithme de Twitter. Or cette reconnaissance n'est pas optimale : les tweets étant limités à 140 caractères[18], les abréviations sont nombreuses, de même que les transcriptions graphiques (smiley, émoticônes), ce qui empêche le fonctionnement optimal des algorithmes habituels (Graham et al., 2014). Les twittos utilisent également des URL, mixent les langues dans un même message, ou emploient des mots argotiques, ce qui rend l'identification de la langue parfois impossible (voir quelques exemples à l'Annexe 2).

Un article sur le blog de Twitter montre la manière des informaticiens de Twitter d'analyser et améliorer les performances (@tm, 2015). Des chercheurs ont étudié ces problèmes en se focalisant sur des langues (arabe et russe (Dias Cardoso, Roy, 2016)) ou par région du monde (Graham et al., 2014). Ces derniers ont comparé une classification manuelle des langues de 4'000 tweets choisis de manière aléatoire parmi 4 grandes villes sur 3 continents avec une classification par 3 algorithmes différents, avec comme constatation finale : « *language identification of tweets is difficult for human and machine coders alike* » (p. 573). Les langues 'européennes' étaient mieux reconnues que l'arabe ou le japonais par exemple, notamment par l'algorithme de Twitter. Les auteurs terminaient en disant que le meilleur

[16] http://www.bfmtv.com/international/un-jihadiste-repere-sur-twitter-grace-a-la-geolocalisation-855388.html
[17] http://geosocialfootprint.com
[18] 280 dès novembre 2017

algorithme dépendrait des questions de recherche spécifique ainsi que de la localisation de la recherche.

2.3 Archivage des tweets

2.3.1 Définition de la nationalité d'un tweet

2.3.1.1 Suisse

La Bibliothèque nationale suisse a pour mission, d'après la loi sur la BNS et son ordonnance[19], de collecter et mettre à disposition les « Helvetica », ainsi que les « e-Helvetica » :

> *« La Bibliothèque nationale suisse (BN) a pour mandat légal de collectionner, de répertorier, de conserver et de rendre accessibles les informations imprimées ou stockées sur d'autres supports que le papier, ayant un lien avec la Suisse. Ce mandat inclut donc également les publications nées numériques ("digitally born") comme par exemple les e-books, les e-journals et les sites web. [20]»*

Selon un entretien avec madame Barbara Signori[21], responsable du service e-Helvetica, une réflexion a eu lieu sur l'archivage des réseaux sociaux, dont le contenu est vu comme important et représentatif des nouveaux modes de communication. Cependant, en plus du défi technique de l'archivage, se posent également des problèmes légaux par rapport au droit d'auteur et de protection de la personne. Pour elle, la définition serait la même que pour les Helvetica, même si la 'suissitude' d'un contenu électronique est plus difficile à estimer. Le contenu serait plus important que la forme : par exemple, un site se terminant en « .ch » n'est pas suffisant pour être considéré comme suisse, si son contenu ne se rapporte pas à ce pays.

Rauchfleisch et Metag (2016) ont analysé l'écosystème politique suisse présent dans Twitter. Les spécificités suisses (parlement semi-professionnel, multipartis, multilinguisme, forte décentralisation du pouvoir dans les cantons et communes ainsi que démocratie directe) font que les études sur les autres pays ne peuvent pas être représentatives pour la Suisse. 81 politiciens (sur 246) avaient un compte en 2013 et avaient écrit 40'026 tweets qui ont été récupérés par l'API de Twitter. Sans surprise, les premiers inscrits étaient de jeunes hommes venant des villes, majoritairement du parti Les Verts. Puis l'utilisation s'est élargie aux femmes, aux politiciens plus âgés et de tous les partis. Mais Twitter n'est pas représentatif des membres de l'Assemblée fédérale : l'UDC est sous-représentée, au contraire du Parti socialiste.

2.3.1.2 Autres définitions de tweets nationaux

Plusieurs recherches ont déjà été menées dans d'autres pays à propos de leur twittosphère, questionnant la nationalité des tweets. Voici les expériences qui nous ont semblé les plus intéressantes pour notre réflexion :

2.3.1.2.1 Australie

En Australie, le projet TrISMA[22] a vu le jour en 2009, fruit d'un partenariat entre plusieurs universités (Queensland University of Technology, Curtin University, Deakin University,

[19] Loi fédérale sur la Bibliothèque nationale suisse (LBNS; RS 432.21) et Ordonnance sur la Bibliothèque nationale suisse (OBNS, SR 432.211)
[20] https://www.nb.admin.ch/snl/fr/home/bn-professionnel/e-helvetica.html
[21] Téléphone du 15 décembre 2017
[22] Tracking Infrastructure for Social Media Analysis

Swinburne University, University of Sydney) et la bibliothèque nationale. Ce projet établit un cadre pour « *tracking, storing, and processing the public social media communication activities of Australian users at very large scale and in close to real time*°»[23]. Bruns, Burgess et Highfield (2014) voulaient trouver une méthode pour recenser tous les comptes australiens : la recherche de *Hashtags* australiens n'étant pas satisfaisante à leurs yeux, ils ont cherché d'autres méthodes :

- Se baser sur les déclarations des utilisateurs dans leurs profils n'est pas suffisamment fiable, ceux-ci pouvant déclarer ce qu'ils veulent, changer leurs déclarations, ou ne rien mettre.
- La géolocalisation ne permet d'atteindre qu'un petit nombre de comptes, comme nous l'avons déjà vu au chapitre 2.2.2.2 et les adresses IP des utilisateurs ne sont pas communiquées par Twitter
- La *timezone* déclarée par les utilisateurs est très utilisée, peut-être parce qu'elle figure en quatrième position dans l'inscription. Au vu de la géographie spécifique australienne, Twitter propose des *timezones* spécifiques pour 8 villes. Cela permet non seulement d'identifier des personnes sur sol australien, mais également une localisation relativement fine.

Cette méthode semblait la plus pertinente, mais aurait nécessité d'analyser des centaines de millions de comptes recensés par *l'API*.

Sur l'hypothèse que les personnes se rassemblent entre connaissances, les chercheurs ont donc préféré récolter les comptes australiens par effet 'boule de neige'. En partant de comptes connus australiens, ou twittant sur des sujets typiquement du pays (élections, émissions de télévision, …), les chercheurs les ont vérifiés en les confrontant à la 'time zone' déclarée par les utilisateurs, puis ils ont rapatrié leurs comptes 'followers' et 'followees'. Cette méthode leur a permis d'identifier près de 4 millions de comptes australiens en 2 ans, d'en archiver les tweets et millions d'interactions entre eux (Bruns, 2017 ; Bruns et al., 2017). Dans sa thèse critiquant les méthodes de recherche sur Twitter, Potting mentionne le manque principal : « *users who did not specify their time zone, were not connected to other Australian users, and did not participate in the selected topics were not included in the dataset.* » (2016, p. 15)

2.3.1.2.2 Autriche

Ausserhofer et Maireder (2013) ont étudié les politiciens autrichiens. Ayant constaté dans la littérature que tous les tweets ne contenaient pas forcément de hashtags ou qu'ils utilisaient une combinaison de hashtags, ils ont procédé par une étude centrée sur l'utilisateur. Afin d'identifier les Autrichiens qui tweetent activement sur la politique, ils ont établi une liste de mots (abréviations des partis politiques, noms de politiciens, sujets politiques d'actualité) qui leur a permis d'identifier des twittos. Pour des questions techniques liées aux limitations de Twitter, ils n'ont gardé que 374 utilisateurs qui ont plus de 100 followers et ont twitté sur les sujets listés. Ils ont ainsi recueilli 145'356 tweets, et les ont séparés en quatre catégories : politiciens, journalistes, experts et citoyens afin de cartographier les sujets, les échanges et les relations.

[23] https://trisma.org

2.3.1.2.3 Grande-Bretagne

La British Library archive depuis 2013 des comptes twittos précis et des tweets regroupés en collections spéciales ou recherchables par hashtag. Il ne nous a pas été possible de trouver les critères de sélection, à part le rapport au pays. (Meikle, 2013)

2.3.1.2.4 Pays-Bas

Potting présente dans sa thèse de master un projet de l'Utrecht Data School[24], qui visait en 2016 à cartographier les tweets hollandais, et particulièrement les écosystèmes des médias locaux. Pour cela, les tweets contenant un des 37'663 termes hollandais listés ont été récupérés, comparés à des comptes utilisateurs de référence, ce qui leur a permis de dire qu'avec cette liste de mots, 58,5% de tous les tweets hollandais avaient été récupérés. Mais : « *It is important to note that research on the Dutch Twittersphere is sampled within the population to those users using Twitter and tweeting in Dutch. Dutch users tweeting in languages other than Dutch, or who did not use one of the keywords, were not included in the research*. » (p. 20)

2.3.1.2.5 Italie

Une équipe de chercheurs a cerné la Twittosphère italienne en s'intéressant aux sujets traités sur Twitter. Pour cela, elle a récupéré les *Trending topics* propres à l'Italie via l'API de Twitter et les a analysés. 40% des tweets concernaient le divertissement, puis du babillage futile, suivis par la politique et le sport à égalité à 12% (Marchetti, Ceccobelli, 2016).

2.3.1.2.6 Suède

Un étudiant a cherché à repérer automatiquement le genre des comptes twittos suédois. Pour cela, il a extrait les *followers* de trois comptes connus dans le pays, puis les a trié manuellement, en enlevant les comptes d'entreprises, ou dont les tweets n'étaient pas en suédois (Matérne, 2017).

2.3.2 Archivage des tweets

2.3.2.1 Archivage complet

Le 14 avril 2010, deux partenariats étaient annoncés :

- Google intégrait les tweets à Google Replay (Replay it: Google search across the Twitter archive (Casey, 2010)), pour revivre un événement comme si on y était. En juillet 2011, le partenariat s'est terminé abruptement sans que les raisons n'en soient clairement définies (Sullivan, 2011).
- la Library of Congress (Raymond, 2010) et Twitter (Stone, 2010) annonçaient l'archivage de la totalité des tweets publiés depuis le début de la plateforme en 2006. En 2013, un autre billet de blog informait les nombreuses personnes attendant des nouvelles de la réalisation que les objectifs fixés étaient atteints : « […] *establish a secure, sustainable process for receiving and preserving a daily, ongoing stream of tweets through the present day; and to create a structure for organizing the entire archive by date*. » (Osterberg, 2013). Les buts suivants étaient alors de faire en sorte que la recherche soit possible au sein de ce corpus, afin de pouvoir l'ouvrir aux chercheurs ou étudiants. Quatre ans plus tard, en 2017, ce n'est pas encore réalisé.

[24] Le rapport du projet étant rédigé en hollandais, il n'a pas pu être utilisé directement ici.

Cette ouverture est pourtant très attendue par les communautés de chercheurs, car Twitter ne permet pas l'accès libre à ses archives. Les contenus peuvent être atteints par des recherches précises sur des *API* (*Application Programming Interface*), soit gratuitement en flux direct par *l'API* Streaming, soit par des recherches dans les *API* historiques payantes[25], soit en interrogeant Firehose, le flux intégral mais vendu très cher. Les accès sont donc plus difficiles et certains chercheurs tentent de trouver d'autres moyens pour arriver à des archives partielles (Gayo-Avello, 2016 ; Brown, 2017 ; Acker, Kriesberg, 2017)

Faute de tweets accessibles, des chercheurs se sont penchés sur les raisons qui font que la Bibliothèque du Congrès n'arrive pas à les rendre publics. Zimmer notamment, identifie deux catégories :

> *« challenges involving practice, such as how to organize the tweets, how to provide useful means of retrieval, how to physically store them; and challenges involving policy, such as the creation of access controls to the archive, whether any information should be censored or restricted, and the broader ethical considerations of the very existence of such an archive, especially privacy and user control »* (2015).

Il attire l'attention sur les problèmes qui doivent être résolus avant que les tweets ne soient mis à disposition du public :

> *« Research has shown that between 40 percent and 50 percent of tweets included information about the author, which might include contact data, other personally identifiable information, locational data, health information, and the like, posing potential privacy threats to users unaware of the fully public nature of their activity or its possible harvesting by researchers.°»*

Pour Kalev Leetaru (2017), archiver des tweets sans les sites mentionnés dans les liens revient à archiver une page html sans les images. Dans sa recherche sur 148 millions de tweets récoltés en janvier 2017, 23% des messages contenaient plus de 12 millions de liens uniques. Après échantillonnage des tweets et séparation par langue du compte utilisateur, les sites pointés par les liens étaient conservés, par Wayback Machine[26] par exemple, avec une grande variété selon la langue, les européennes étant plus archivées que les autres : de 8,6% pour le philippin à 44,8% pour le suédois.

Etant donné les sommes encaissées[27] par Twitter en vendant l'accès à ses tweets via ses API, le réseau n'est sûrement pas pressé de voir l'entier des archives accessibles gratuitement[28]. Le défi semble tout de même trop gros pour la Bibliothèque du Congrès (McGill, 2016), en décembre 2017, celle-ci annonce la fin de l'archivage complet des tweets, qui continuera seulement de manière sélective (Osterberg, 2017). Les raisons évoquées sont triples :

- la nature de Twitter qui a changé : volume et longueur des tweets en forte augmentation, tweets plus visuels alors que la LoC ne conservait que les textes ;
- les douze premières années du réseau représentent les débuts et l'essor de ce réseau ;

[25] https://developer.twitter.com/en/docs/tutorials/choosing-historical-api
[26] Archive.org
[27] En 2010, l'accès à 50% des tweets s'achetaient 360'000 $ par année : http://readwrite.com/2010/11/17/twitter_to_sell_50_of_all_tweets_for_360kyear_thro
[28] Encore étendues dès novembre 2017, avec des API Premium's : https://blog.twitter.com/developer/en_us/topics/tools/2017/introducing-twitter-premium-apis.html

- la fin de l'exception de l'exhaustivité : les tweets seront archivés comme les sites internet, selon la politique d'acquisition de la bibliothèque.

Cependant, les milliards de tweets récoltés restent encore sous embargo, jusqu'à ce que les problèmes d'accès soient résolus.

2.3.2.2 Archivages partiels

A défaut d'archivage total, certaines initiatives existent pour archiver des tweets sur un sujet, ou d'un utilisateur précis. La WayBack Machine (cf note 26) archive des comptes Twitter régulièrement, cependant à des intervalles très différents selon les comptes, et sans garantie d'exhaustivité. Voici quelques autres projets intéressants :

2.3.2.2.1 Grande-Bretagne

En 2013, The Guardian annonçait qu'un projet d'archivage de tout le web du « domaine uk » allait débuter, mené par 6 bibliothèques anglaises et irlandaise, avec pour but de préserver les enregistrements des événements culturels et intellectuels, « *copies of every public tweet and Facebook entry in the UK could eventually be included »* (Meikle, 2013). Quatre ans après, 17 millions de référence, des comptes précis ou des tweets regroupés en collections spéciales ou recherchables par *hashtag*, sont accessibles physiquement depuis les bibliothèques participantes[29]. D'autres tweets sont accessibles via leurs IDs dans le « UK Data Service » : destinées à la recherche, actuellement 22 collections[30] sont téléchargeables – les tweets originaux devant être téléchargés via les API de Twitter. (Thomson, 2017)

2.3.2.2.2 Tweets politiques

Les tweets écrits ou diffusés par des personnalités politiques sont particulièrement observés. L'élection de Donald Trump a accéléré le mouvement d'archivage partiel de ces tweets, le plus souvent par pays. Certains sites se spécialisent dans l'archivage et la mise à disposition des tweets effacés des politiciens, comme Politwoops[31], qui signale les tweets effacés par des politiciens en Europe et en Suisse. Les tweets de D. Trump sont archivés sur plusieurs sites, notamment Trump Twitter Archive[32] ou l'équivalent américain de Politwoops[33].

2.3.2.2.3 Archivage à des fins scientifiques

Les chercheurs ayant travaillé avec des tweets, géolocalisés ou non, 'archivent[34]' leur corpus de messages. Différentes motivations les guident : preuve de leurs recherches, mais également mise à disposition de données pour d'autres chercheurs afin de reproduire, comparer ou tester leurs propres corpus (Kinder-Kurlanda et al., 2017). D'après la politique de Twitter pour les développeurs, seuls peuvent être communiqués les « *Tweet IDs, Direct Message IDs, and/or User IDs* », avec un maximum fixé à 1,5 million. Depuis le 3 novembre 2017, Twitter a changé sa politique et indique :

> *« You may not distribute more than 1,500,000 Tweet IDs to any entity […] within any given 30 day period,* ***unless you are doing so on behalf of an academic institution*** *and for the sole purpose of non-commercial research […] » (Twitter Inc., 2017a, p. F2bi)* [c'est nous qui surlignons]

[29] https://beta.webarchive.org.uk
[30] https://discover.ukdataservice.ac.uk/?q=twitter
[31] https://www.politwoops.eu
[32] http://www.trumptwitterarchive.com
[33] https://projects.propublica.org/politwoop
[34] Entre guillemets, car ce n'est le plus souvent pas un archivage au sens professionnel

3. Méthodologie

3.1 Population, échantillonnage, collecte et stockage

La recherche porte sur deux populations distinctes :

- La première partie de la recherche, soit l'exploration d'un million de tweets géolocalisés en Suisse, a comme population les tweets géolocalisés publics en Suisse émis entre le 18 février et le 31 août 2017 (population 1).
- La seconde partie de la recherche, soit la définition de la notion de « tweet suisse », a comme population la totalité des tweets suisses (population 2).

L'échantillon choisi est identique pour les deux populations, mais les méthodologies de l'échantillonnage, les biais que celles-ci induisent, ainsi que leurs niveaux de confiance respectifs diffèrent considérablement.

3.1.1 Comptes humains et comptes robots

En réalisant la revue de littérature, nous avons trouvé de nombreux articles sur les moyens pour identifier et retirer les tweets automatisés envoyés par des robots. Afin d'estimer la part de robots dans notre échantillon, nous avons extrait le champ *source_keyword* et réparti les sources mentionnées entre robots et humains, en nous basant sur la littérature explorée, notamment Moon (2017) et Tsou et al. (2017). Par manque de temps, le site Botometer n'a pas pu être utilisé, de même que les calculs de distance entre deux tweets émis à quelques secondes ou minutes d'écart (Kumar et al., 2017).

Après les avoir identifiés, nous avons tout de même décidé d'analyser tout l'échantillon sans séparer les robots, tout en étant conscientes du biais possible sur certaines analyses. Le chapitre 5.1.2.4 revient plus en détail sur ces comptes robots.

3.1.2 L'échantillonnage de la population 1

L'échantillonnage de la population 1 est probabiliste. Il comprend tous les tweets disponibles gratuitement par l'*API* Streaming de Twitter ayant dans les *Champs* '*place'* une valeur correspondant à une coordonnée géographique située en Suisse, et qui portent la valeur « false » dans le champ *source.user.protected*, autrement dit dont l'utilisateur n'a pas de statut protégé et est donc supposé consentir à la divulgation de ses métadonnées (volontaire).

Dans le cas de la population 1 cette méthode d'échantillonnage est justifiée car l'échantillon ainsi constitué doit comprendre **la totalité des tweets publics** de la période de collecte dont l'utilisateur permet la géolocalisation selon l'une ou l'autre des deux modalités mises à disposition par Twitter. L'API de Twitter permet de moissonner 1% des tweets mondiaux et alerte en cas de dépassement du seuil. Comme nous n'avons pas reçu une telle alerte, nous pouvons penser que nous avons eu l'intégralité des tweets géolocalisés en Suisse durant la période concernée. Toutefois, elle comporte également deux sources potentielles de biais : la fausse attribution géographique et des trous techniques dans la collecte.

3.1.2.1 Biais dû à une fausse attribution géographique

L'API utilise la méthode suivante pour déterminer si un Tweet se trouve dans une zone de délimitation *(Bounding.box*)[35] :

[35] https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters

- Si le champ *coordinates* est rempli (*Géolocalisation fine*), ces valeurs seront testées par rapport à la *bounding box*.
- Si le champ *coordinates* est vide mais que celui de *place* est rempli (*Géolocalisation manuelle*), c'est la région définie dans ce champ qui est vérifiée pour l'intersection avec la *Bounding.box*.

Pour exclure des correspondances de lieu ou inclure uniquement des lieux qui tombent complètement dans la zone de délimitation, il est indispensable d'effectuer une étape de filtrage supplémentaire après avoir lu le flux filtré. Nous avons appliqué le filtre CH pour le champ *place.country_code* afin d'exclure les tweets qui tout en tombant dans le *Bounding.box* proviennent des pays limitrophes.

Cette méthode a l'avantage d'être simple et rapide, ce qui répond à la contrainte du temps du projet. Toutefois elle ne garantit pas un résultat 100% fiable à cause de la méthode d'alimentation des *Champs 'place'*. Puisque, en cas de *Géolocalisation fine*, ces champs sont alimentés automatiquement par Twitter selon la proximité géographique des coordonnées exactes avec une *place.id* de sa base de données, et que la méthode exacte de cette attribution n'est pas connue, il n'est pas exclu qu'un tweet émis à proximité des frontières soit faussement attribué à une *Localité* proche mais se situant de l'autre côté de ladite frontière. **Cela peut à la fois inclure dans notre corpus des tweets émis dans un pays limitrophe, et en exclure certains émis en Suisse. La contrainte de temps de notre projet n'a pas permis d'investiguer la question pour mesurer l'impact potentiel sur nos résultats.**

3.1.2.2 Biais dû à des trous techniques dans la collecte

Des problèmes techniques (panne de serveur) étant intervenus à plusieurs moments de la collecte, les tweets récoltés ne recouvrent pas les 100% potentiels. En tenant compte des minima, des maxima et des moyennes journaliers pour les mois concernés, cela représente un manque de 23'353 à 35'574 tweets, soit 2,2 à 3.3% de la collecte potentielle (Tableau 1).

Tableau 1 : Jours de collecte partielle ou manquante

Collecte partielle	Collecte manquante
Mars : 28 Mai : 11, 12, 17	 Mai : 14, 15, 16

3.1.2.3 Marge d'erreur et niveau de confiance de l'échantillonnage de la population 1

Compte tenu du nombre important de notre échantillon qui couvre presque en entier la population, la marge d'erreur reste extrêmement faible, autour de 0.02% à un niveau de confiance de 99% (estimation de fr.CheckMarket.com).

3.1.3 L'échantillonnage de la population 2 (« les tweets suisses »)

L'échantillonnage de la population 2 est une tâche délicate car cette partie de la recherche présente la particularité que sa population est constituée d'une catégorie qu'elle ambitionne précisément à définir. L'échantillonnage d'une population aux contours mouvants comporte forcément un risque très élevé de biais qu'il est dès lors nécessaire de prendre en considération.

Bien que notre échantillon soit composé des mêmes données que celui de la population 1, la méthode d'échantillonnage n'est pas identique. Celui-ci est non probabiliste, de commodité,

volontaire et au jugé. L'échantillonnage repose sur un parti-pris : la supposition qu'un lieu d'émission en Suisse puisse être considéré comme un des critères possibles de la « suissitude » des tweets. Comme l'échantillon est composé uniquement des tweets géolocalisés publics, en sont donc exclus à la fois les tweets non géolocalisés et les *Tweet privé*s.

Aux deux sources potentielles de biais partagées avec l'échantillonnage de la population 1, s'ajoute ainsi pour celui de la population 2 l'absence des tweets privés et le critère de sélection retenu, c'est à dire celle du lieu d'émission en Suisse.

3.1.3.1 Biais dû à l'absence des *Tweet privés*

Le caractère privé des tweets et des comptes protégés rend l'estimation de leur proportion par rapport à l'ensemble des tweets et des comptes extrêmement difficile. L'étude de Meeder et al. (2010) donne le chiffre de 8,4% au niveau mondial. Selon une estimation de 2011, 9,5% des twittos français protégeaient leur compte (Aminedigirep, 2011). En octobre 2012 la statistique de Twitter annonce 11,84% d'utilisateurs avec le statut « protégé » (Beevolve Inc., 2012). L'étude de Liu et al. (2014), menée sur un échantillon de 2 millions de *user.id*, démontre une baisse continue et significative de la proportion des utilisateurs protégés : autour de 15% en janvier 2007, elle est d'environ 5% en janvier 2014. Toutefois il n'est pas possible de déduire directement la proportion des tweets privés de celle des comptes privés. En effet, aucune étude n'ayant été menée sur les habitudes des utilisateurs privés, nous ne possédons pas de base établie de manière scientifique pour estimer la quantité et la fréquence de leurs émissions, sans parler de leur attitude vis-à-vis de la géolocalisation. La proportion des tweets privés qui échappent à notre collecte étant difficile à estimer, nous renonçons à la chiffrer.

3.1.3.2 Biais possible dû au critère de sélection

Le critère retenu pour constituer l'échantillon de la population 2 a été choisi en fonction d'un projet potentiel d'archivage de « tweets suisses ». Il correspond à un des critères de sélection de la collection Helvetica de la BN : le lieu d'édition en Suisse (voir aussi chapitre 5.3). Cependant, si ce critère est sans aucun doute pertinent et peut constituer une base de réflexion pour définir l'objet potentiel de l'archivage, il est également problématique.

L'échantillon constitué de tweets géolocalisés en Suisse comprend 1-2% de la totalité des tweets publics émis à l'intérieur des frontières durant la période de collecte (voir chapitre 2.2.2.2). S'il s'agissait d'un échantillonnage probabiliste cette proportion n'aurait pas d'incidence sur le niveau de confiance de la recherche grâce au nombre important de la population (53 à 107 millions de tweets uniquement pour les tweets publics). Cependant le fait même que les auteurs de 98 à 99% des tweets choisissent de ne pas géolocaliser leurs messages peut être **problématique au niveau de la représentativité**. Sans étude préalable sur l'usage des services de géolocalisation par les Suisses en général et plus particulièrement de celui de Twitter, ou plus largement sur la sensibilité des Suisses vis-à-vis de la protection des données personnelles et de la vie privée numérique, nous ne possédons pas de point de repère pour corriger les valeurs éventuellement dues à des différences de pratiques entre les cantons, entre les régions linguistiques, ou encore entre ville et campagne. **Certaines parmi celles-ci pourraient être déduites des résultats de cette recherche, mais nécessiteront des études ultérieures pour les confirmer ou infirmer.**

Un autre problème de l'échantillon est le fait que les *retweets*, ces messages reçus d'autres twittos et retransmis à ses propres *followers* en sont exclus de fait que ceux-ci ne peuvent pas être géolocalisés. En effet :

> *« Only tweets with original content can be geotagged. Retweets generated by invoking the retweet command in the Twitter user interface are not classed by Twitter as original content and are never geotagged. However, retweets generated by copying and pasting the content of a tweet into the tweet-composition box are classed as original content and can be geocoded (if the user chooses) ».* (Sloan, Morgan, 2015)

Nous l'avons confirmé parmi les tweets collectés : nous n'avons pas de retweets mentionnés comme tels. Si cela n'est pas problématique pour l'étude de la population 1 (qui ne concerne que les tweets géolocalisés), ça l'est davantage pour définir la population totale des tweets suisses.

3.1.3.3 Marge d'erreur et niveau de confiance de l'échantillonnage de la population 2

L'échantillonnage au jugé ne permet pas de faire une estimation sur le niveau de confiance de l'échantillonnage pour la population 2.

3.1.4 Collecte et stockage

Afin de récolter tous les tweets géolocalisés publics en Suisse, Bastien Berger, assistant HEG, a composé une requête dans l'API Streaming de Twitter, qui comprend les tweets géolocalisés à l'intérieur d'une zone de délimitation de forme rectangulaire (*bounding box*) entourant la Suisse et défini par les coordonnées géographiques :

- long. 5.95 ; lat. 45.81
- long. 10.50 ; lat. 47.82

Les messages ayant des coordonnées à l'intérieur de ce rectangle sont capturés de manière automatisée depuis l'API Streaming de Twitter via le plugin dédié sur logstash[36], puis stockées dans Elasticsearch. Les fichiers .json issus de ces captures sont enregistrés automatiquement dans un dossier sur le serveur Goldorak de la Haute école de gestion de Genève. Suite aux trous de récolte (voir chapitre 3.1.2.2), une récupération similaire des tweets avec leurs métadonnées a été mise en place sur un autre serveur. Ainsi, il a été possible de réindexer des données manquantes depuis ce serveur miroir vers Goldorak.

De plus, les données ont été sauvegardées sur SwitchDrive tout au long du projet (voir aussi chapitre 3.5.2).

3.2 Les champs de métadonnées

Twitter a énormément de données sur les personnes utilisant ses services, et en fournit un certain nombre[37], variable selon les tweets, via ses *API*. Après avoir pris connaissance des 88 champs de métadonnées que nous avions à disposition, nous avons décidé d'en retenir 30 (liste complète en Annexe 3) qui pouvaient nous aider à répondre aux questions de recherche de notre travail (identification du tweet et du twittos, géolocalisation, langues),

[36] https://www.elastic.co/guide/en/logstash/current/plugins-inputs-twitter.html

[37] Mais pas tout … notamment le pays d'origine ou les autres informations détectées automatiquement, comme le sexe, l'âge les lieux fréquentés, les centres d'intérêt et pour ceux utilisant une application Twitter, la liste des applications permettant de twitter.

ainsi que quelques-unes pouvant être à priori intéressantes (texte, *hashtags*, réseau ami-*follower-followee*).

3.3 Les *Localités* Twitter et les localités OFS

Twitter permet de géolocaliser les tweets au travers d'une liste d'endroits, fournie par Foursquare[38]. Afin d'associer les *place.id* avec les cantons, nous avons dû établir un fichier de concordance en effectuant les opérations suivantes :

- extraction de Kibana des métadonnées *place.id* et *place.fullname*
- regroupement des *place.id* par numéro et dédoublonnage
- fusion du fichier avec la Liste historisée des communes de la Suisse (Office fédéral de la statistique, 2017a). Récupération du numéro OFS et du canton.
- comparaison de la liste obtenue avec le Répertoire officiel des communes de Suisse de l'OFS (Office fédéral de la statistique, 2017c). Correction manuelle pour les erreurs de saisie et les communes fusionnées
- ajout de la population résidante par commune (Office fédéral de la statistique, 2017b)
- ajout des codes postaux et des coordonnées géographiques depuis le Répertoire officiel des localités de cadastre.ch (Swisstopo, 2017).

Le tout a consisté en un fichier Excel (voir extrait dans l'Annexe 4) permettant de croiser des données d'autres champs de Twitter avec la population ou le canton.

Nos tweets sont associés à 2'481 *place.id* en Suisse représentés par 3'197 *place.fullname* (il y a des doublons car les lieux sont disponibles dans plusieurs langues : Genève, Genf, Geneva, …), Les communes étaient 2'240 en 2017 en Suisse, mais ont représenté 3'388 entrées dans la liste historisée listant les fusions.

3.4 L'algorithme de reconnaissance des langues de Twitter

Les imperfections du système d'identification automatique des source.lang de Twitter sont régulièrement pointées du doigt dans la littérature scientifique (voir chapitre 2.2.3).

Le premier problème concerne le nombre des langues détectées. L'algorithme peut en identifier 60 (59 langues vivantes + 1 « indéterminée »). Or rien qu'en Europe on en dénombre entre 200 et 300, et dans le monde plus de 7'000[39]. Twitter propose donc d'identifier moins de 1% des langues vivantes actuelles.

Le second problème concerne les erreurs d'identification. La limitation de la longueur des tweets incite les *Twittos* à recourir à des abréviations, à des syntaxes tronquées, à un mélange de langues pour trouver des mots plus courts, ou encore à des émoticons. Les photos et les URL ne prêtent pas non plus à l'identification automatique des langues. Pourtant Twitter ne les classe pas systématiquement parmi les langues indéterminées (und), mais leur assigne souvent une langue.

Afin de pouvoir estimer la marge d'erreur pour notre échantillon nous avons procédé à trois tests :

[38] https://help.twitter.com/fr/safety-and-security/tweet-location-settings
[39] http://www.ethnologue.com

- Vérification manuelle sur 1'000 tweets sur une journée collectés le 31 août 2017 (TL1)
- Vérification manuelle sur 100 tweets « indonésiens » extraits de Kibana avec le filtre « l*ang.keyword* :in » (TL2)
- Vérification systématique sur les tweets d'un grand twittos humain (13'968 messages), attribués par Twitter à 26 source.lang différentes (TL3)

Nous avons utilisé les règles suivantes pour les trois tests :

- Nous avons utilisé les champs *text.keyword, lang.keyword* et *User.lang* .
- Pour identifier les langues anglais (en), français (fr), allemand (de), italien (it) et hongrois (hu) nous avons utilisé nos connaissances personnelles.
- Pour cas de doute à cause d'un message trop court rédigé dans une langue proche à une autre (p.ex. espagnol / portugais) nous avons pris pour critère de choix la *user.lang*.
- Pour les messages écrits avec des alphabets non latin nous avons considéré que la langue correspond à l'alphabet (arabe, japonais, thai, etc.)
- Pour toutes les autres nous avons utilisé Google translator.
- Les messages avec des caractéristiques suivants ont été considérés comme 'indéterminés' : 1. uniquement #, emojis, @, mot unique, noms propres, marques, noms géographiques, chiffres / plusieurs langues mélangées sans langue dominante.
- Les messages avec un mélange de langues mélangées avec une langue dominante ont été classés sous la langue dominante

3.4.1.1 Test linguistique (TL1)

L'échantillon est composé de 1'000 tweets consécutifs produits le 31 août 2017. 26 langues y sont détectées par Twitter. Sur les 1'000 tweets, 841 ont une attribution de langue correcte (84,1%), 147 une attribution fausse (14,7%) et dans le cas de 12 tweets il était impossible de déterminer avec notre méthode si l'attribution était juste ou fausse (1,2%). (Voir en Annexe 5, Tableau 8 : Langues détectées par Twitter dans l'échantillon du test (TL1).

La proportion de la fausse attribution atteint les 100% dans 7 langues, mais parmi celles-ci 6 étaient représentées par moins de 4 tweets et la dernière, le hollandais, seulement par 25. De ce fait l'échantillon n'était pas représentatif.

Parmi les langues avec plus de 50 tweets, l'anglais (16% sur 293 tweets) et l'allemand (15% sur 82 tweets) sont les moins bien détectés. Ils sont suivis par l'espagnol (10% sur 50), le français (8% sur 132), le portugais (8% sur 60), l'italien (5% sur 66) et l'indéterminé (5% sur 130). L'arabe est la seule langue de cette catégorie qui montre une marge d'erreur de 0% (sur 91 occurrences).

La marge d'erreur de la détection automatique des langues des tweets selon le (TL1) est donc d'environ 15%, valable également pour les grandes langues européennes. Pour les trois langues officielles de la Suisse ce taux se situe entre 5-15%.

3.4.1.2 Test linguistique (TL2)

L'indonésien figurant en 10^{e} langue des tweets suisses nous a fortement étonnées et nous l'avons testé, même si d'après Sloan et Morgan (2015), les Indonésiens se géolocalisent beaucoup. Nous avons pris un échantillon composé de 100 tweets ayant « in » dans le

champ *lang.keyword*. Après vérification, il a été possible d'y identifier 8 langues avec certitude parmi lesquelles non seulement des langues asiatiques, proches de l'indonésien comme le tagalog, mais également trois langues africaines, l'arabe, l'hindi, et même l'allemand et l'anglais. 24% des langues restent indéterminées par notre méthode, tandis que 8% sont soit réellement de l'indonésien, soit du tagalog (voir en Annexe 5, Figure 29 : Réattribution des langues "in").

La marge d'erreur sur l'échantillon, certes non représentatif, est donc de 92%. Il est intéressant à relever la proximité de ce résultat (8% d'attribution probable à la langue indonésienne) avec la proportion des utilisateurs ayant signalé l'Indonésie dans le champ *timezone* de leur compte par rapport à la totalité des utilisateurs de l'échantillon pour la population 1 (9,4%).

3.4.1.3 Test linguistique (TL3)

Le quatrième plus grand twittos de notre collection est un humain, dont les 13'968 tweets sont signalés en 26 langues différentes et en indéterminé. Nous avons vérifié l'entier de ses messages afin de leur attribuer la bonne langue correspondante. La majorité des 27 langues a été faussement attribuée par l'algorithme de Twitter : le twittos n'a pas écrit en basque, créole haïtien, danois, estonien, finnois, gallois, hindi, hongrois, indonésien, letton, lituanien, néerlandais, norvégien, polonais, portugais, roumain, slovène, suédois, tagalog, tchèque ni en turc, mais seulement en allemand (16 tweets au lieu de 29), anglais (420 au lieu de 550), espagnol (7 au lieu de 99), français (11'651 au lieu de 11'656), indéterminé (1'280 au lieu de 1'305) et italien (1 au lieu de 66). Les messages en français étaient généralement plus longs, ce qui explique peut-être la bonne reconnaissance de cette langue.

La marge d'erreur sur la totalité de ses messages est de 4,25%, mais en ne considérant que les 5 langues plus l'indéterminée dans lesquelles sont réellement écrits les messages, elle descend à 2,41%.

3.5 Gestion des données

3.5.1 DMP

A la suite du cahier des charges, un *DMP* (Data management plan) a été établi afin de planifier les questions de format des données, de sauvegarde et de stockage.

3.5.2 Stockage et sauvegarde

Bastien Berger nous a fourni régulièrement des fichiers csv contenant les 30 champs de métadonnées choisis, que nous avons sauvegardé sur SwitchDrive, avec nos fichiers d'analyse. Le rapport et les autres livrables étaient sur DropBox.

3.5.3 Kibana

Les tweets récoltés sont recherchables et visualisables sur un navigateur à l'aide de l'outil Kibana (vers.5.0.0), un greffon développé expressément pour visualiser les données d'ElasticSearch.

Nous l'avons utilisé pour nos analyses, en affichant les champs souhaités et en les extrayant afin de pouvoir les travailler.

3.5.4 Microsoft Excel

Si au début nous pensions pouvoir traiter les 180 fichiers csv contenant la totalité de nos données brutes dans Excel, nous avons dû y renoncer au vu de la masse des données.

Nous l'avons donc utilisé sur les extractions de Kibana afin de compléter, affiner, analyser nos données avant d'en faire des graphiques ou des tableaux croisés dynamiques.

3.6 Limites de la recherche

3.6.1 Archivage des données

Twitter a longtemps bloqué les possibilités d'archivage des tweets utilisés dans des projets de recherche (Sloan, Morgan, 2015). Comme mentionné dans la littérature (voir 0), seuls les IDs des tweets ou des comptes peuvent être archivés, avec un maximum de 1,5 million. Notre population étant moindre, nous pourrions donc archiver l'intégralité des tweets. Kinder-Kurlanda et al. (2017) ont testé et décrit comment archiver des tweets géolocalisés afin de respecter les impératifs de Twitter de même que les considérations éthiques.

3.6.2 Reproductibilité

Les tweets récoltés pour notre recherche ont été récupérés de l'API Streaming de Twitter après leur publication en ligne. Mais certains n'existent déjà plus de manière publique : tweets supprimés ou rendus privés, comptes de Twittos suspendus et donc messages indisponibles[40], … Etant donné que la communication et l'archivage des tweets ne peut se faire qu'avec leurs IDs (voir 2.3.2.2.3), ces modifications, que nous devrions reporter sur la copie des tweets enregistrés localement[41], empêchent tout futur chercheur d'accéder à notre population de tweets et donc notre recherche ne peut pas être reproduite, ni comparée à d'autres.

3.6.3 Ethique

Au vu des informations sensibles auxquelles nous avions accès, nous permettant d'identifier des personnes, voire les suivre littéralement 'à la trace', nous avons dès le début de la recherche décidé de nous interdire ces pratiques et de ne pas publier des informations permettant à d'autres personnes de le faire.

De même, nous ne sommes pas intervenues sur notre corpus de tweets, mis à part 20 messages écrits le 10 mars 2017 pour tester la (géo)-localisation et la reconnaissance des langues.

[40] Les *retweets* de messages supprimés (volontairement ou par suppression du compte) sont également supprimés automatiquement. Voir : https://help.twitter.com/fr/using-twitter/retweet-faqs

[41] «°*If content is deleted, gains protected status, or is otherwise suspended, withheld, modified, or removed from the Twitter Service (including removal of location information), you will make all reasonable efforts to delete or modify such Content (as applicable) as soon as reasonably possible, and in any case within 24 hours after a request to do so by Twitter or by a Twitter user with regard to their Content, unless otherwise prohibited by applicable law or regulation, and with the express written permission of Twitter.*°» (Twitter Inc., 2017a, p. C3)

4. Les tweets géolocalisés en Suisse

4.1 Répartition spatiale : question de recherche (QR1)

Question de recherche (QR1) : Comment peut-on exploiter les données de géolocalisation fournies par Twitter dans le contexte suisse ?

Dans notre échantillon la proportion des tweets avec *Géolocalisation fine* est de 22%, tandis que 78% des tweets ont une *Géolocalisation manuelle*. Autrement dit, à peine plus d'un cinquième de l'échantillon possède des coordonnées géographiques précises dans le champ *coordinates.coordinates*. Bien qu'il soit intéressant de rechercher les éventuelles différences et similitudes dans le comportement des *Twittos* et leurs raisons possibles pour choisir l'une ou l'autre méthode, pour des raisons de contraintes de temps nous avons renoncé à poursuivre sur cette piste. Toutes nos analyses sont donc basées sur les valeurs des *Champs* '*place'*.

Les 2'480 *place.id* présents dans notre échantillon correspondent à 2'202 entités administratives actuelles, soit à 1 pays, 26 cantons et 2'175 communes. La différence s'explique par fait que la base de données utilisée par Twitter n'est pas à jour : les fusions de communes ne se répercutent pas sur le nombre des *Localités* Twitter, qui continue de proposer pour la géolocalisation manuelle et à attribuer aux géolocalisations fines les communes qui n'existent plus de manière autonome.

Cette base de données contient également une erreur manifeste pour la Suisse : quatre *place.id* correspondent aux cantons Bâle-Ville et Bâle-Campagne. Ils se confondent encore plus dans les différentes traductions qui leur correspondent dans les champs *place.name* et *place.full.name*. L'utilisateur ne verra pas donc forcément la différence en faisant son choix lors de la géolocalisation manuelle. Ce genre d'incertitude peut donc constituer un biais pour l'analyse des données concernant Bâle-Ville et Bâle-Campagne (Tableau 2).

Tableau 2 : Les *place.id* de Bâle et leurs dénominations Twitter

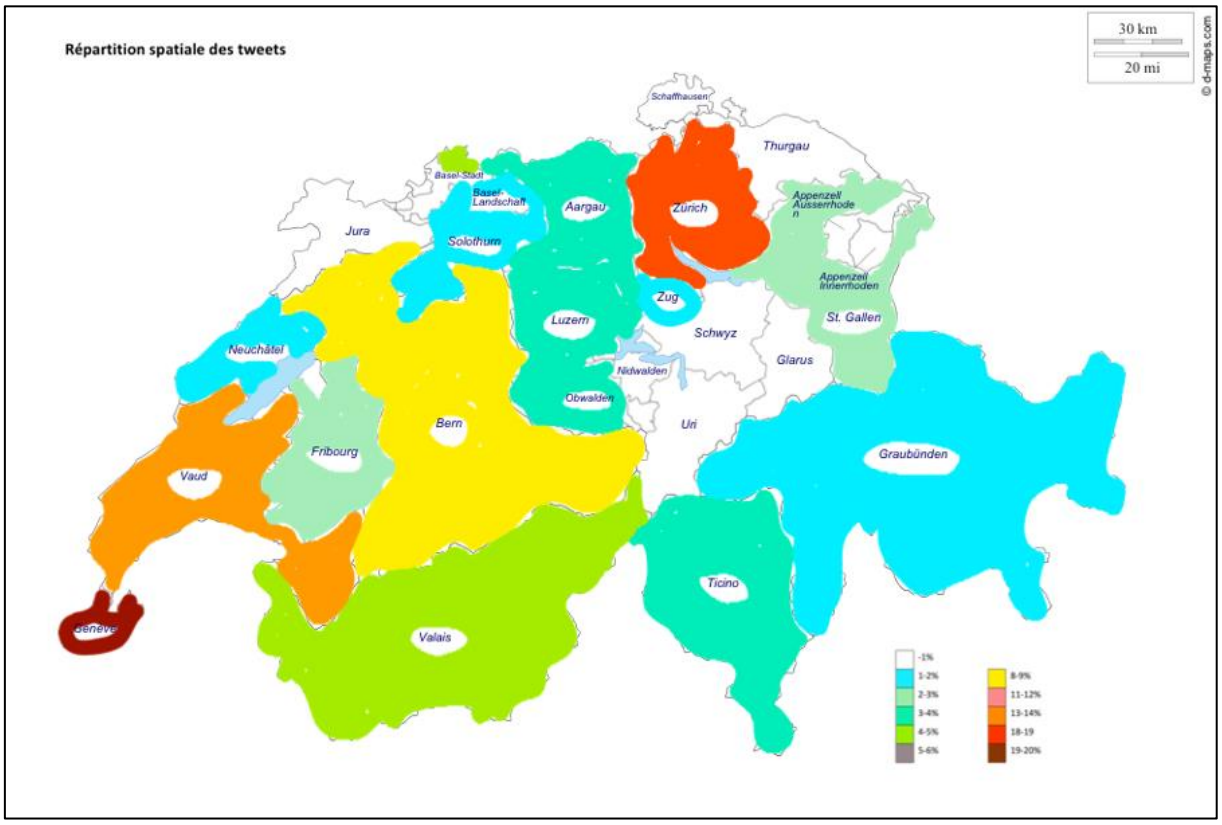
Place.id	**Intitulé Twitter (*place.full.name*)**
508f024bc856afc5	Basel-City, Switzerland (admin) Bâle, Suisse (admin)
40855e5e30f9a627	Basel-Country, Switzerland (admin) Bâle-Ville, Suisse (admin) Basileia, Suíça (admin)
bbf6c74e4f26f23d	Bâle, Suisse (city) Basileia, Suíça (city)
6ff8cdd9bab371df	Basel-Stadt (admin) Basel (admin)

Pour pouvoir exploiter les données de la géolocalisation manuelle nous avons opté pour la création et l'utilisation d'une concordance entre les *place.id* et la liste officielle des communes suisses actuelles (voir le chapitre 3.3). Nous avons analysé la répartition spatiale des tweets et des twittos uniquement au niveau des cantons. Une analyse avec une granularité plus fine sera souhaitable pour faire suite à cette étude. Pour des questions de commodité nous avons pris le parti d'inclure les tweets géolocalisés uniquement au pays (Suisse (country) : 4e7c21fd2af027c6) dans les analyses au niveau des cantons avec le sigle CH.

4.1.1 Répartition spatiale des tweets

Les 1'039'819 tweets moissonnés durant la période de collecte sont répartis sur tout le territoire suisse. Aucun canton n'est vierge de tweets géolocalisés. Toutefois la répartition est très inégale. Alors que 4 cantons donnent plus de 60% des tweets (GE, ZH, VD, BE), dans 13 cantons le nombre de tweets produits représente moins de 5% chacun (ZG, BL, SO, GR, NE, FR, SG, AG, OW, LU, TI, BS, VS), et dans 9 cantons cette proportion n'atteint pas 1% (GL, AI, AR, UR, JU, NW, SZ, SH, TG). Les tweets dont la géolocalisation comprend uniquement le pays représentent 2,18% de l'échantillon (Figure 1).

Figure 1 : Répartition des tweets par canton

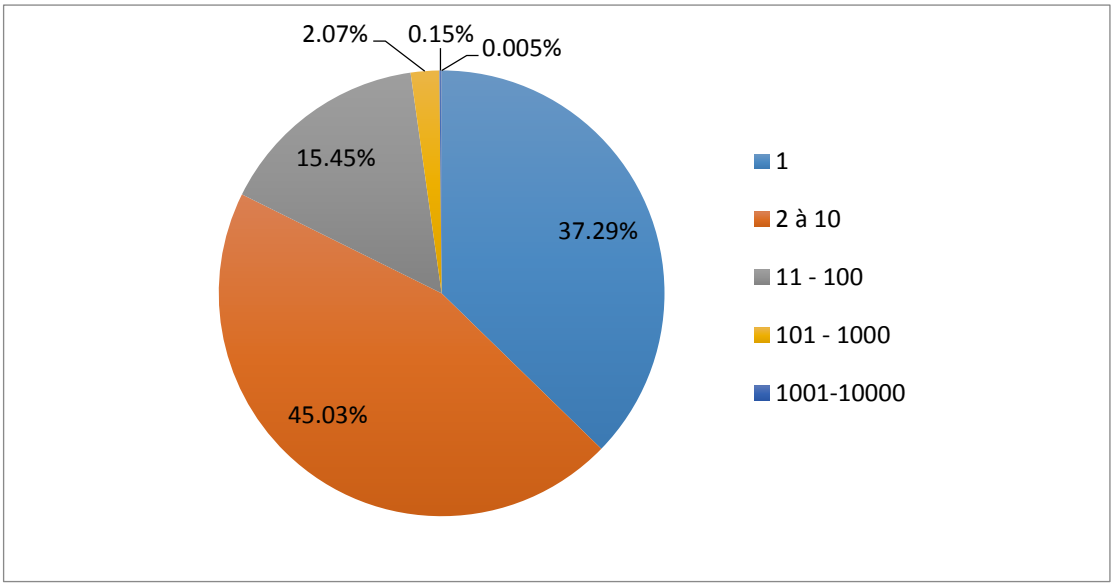


4.1.2 Répartition spatiale des comptes

65'221 twittos ont produit les 1'039'819 tweets de notre échantillon.

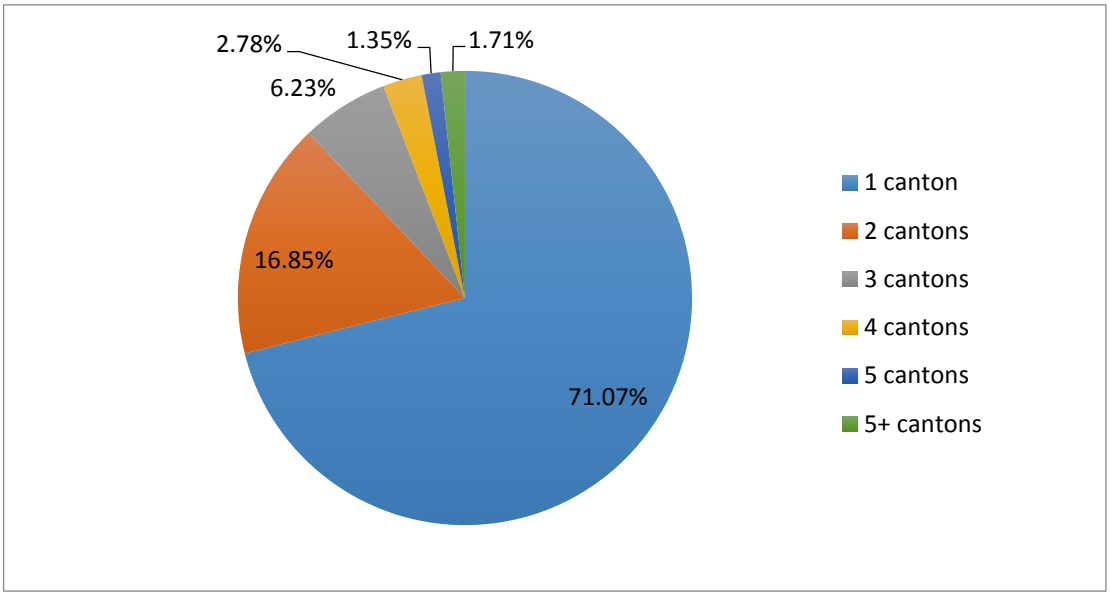
82% des comptes ont produit chacun moins de 10 tweets durant toute la période de collecte, dont près de la moitié, soit 24'320 utilisateurs, un seul tweet. (Figure 2).

Figure 2 : Proportion des comptes groupés par nombre de tweets



Seuls 29% des comptes sont actifs dans plusieurs cantons, tandis que l'activité de 71% d'entre eux se limite à un seul canton. (Figure 3).

Figure 3 : Proportion des comptes par l'étendue géographique de leur activité



La carte de répartition du nombre des comptes ayant produit au moins un tweet dans un canton ressemble beaucoup à celle des tweets. Toutefois 9 cantons montrent des proportions légèrement différentes. À Obwald, Bâle-Ville, Vaud et Genève la proportion des tweets est plus importante que la proportion des twittos : autrement dit les twittos de ces cantons produisent plus de tweets que la moyenne du pays. En Thurgovie, aux Grisons, au Tessin, en Valais et à Berne, au contraire, la proportion des tweets est moins importante que celle des twittos : dans ces cantons plus de twittos twittent donc moins (Figure 4).

Figure 4 : Répartition spatiale des twittos par canton

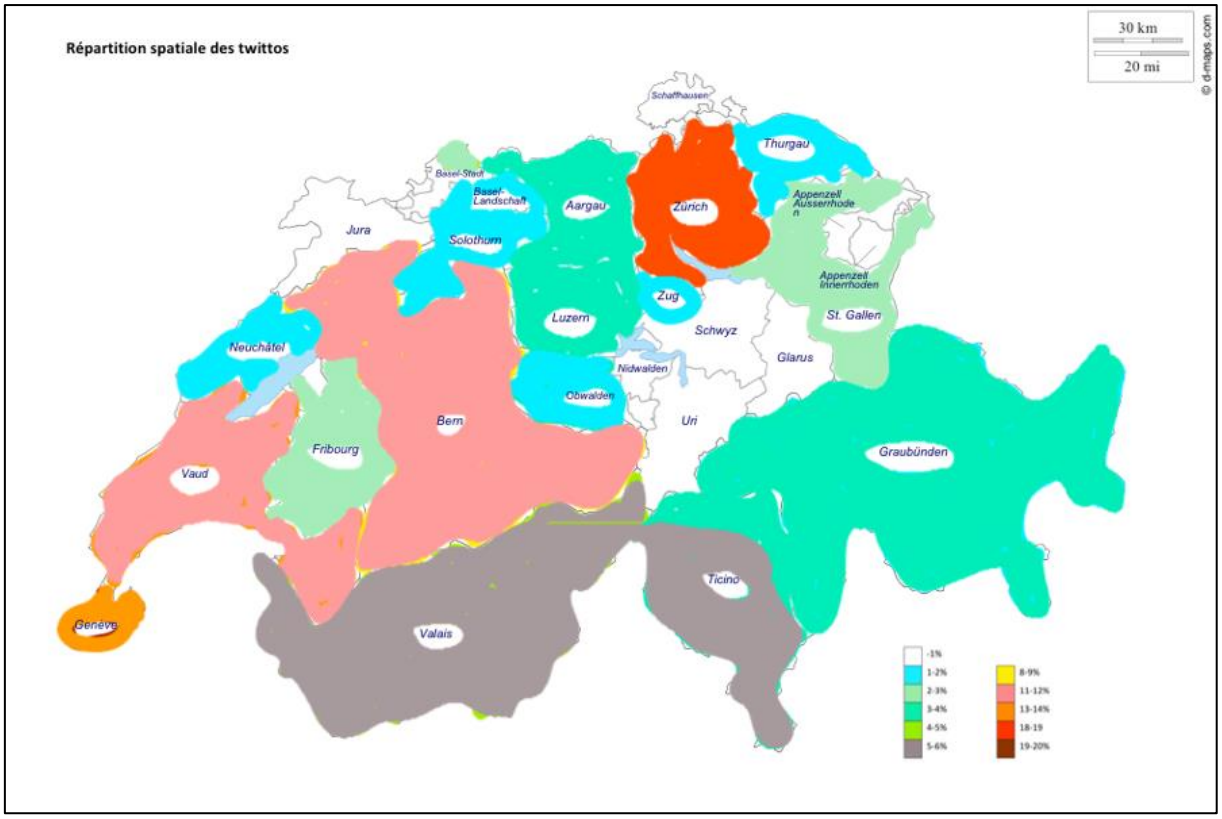
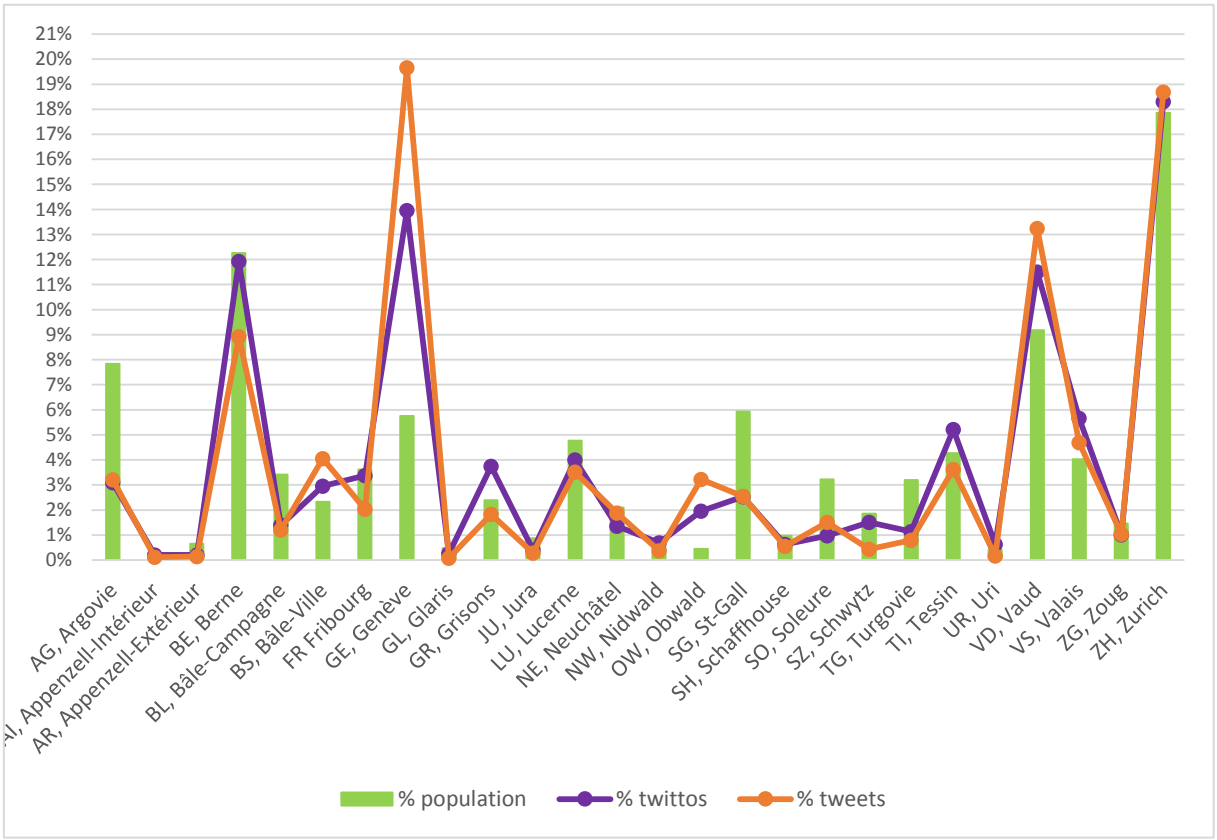


Figure 5 : Proportions de la population résidente de plus de 15 ans, des tweets et des twittos actifs par canton



Il est à noter que la répartition spatiale des twittos n'est pas proportionnelle au nombre des résidents des cantons (voir Figure 5 et Annexe 7, Tableau 9). Tandis le nombre des résidents des cantons alémaniques est de 70% de la population suisse de plus de 15 ans, seuls 57% des comptes sont actifs dans ces cantons[42]. À l'inverse, pour les 26% de la population résidente dans les cantons romands la proportion des comptes actifs s'élève à 37%. Nous avons avancé trois hypothèses pour expliquer ce fait sans avoir eu l'occasion de les tester dans le cadre de cette étude :

- Les résidents des cantons alémaniques ont moins de compte Twitter
- Les résidents des cantons alémaniques utilisent moins la géolocalisation[43]**.**
- Les cantons romands et le Tessin sont plus visités par des touristes qui twittent avec géolocalisation

Il serait toutefois intéressant de poursuivre les recherches pour les vérifier.

4.2 Répartition linguistique et spatio-linguistique : question de recherche (QR2)

Question de recherche (QR2) : Dans quelle mesure le système d'identification automatique des langues par Twitter permet-il d'obtenir une image réelle de la diversité linguistique de la Suisse ?

Les imperfections du système d'identification automatique des source.lang de Twitter, régulièrement pointées du doigt dans la littérature scientifique (voir chapitre 2.2.3), sont particulièrement problématique pour la Suisse, car le romanche, pourtant langue nationale, est absent de la liste des langues détectées. Le fait que les tweets rédigés dans cette langue échappent à 100% à l'identification automatique fausse d'emblée l'image réelle de la diversité linguistique de la Suisse examinée du point de vue de la langue de rédaction des tweets.

Le second problème concerne les erreurs d'identification. Les trois tests effectués confirment une marge d'erreur important, entre 4,25 et 15% pour les grandes langues européennes, et allant jusqu'à 92% pour l'indonésien (voir chapitre 3.4 et Annexe 5). Afin de pouvoir obtenir une image plus proche de la réalité de la diversité linguistique virtuelle des twittos de notre échantillon, nous avons pris la partie de recouper les données sur les langues productrices (source.lang) avec celles sur les langues des comptes (*User.lang*).

4.2.1 Les langues productrices (source.lang)

4.2.1.1 Répartition spatio-linguistique des tweets : Hypothèse linguistique (HSL1)

Hypothèse linguistique (HSL1) : La Suisse est un pays avec trois langues officielles reconnues par Twitter, mais la pratique virtuelle linguistique montre quatre langues principales, avec l'anglais comme deuxième langue dans tous les cantons et première langue au niveau de la Suisse.

[42] Pour respecter les contraintes temporelles du projet nous avons renoncé à entrer dans les détails des cantons bi- et trilingues.

[43] Un indice pourrait être en ce sens le fait que, la proportion des comptes germanophones de notre échantillon (15%) dépasse légèrement celle des comptes francophones (13%), tandis que celle des tweets qu'ils produisent est quasi identique (respectivement 21,2% et 20,8%). Les comptes francophones twittent donc légèrement plus avec géolocalisation que les germanophones (voir chapitre 4.2.2.1).

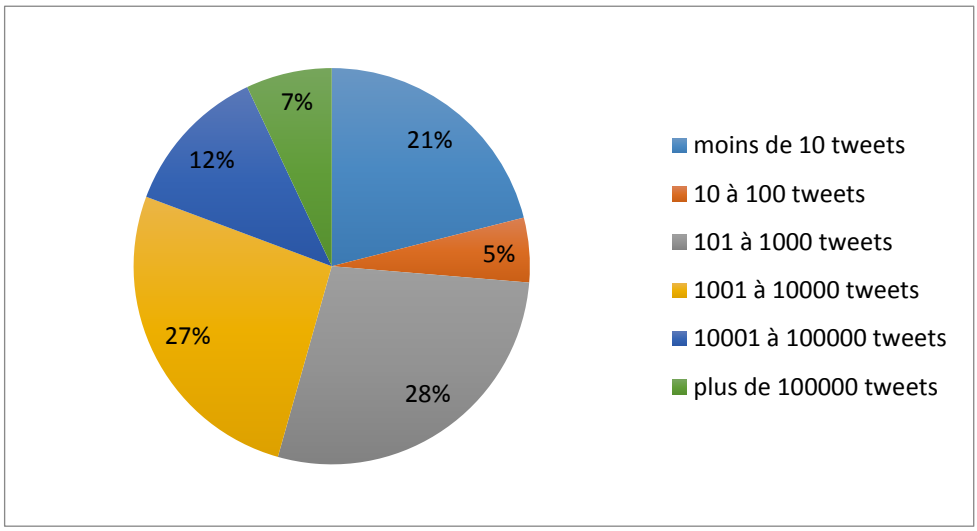
L'analyse de la répartition spatiale des langues des tweets est l'occasion de reprendre pour la Suisse l'hypothèse (H 1.1a) du projet GEoTweet : « *Genève est un canton francophone, mais la pratique virtuelle linguistique via Twitter s'approche plutôt de celle d'une région bilingue (franco-anglaise) associée à la présence d'une importante variété linguistique* ».

Pour cartographier la répartition spatio-linguistique des tweets nous avons extrait de Kibana le nombre des tweets par *source.lang* (*lang.keyword*) et par *Localité* (*place.id*), et les avons compilés par la suite par canton grâce au tableau de concordance avec les N°OFS (voir chapitre 3.3).

4.2.1.1.1 Résultats

Parmi les 60 langues détectées par Twitter comme *source.lang* 57 sont présentes dans l'échantillon, y compris la catégorie des langues indéterminées (und). Selon les valeurs obtenues par la collecte, 12 langues (21%) de notre échantillon sont utilisées chacune moins de 10 fois pour rédiger des tweets. Seules 4 langues (7%) dépassent les 100'000 tweets, parmi lesquelles le groupe des langues indéterminées (und) (Figure 6).

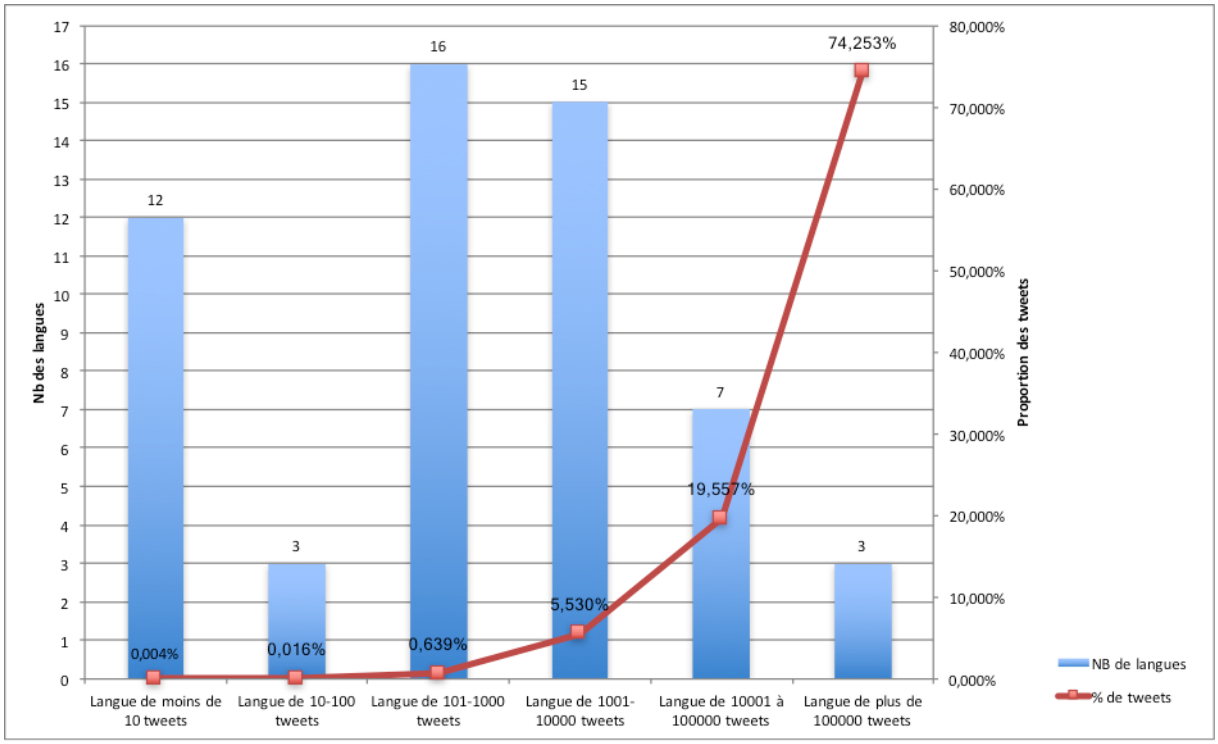
Figure 6 : Proportion des *source.lang* groupées par nombre de tweets (avec 'und')



La proportion de la catégorie (und) s'élève pour la Suisse à plus de 11% (118'524 tweets). Après l'avoir écartée, l'analyse montre que tandis que 46 langues produisent environ 6% des tweets, **les 10 langues les plus fréquentes en produisent 93,81%.**

Les dix langues les plus utilisées selon Twitter dans notre échantillon, sans prendre en compte l'indéterminée, sont l'anglais (40%), le français (18%), l'allemand (15%), l'espagnol (4,75%), le portugais (3,7%), l'italien (3,1%), l'arabe (2,8%), le turc (2,4%), le japonais (1,4%) et l'indonésien (1,3%). Le taux d'erreur sur l'identification automatique de ce dernier étant très élevé, il est toutefois raisonnable de l'écarter du palmarès (voir chapitre 3.4.1.2) Les trois premières langues totalisent près de 75% des tweets de notre échantillon (Figure 7).

Figure 7 : Pourcentage des tweets selon la fréquence de la *source.lang* (sans und)



L'hypothèse HSL1 s'est donc vérifiée en ce qui concerne la priorité de l'anglais comme *source.lang* au niveau du pays. En revanche, les prévisions de l'hypothèse ont été largement dépassées concernant la place de l'anglais au niveau des cantons. S'il est vrai que nous retrouvons une proportion élevée de l'anglais dans les trois régions linguistiques, nous ne nous attendions pas à le découvrir en première position dans 17 cantons, parmi lesquels Genève où sa proportion s'élève à plus de 50%.

Cette tendance est bien plus marquée en Suisse alémanique qu'au Tessin ou en Suisse romande, où Genève fait figure d'exception (Figure 8). Il faut descendre au niveau des secondes langues utilisées pour obtenir une carte comparable à la carte linguistique réelle de la Suisse. En effet, une langue nationale suit en deuxième position l'anglais dans tous les cantons où elle n'est pas à la première place (Figure 9).

La proportion de l'anglais est en général plus élevée dans les cantons alémaniques et augmente sensiblement dans les cantons de la Suisse centrale (Figure 10 : Proportion des tweets en anglais par canton). Pour 6 cantons (AR, BL, GL, SG, SO, TI) peu de différence sépare la première et la deuxième langue, chacune représentant environ un tiers des tweets du canton. En revanche dans les autres cantons la première langue précède d'au moins 10% la seconde. Ce sont les cantons d'Obwald et de Neuchâtel qui montrent les plus grands écarts : à Obwald 67% de tweets en anglais contre 15% en allemand, tandis qu'à Neuchâtel 65% de français contre seulement 19% d'anglais. Des différences très fortes en faveur de l'anglais face à une langue nationale sont à relever également à Zoug (en : 52% / de :22%), à Uri (en : 52.5% / de : 22.5%), à Schaffhouse (en : 51.5% / de : 26%), à Genève (en : 52% / fr : 31%), à Bâle-Ville (en : 48% / de : 21,5%) et même au Grison (en : 45% / de : 27%). Une des langues nationales est au contraire largement favorisée face à l'anglais en Appenzell Rhodes-Extérieures (de : 48% / en : 26%), à Fribourg (fr : 42,5% / en : 32.5%), au Jura (fr : 45% / en : 29%), dans le canton de Vaud (fr : 45% / en : 30%) et en Valais (fr : 45% / en : 35%).

Figure 8 : Première langue des tweets par canton

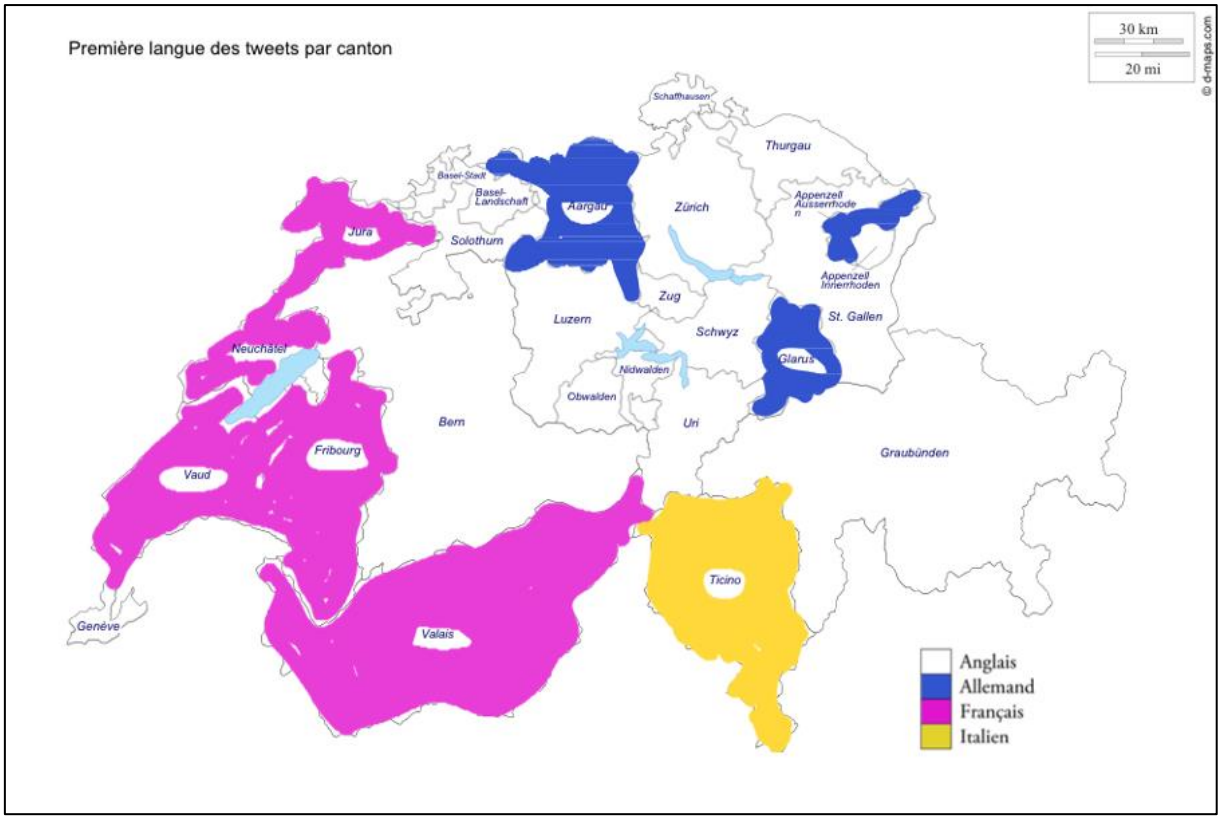


Figure 9 : Première langue nationale comme *source.lang* par canton

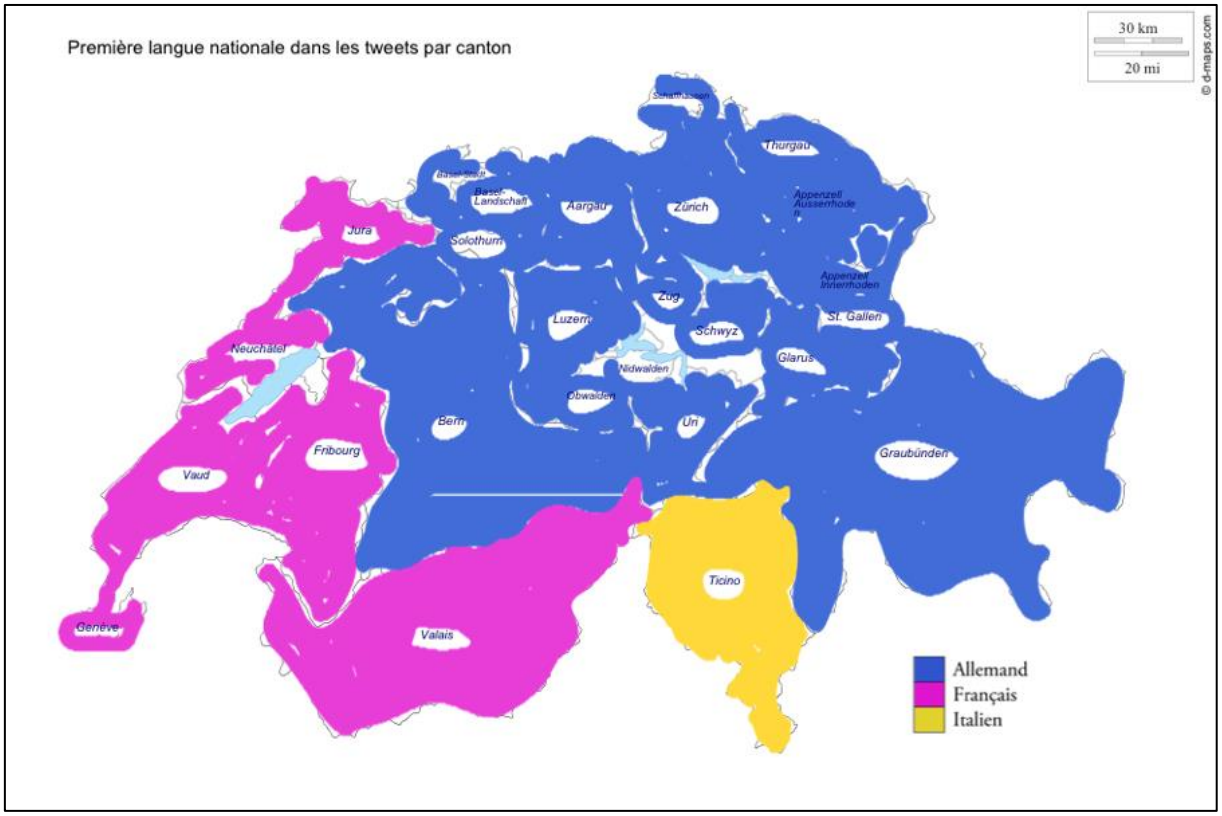
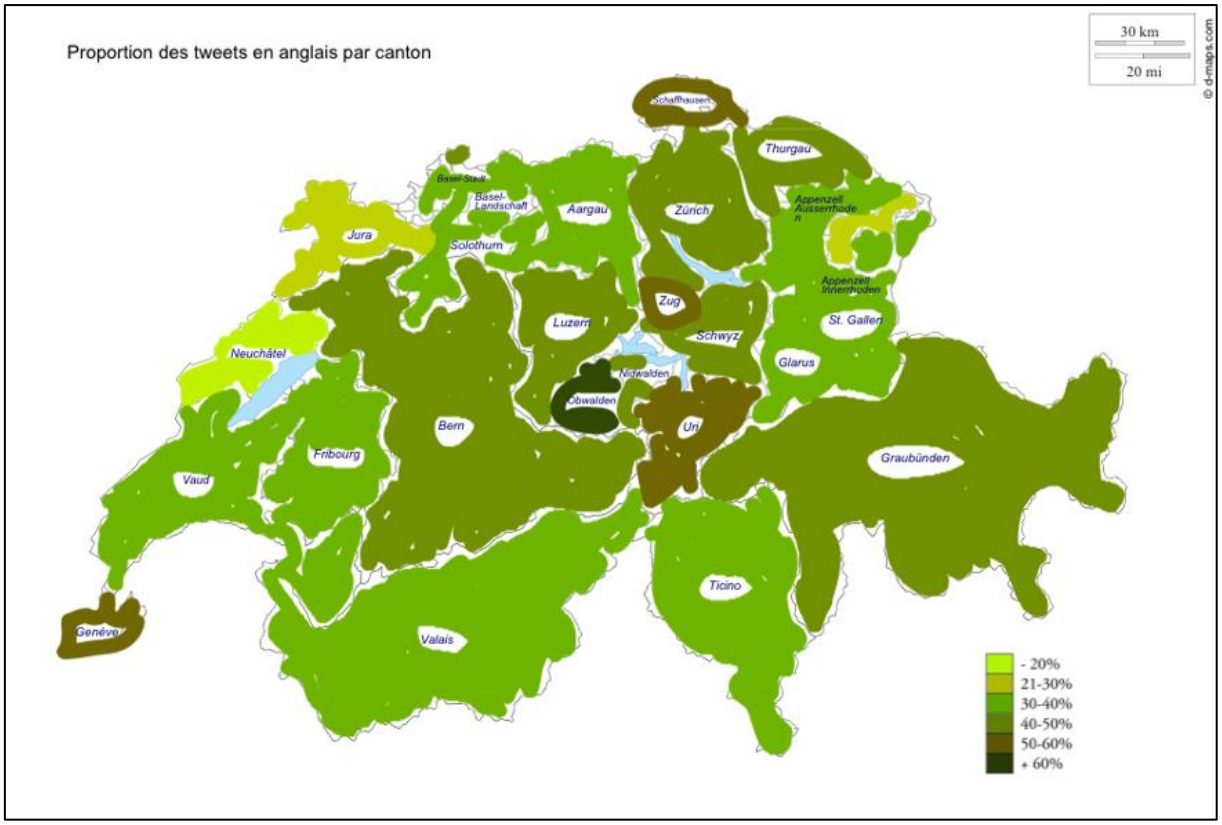


Figure 10 : Proportion des tweets en anglais par canton



4.2.2 source.lang et *User.lang* : Hypothèse linguistique (HSL2)

Pour expliquer cette proportion très élevée de l'anglais comme *source.lang* dans les cantons alémaniques nous avons décidé de faire appel au contenu du champ *user.lang*, c'est-à-dire la langue des comptes déclarée par l'utilisateur. Nous avons formulé l'hypothèse suivante :

Hypothèse linguistique (HSL2) : La proportion plus élevée de l'anglais comme langue productrice dans les cantons alémaniques s'explique en partie par une propension plus élevée des germanophones à twitter en anglais par rapport aux Suisses francophones et italophones.

4.2.2.1 Résultats

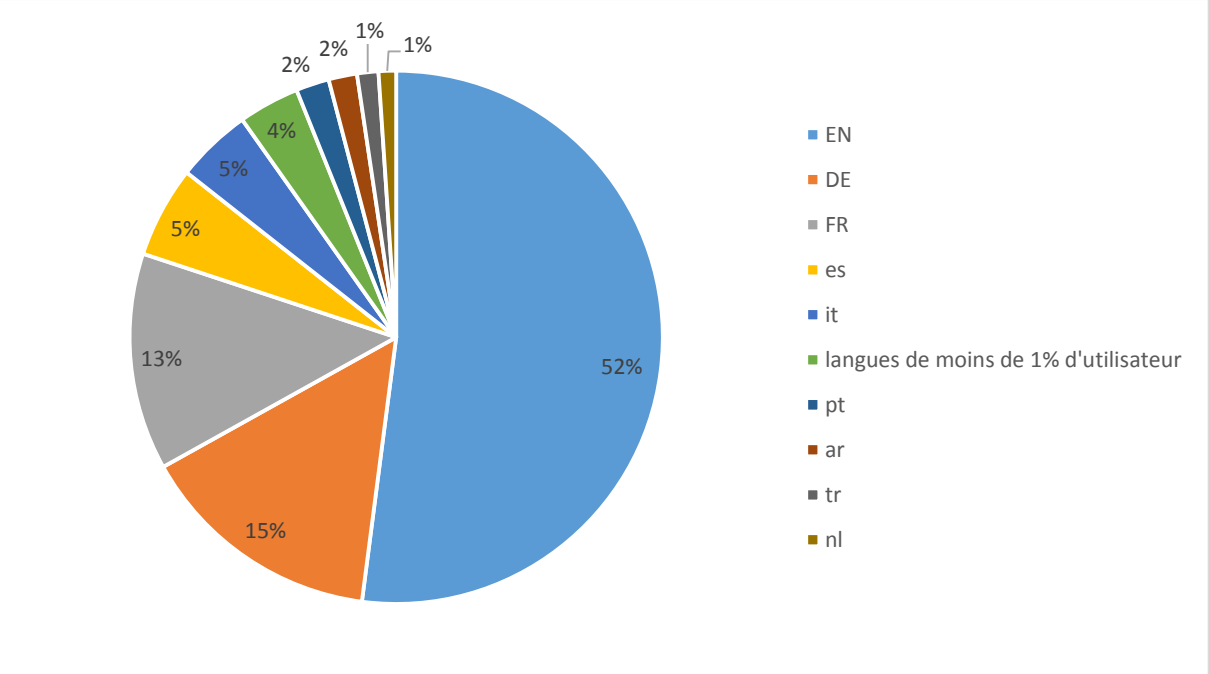
Le champ *user.lang* est plus libre que celui de *lang.keyword* qui contient la *source.lang.* L'utilisateur peut choisir la langue de son compte qui sera codée par Twitter selon la norme d'ISO 639-1 à deux lettres ou selon celle ISO 639-2 à trois lettres. Cela permet, ou du moins a permis à un certain moment de l'histoire de Twitter, plus de précision que la *source.lang* : un utilisateur suisse aurait pu opter pour gsw ou de-CH pour le suisse allemand, ou encore pour fr-CH au lieu de fr pour le français[44]. Pourtant on constate d'emblée que peu d'utilisateurs l'ont choisi pour la Suisse. Dans notre échantillon ces trois langues n'ont été définies que par un seul utilisateur chacune. Pour les besoins de nos analyses et pour simplifier les comparaisons, nous avons considéré comme une seule *user.lang* toutes les variantes d'une langue vivante. Ainsi (en, en-GB, en-AU) = anglais, (de, de-CH, gsw) =, allemand, (fr, fr-CH, fr-CA) =français.

[44] Cela n'est plus valable aujourd'hui : le twittos ne peut choisir que parmi 48 langues. Les comptes concernés dans notre étude ont été créés entre janvier et mai 2015. Nous n'avons pas trouvé plus d'informations à ce sujet

Le nombre des différentes *user.lang* de notre échantillon est identique à celui des langues productrices détectées par Twitter : 57. Toutefois il ne s'agit que d'une coïncidence, car les deux listes ne se recouvrent que partiellement (voir Annexe 6).

Dans l'ensemble de notre échantillon plus de la moitié des comptes (52%) sont déclarés en anglais, tandis que **l'allemand, le français et l'italien totalisent environ un tiers des comptes (33%)**. Aucun compte n'est déclaré en romanche, mais nous ne savons pas si cela a été possible par le passé. La proportion des comptes des cinq premières *user.lang* est de 90,5%. Ceux-ci produisent 90,89% des tweets (Figure 11).

Figure 11 : Proportion des comptes par *user.lang*



Au niveau suisse la majorité des comptes sont donc en anglais et les comptes affichant les trois langues officielles en *user.lang* sont en minorité. Ce constat se vérifie également au niveau des cantons. L'allemand n'est la première *user.lang* que dans quatre cantons (AG, AR, SO, TG), le français dans deux (JU, NE), et l'italien nulle part (Figure 12). **La première cause de la proportion très élevée de l'anglais comme langue productrice au niveau de la Suisse et au niveau des cantons est donc sans doute la proportion très élevée des comptes anglophones**. Cependant cela n'explique pas pourquoi l'anglais n'est pas la première langue des tweets dans la majorité des cantons francophones et italophones. D'autant plus que les comptes anglophones produisent moins de tweets par rapport à leur proportion que les comptes en allemand ou en français. Les comptes affichant l'anglais en *user.lang* produisent seulement 36,9% des tweets au niveau de la Suisse et les comptes en italien 2,7%, ce qui est bien inférieur à leurs proportions respectives parmi les twittos. **Les comptes en allemand et en français semblent plus actifs, la proportion des tweets qu'ils produisent (respectivement 20,70% et 21,19%) dépassant largement leur proportion parmi les comptes (15% et 13%) (Figure 13).**

Figure 12 : Première *user.lang* par canton

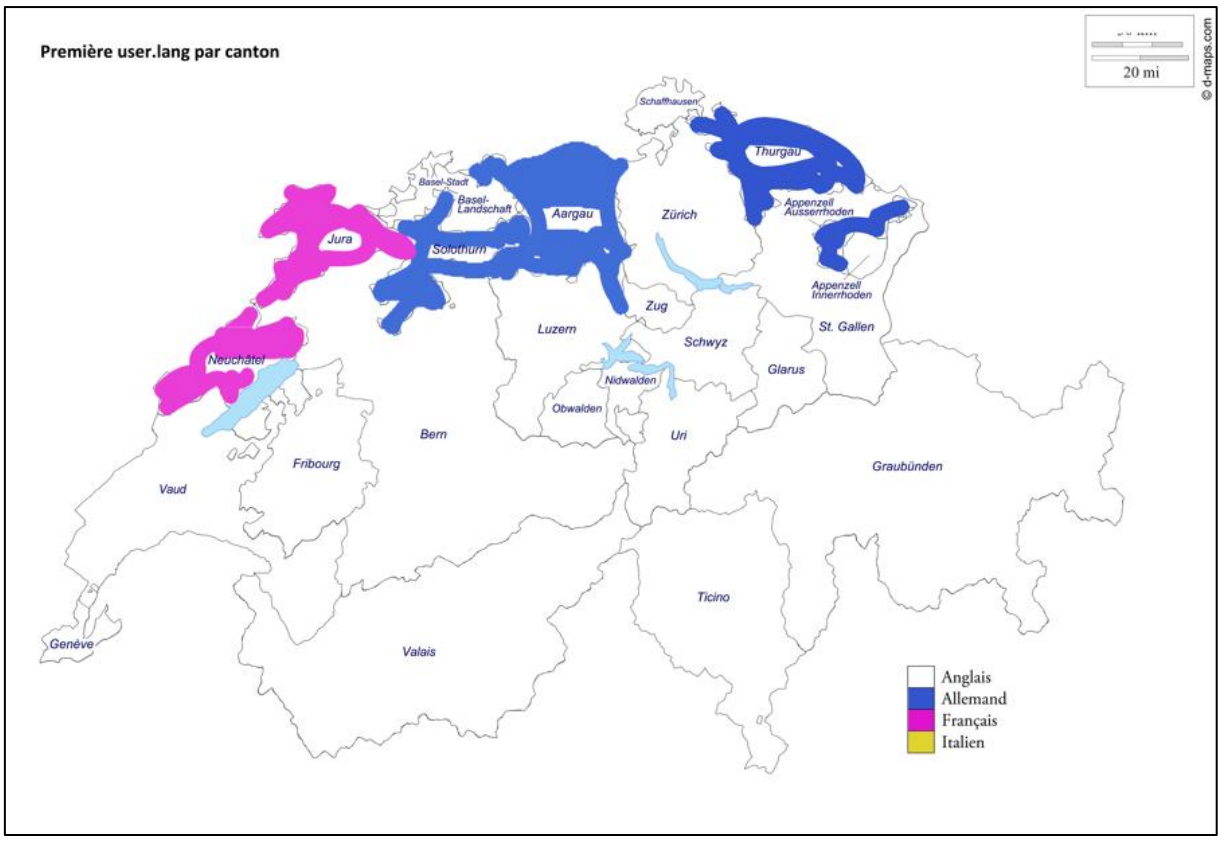
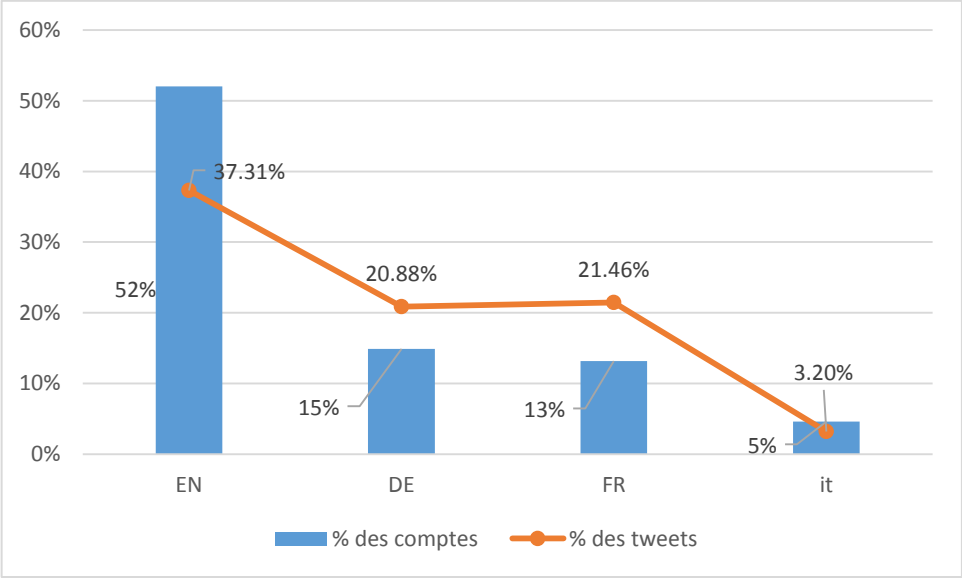


Figure 13 : Proportion des tweets classés par *user.lang* par rapport à la proportion des comptes



Le nombre des source.lang utilisées par les comptes pour rédiger les messages est plutôt élevé quelle que soit la *User.lang* déclarée. En ce qui concerne l'anglais et les trois langues nationales, 35 langues différentes détectées dans les tweets produits par les comptes « en italien », 42 dans ceux des comptes « en français », 45 pour les comptes « en allemand » et 56 pour les comptes « en anglais ».

En examinant les préférences linguistiques des comptes germanophones, francophones et italophones, on constate que la proportion des comptes qui twittent en anglais parmi les

comptes affichant une des trois langues nationales en *user.lang* est d'environ 45%. Toutefois si chez les comptes francophones elle est seulement de 43%, chez les germanophones plus de la moitié des comptes sont concernés (53%). De plus, non seulement une plus grande proportion des comptes germanophones et italophones twittent occasionnellement ou uniquement en anglais, mais ils produisent également une proportion plus élevée de leurs tweets en anglais. Tandis que les comptes en français ne produisent que 16% de leurs tweets en anglais, les italophones sont à 21% et les germanophones à 26%. **Plus d'un quart de la production des comptes germanophones est donc en anglais** (Figure 14). En comparant les proportions des différentes source.lang des tweets produits par des comptes affichant la même *user.lang*, on constate que plus la proportion de l'anglais est élevée parmi les tweets, plus la proportion des tweets produits dans la langue identique à celle du compte diminue. **Ainsi les comptes francophones produisent 65% de leurs tweets en français, tandis que les germanophones moins de 50% en allemand (**Figure 15**).**

Autrement dit, les comptes germanophones twittent plus volontiers en anglais que les comptes francophones, et en même temps moins volontiers en allemand que les comptes francophones en français. L'hypothèse (HSL2) s'est donc vérifiée sur notre échantillon.

Figure 14 : Proportion des tweets en anglais classés par *user.lang* comparée à la proportion des comptes qui twittent en anglais

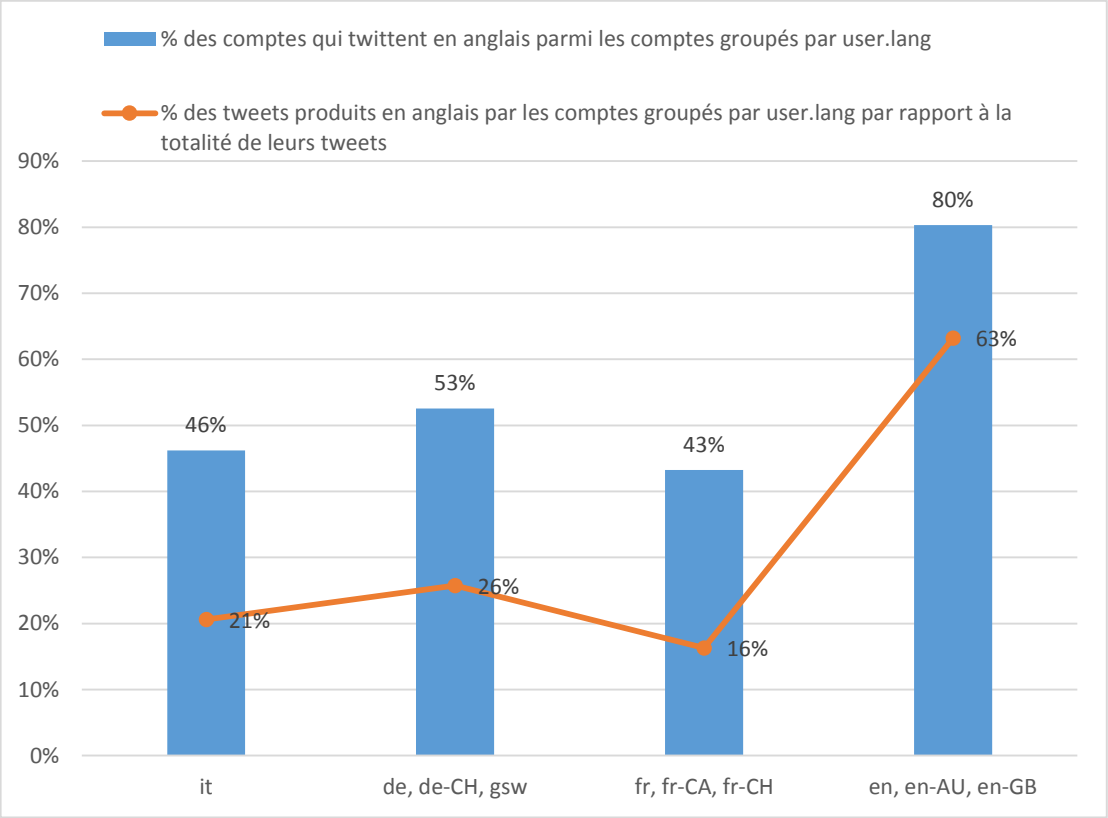
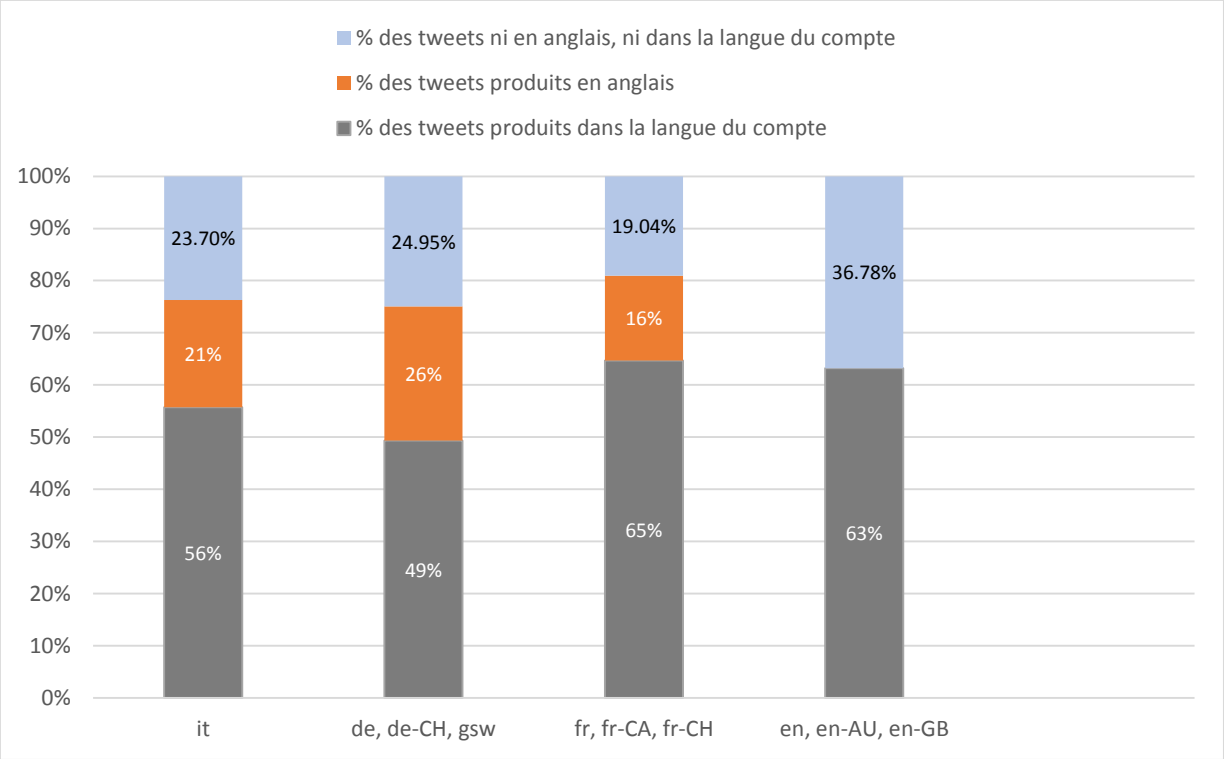


Figure 15 : Proportion des *source.lang* utilisées par les comptes classés par *user.lang*



4.3 Répartition temporelle, spatio-temporelle et linguistique

Hypothèse linguistico-temporelle (HLT1) : La quantité des tweets ainsi que les communautés linguistiques virtuelles sur Twitter évoluent différemment selon les mois.

L'analyse de la répartition temporelle des langues des tweets est l'occasion de reprendre pour la Suisse l'hypothèse H 1.1b du projet GEoTweet : « La quantité de tweets varie en fonction des heures de la journée et des jours de la semaine. La variation des communautés linguistiques virtuelles sur Twitter évolue différemment selon les heures de la journée, les jours de la semaine et les dates ».

Pour le vérifier nous avons effectué des analyses croisées sur les *source.lang* et sur les *user.lang* de notre échantillon.

4.3.1 Analyses et résultats pour l'hypothèse (HLT1)

4.3.1.1 Répartition temporelle et linguistico-temporelle des tweets

4.3.1.1.1 Méthode

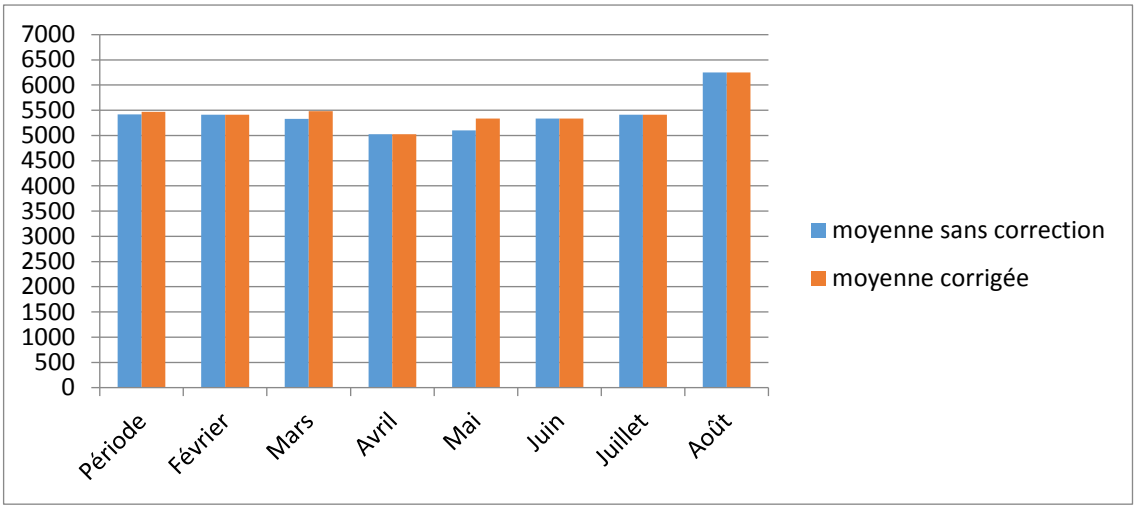
L'analyse de la répartition mensuelle du nombre des tweets est problématique en raison des jours de collecte manquants ou partiels. Une comparaison entre les résultats sur les valeurs réelles et les valeurs corrigées le montre très clairement (voir les tableaux en Annexe 8, Tableau 10 et Tableau 11). Si les médianes changent très peu, les moyennes des mois de mars et de mai touchées par les lacunes augmentent considérablement après correction.

4.3.1.1.2 Répartition mensuelle du nombre des tweets

La moyenne d'avril est la plus basse dans notre échantillon avant comme après correction. La différence entre sa moyenne et celle du mois d'août, le plus fructueux pour la collecte,

s'élève à 24,8%. Après correction c'est le mois de mars qui s'avère être le second plus riche en tweets géolocalisés en moyenne, suivi de près du mois de février qui est quasi à l'égalité avec le mois de juillet. Mai et juin occupent, également à l'égalité, l'avant dernière place (Figure 16).

Figure 16 : Répartition mensuelle des tweets (moyennes)

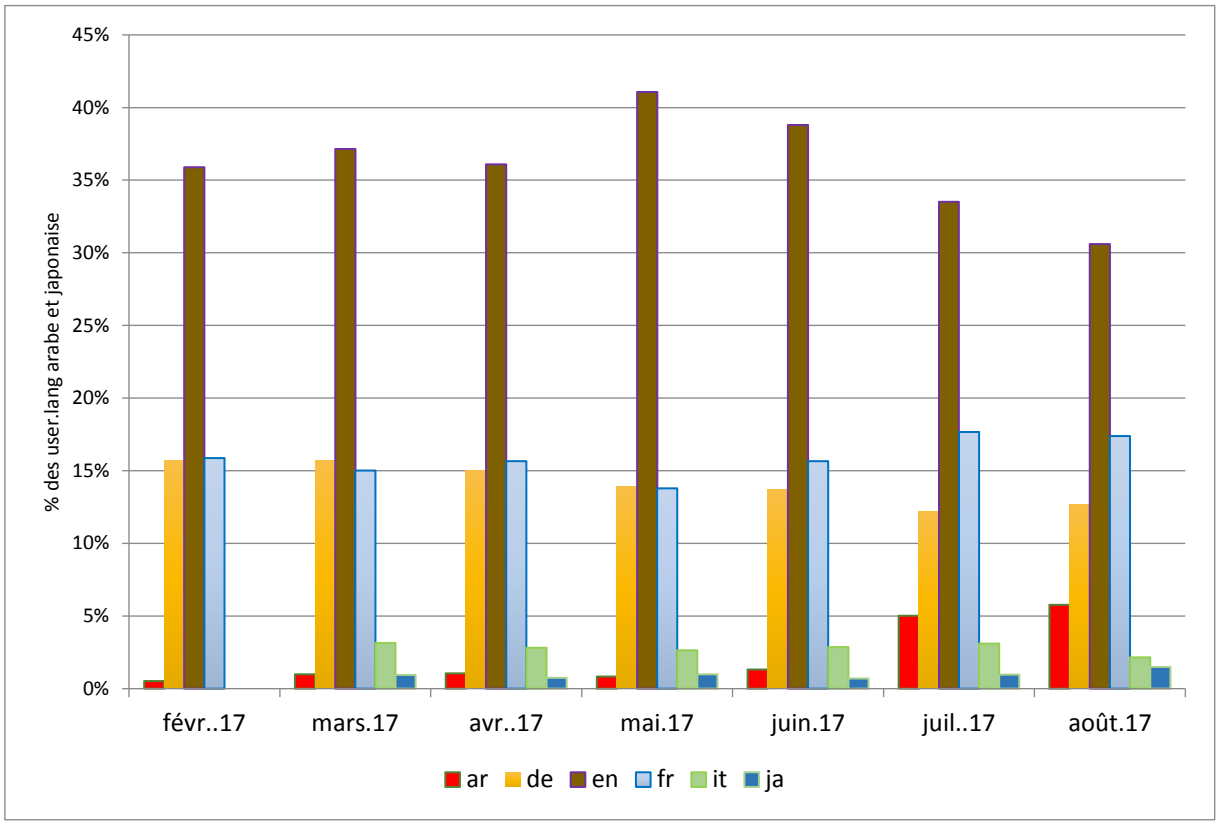


4.3.1.1.3 Répartition mensuelle des langues des tweets (source.lang)

Le nombre des *source.lang* des tweets varie peu par mois : 44 *source.lang* en février, 46 en août, 47 en avril et en juillet, 48 en mars et en mai, et enfin 49 en juin.

On constate en revanche d'importantes variations dans la répartition mensuelle des *source.lang*, et notamment dans leurs proportions respectives parmi celles du mois. Tandis que l'anglais est le plus présent dans les tweets du mois de mai (41%) et moins présent au mois d'août (30%), et la proportion de l'allemand et de l'italien sont est en baisse continue de février à août (de 15,87% à 12,67% pour l'allemand, de 3,13% à 2,15% pour l'italien), le français est à son plus haut niveau précisément durant les mois d'été. **Deux langues typiques des touristes, l'arabe et le japonais montrent en revanche des progressions spectaculaires : de 0.5% en février la proportion de l'arabe monte à 5,7% au mois d'août, et le japonais, pas du tout présent en février, culmine à 1,5% en août (**Figure 17**)**.

Figure 17 : Répartition mensuelle des proportions de six *source.lang (tweets)*

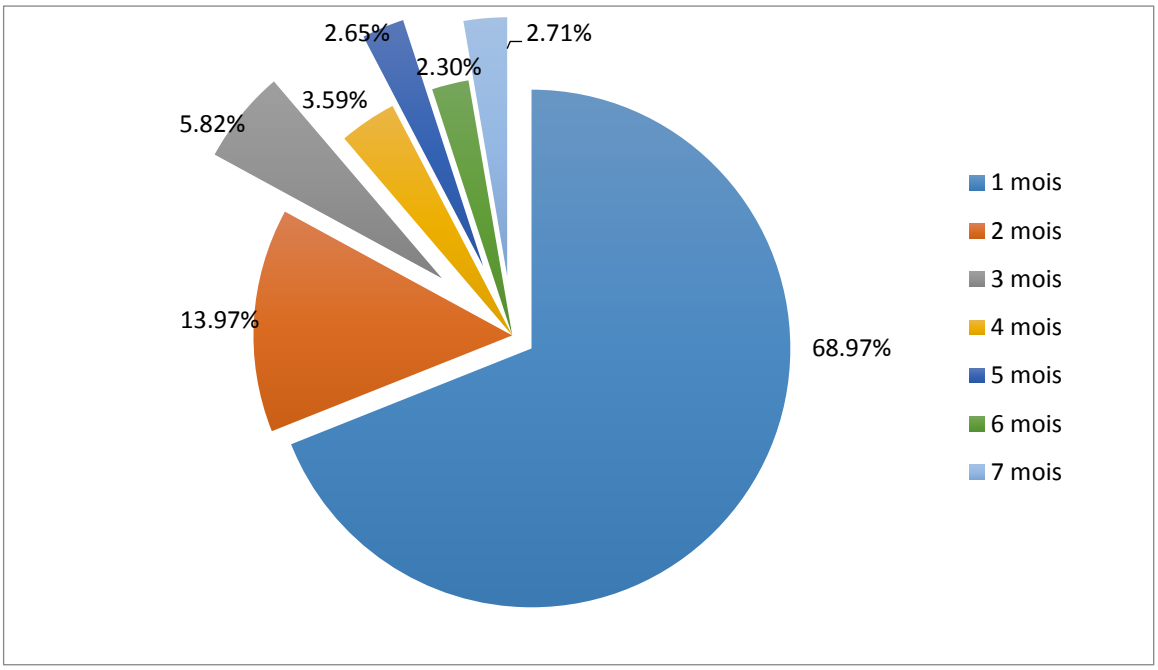


4.3.1.2 Répartition mensuelle des comptes et des *User.lang*

4.3.1.2.1 Méthode

L'analyse de la répartition mensuelle du nombre des *Utilisateur actif*s est problématique en raison des jours de collecte manquants ou partiels. La correction, bien que très délicate, est toutefois nécessaire en raison de la proportion très élevée des comptes dont la période d'activité ne dépasse pas un mois. En effet, **la *Période ou durée d'activité* de plus de 2/3 des twittos, 68,97% (44'980 comptes), se limite à un mois civil**. À l'inverse, parmi les 65'221 utilisateurs uniques seuls 2,71% (1'767 comptes) ont twitté tout au long de la période de collecte et montrent donc une durée d'activité de 7 mois. (Figure 18). Cette proportion très élevée des utilisateurs qui ne sont actifs que durant une brève période laisse penser que les lacunes techniques de la collecte ont également fortement diminué le nombre des comptes moissonnés. Cette proportion élevée des utilisateurs dont la période d'activité se limite à une période courte indique un comportement où le choix de twitter ou celui de géolocaliser ses tweets en Suisse dépend de facteurs liés à des contraintes ou opportunités temporels. Malgré l'incertitude de cette méthode, et **uniquement pour les besoins de l'analyse du nombre mensuel des utilisateurs actifs**, nous avons donc décidé de « combler » non seulement les jours de collecte lacunaires, mais également le mois de février en y ajoutant la moyenne journalière du nombre des twittos sur 17 jours. **Toutes les autres analyses concernant la qualité des comptes (*user.lang*, *Période ou durée d'activité*, etc.) sont effectuées sur les valeurs réelles**.

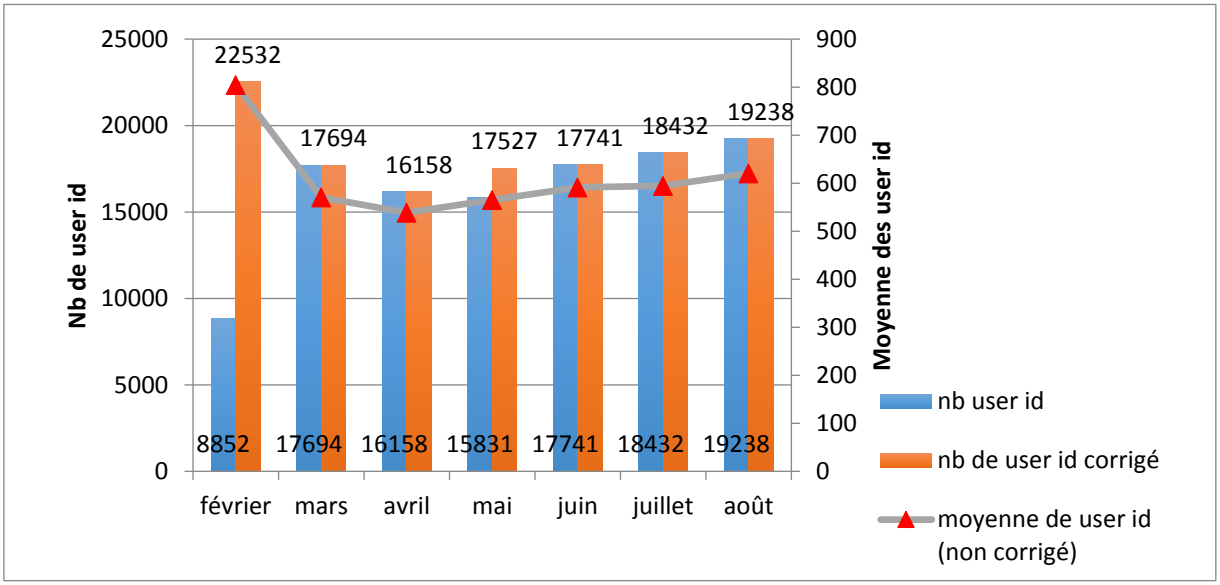
Figure 18 : Proportion des utilisateurs selon la durée de leur période d'activité



4.3.1.2.2 Nombre d'utilisateurs actifs par mois (valeurs corrigées)

Une comparaison entre les résultats sur les valeurs réelles et les valeurs corrigées du nombre mensuel des utilisateurs actifs montre des différences très nettes (Figure 19). Tandis qu'avant correction c'est le mois d'août qui totalise le plus grand nombre d'utilisateurs (19'238), après correction il n'occupe que la seconde place après le mois de février (22'532). Ils sont suivis du mois de juillet (18'142), puis par juin (17'741), mars (17'694) et mai (17'527). Le mois d'avril est le dernier avec 16'158 utilisateurs actifs.

Figure 19 : Répartition du nombre des utilisateurs actifs par mois, avant et après correction, comparée à la moyenne des twittos



4.3.1.2.3 Les twittos à une durée d'activité d'un mois et leur proportion selon le nombre de leurs tweets (valeurs réelles)

Le mois qui totalise la plus grande proportion d'utilisateurs à une *Période ou durée d'activité* d'un mois est celui d'août (49%), suivi à quasi égalité par juillet (42%) et mars (41%). En juin,

mai et avril leur proportion descend en dessous de 40%, et en février en dessous de 30% (Figure 20). Il est cependant probable que les résultats pour le mois de février soient pénalisés par le période de collecte plus courte (11 jours du 18 au 28). Plus de la moitié (54%) des utilisateurs à une durée maximale d'activité d'un mois twittent une seule fois. Cela représente 37% de la totalité des comptes, soit 24'320 utilisateurs uniques. 40% des utilisateurs d'une durée d'activité d'un mois, soit 17'645 comptes, twittent entre 2 et 9 fois. Ces deux catégories couvrent donc 64% de la totalité des comptes et même 93,3% de ceux à une période d'activité d'un mois. Les utilisateurs qui twittent plus de dix fois constituent une petite minorité : 2'912 comptes (6.5%) twittent 10 à 99 fois, et 104 comptes (0.2%) plus de 100 fois pendant ce court laps de temps. Les analyses montrent une augmentation très nette de leur proportion par rapport à la totalité des utilisateurs du mois d'août (5,13%) et dans une moindre mesure du mois de juillet (3,01%) (Figure 21).

Figure 20 : Proportion des utilisateurs à une période d'activité d'un mois (répartition mensuelle)

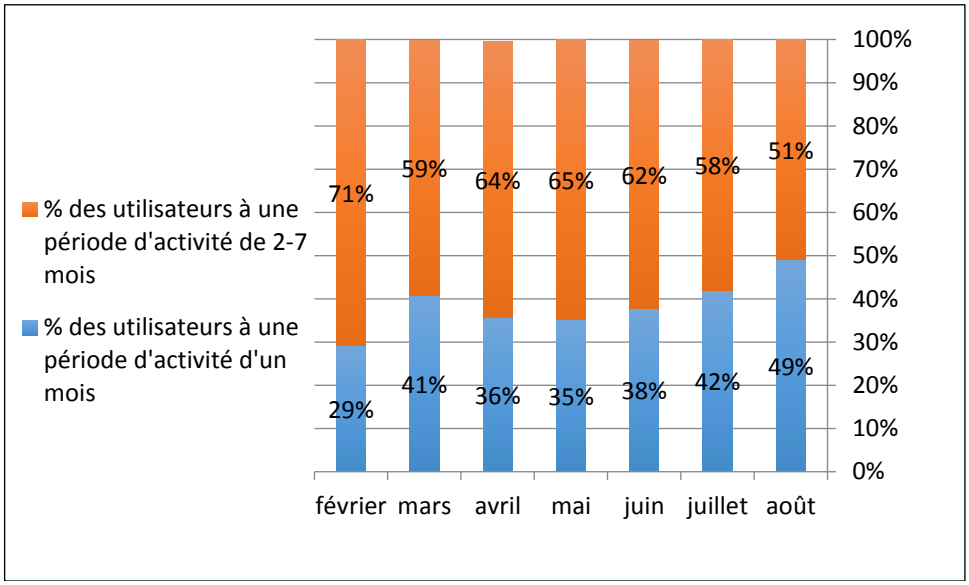
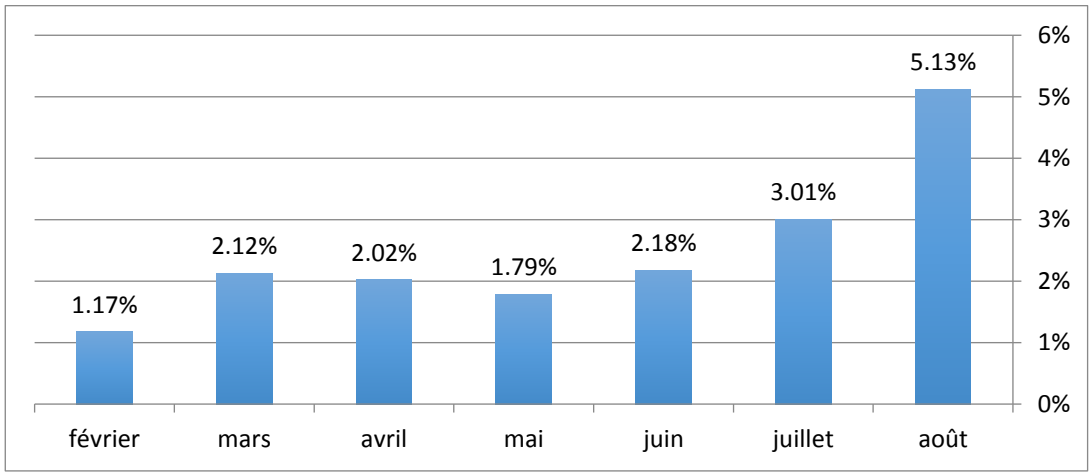


Figure 21 : Proportion des utilisateurs à une période d'activité maximale d'un mois, ayant twitté au moins 10 fois, par rapport à la totalité des utilisateurs du mois

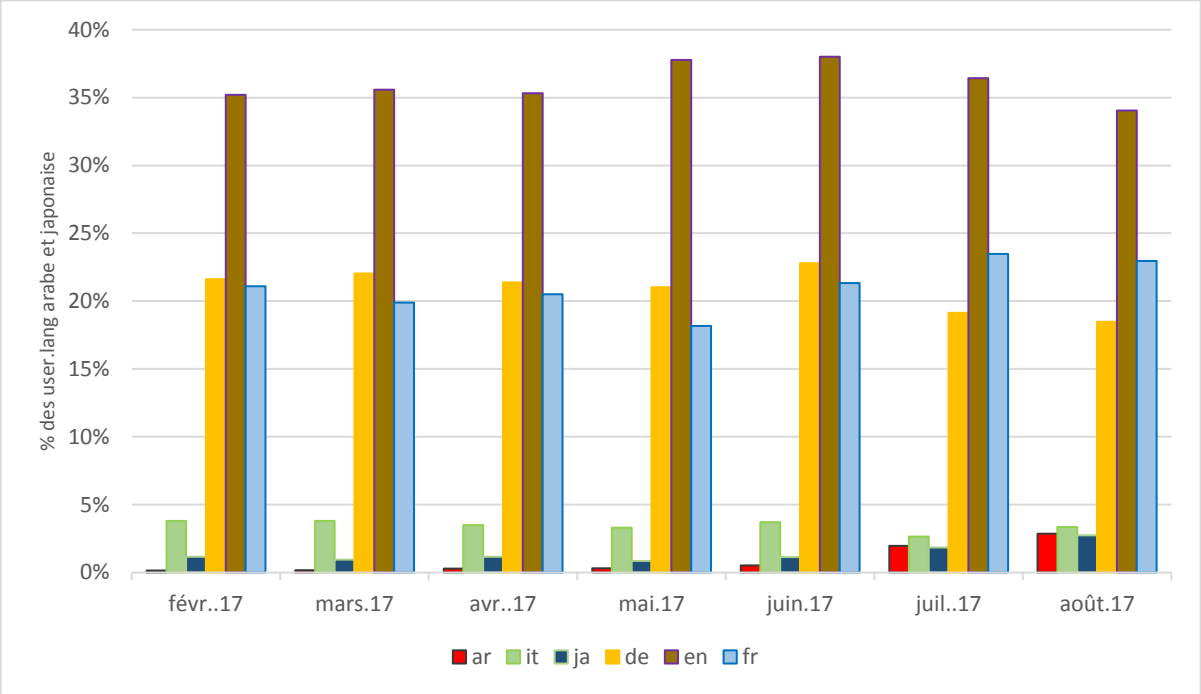


4.3.1.2.4 Répartition mensuelle des langues des comptes (User.lang) (valeurs réelles)

Le nombre des *user.lang* varie plus selon les mois que celui des source.lang. 36 au mois de février, 41 en mai, 44 en avril, 46, en mars, 48 en juin et août et 49 en juillet. En revanche la proportion des grandes *user.lang* varie bien moins que celle des grandes *source.lang* et

reste particulièrement stable pour les trois langues nationales. Nos deux langues-test pour les touristes, l'arabe et le japonais, montrent une augmentation nette pour les mois d'été (Figure 22).

Figure 22 : Répartition mensuelle de quatre « grandes » *user.lang* et les deux langues-test pour touristes (comptes)



4.3.1.3 Résultat de l'hypothèse (HLT1)

L'hypothèse (HLT1) concernant l'évolution différenciée de la quantité des tweets ainsi que les communautés linguistiques virtuelles selon les mois s'est donc parfaitement vérifiée. De plus les résultats des analyses nous ont permis de formuler une seconde hypothèse linguistico-temporelle que nous traiterons dans le chapitre suivant.

4.3.2 Hypothèse linguistico-temporel (HLT2)

Hypothèse (HLT2) : l'augmentation du temps libre (congés), les expériences inédites (vacances, loisirs), et l'apport extérieur des touristes intérieurs et extérieurs ont un effet positif sur l'augmentation du nombre des tweets.

L'hypothèse (HLT2) fut formulée en tenant compte des résultats suivants des analyses liées à la vérification de l'hypothèse (HLT1) dans le chapitre précédent :

- Les mois qui produisent le plus de tweets géolocalisés en moyenne journalière sont ceux des vacances : février-mars pour les vacances de ski et juillet-août pour les vacances d'été.
- La proportion très élevée des utilisateurs dont la période d'activité se limite à une période courte indique un comportement où le choix de twitter ou celui de géolocaliser ses tweets en Suisse dépend de facteurs liés à des contraintes ou opportunités temporels[45]. Les analyses montrent une augmentation très nette de leur proportion en été.

[45] Un utilisateur peu actif peut être incité à twitter ponctuellement également sous le coup d'une émotion forte, sans que ce changement de comportement soit lié aux vacances et aux congés. Une influence de la fête nationale du 1er août est par exemple

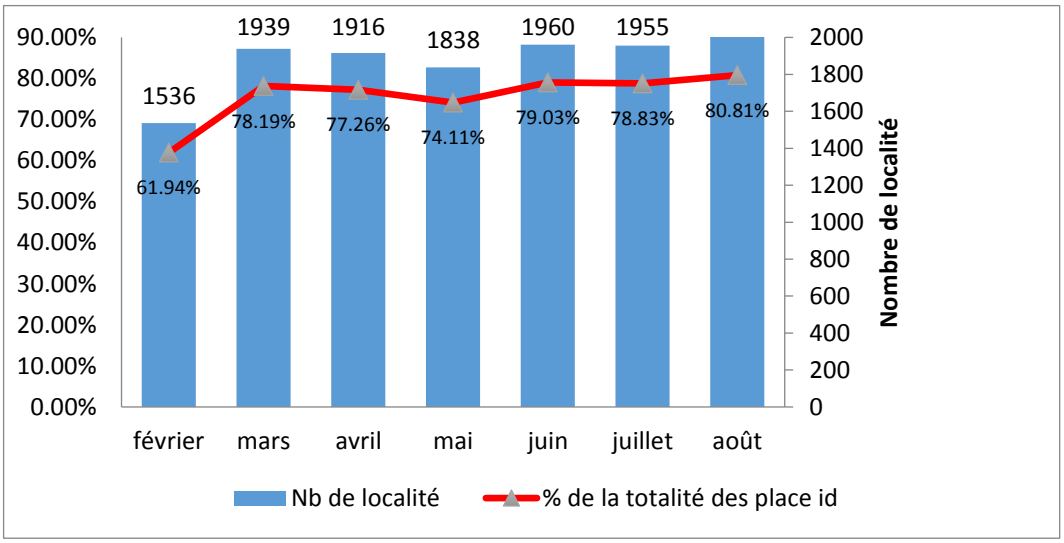
- On constate d'importantes variations dans la répartition mensuelle des source.lang. Tandis que la proportion de l'anglais, de l'allemand et de l'italien baisse, celle de certaines langues « typiques » des touristes augmente en été.
- Tandis que la proportion des grandes *User.lang* et particulièrement celle des trois langues nationales reste stable entre février et août, la proportion des comptes affichant certaines langues « typiques » des touristes augmente.

Pour tester cette hypothèse, en plus du nombre des indices recueillis par les analyses précédentes, nous avons effectué des analyses sur la répartition temporelle des *place.id* (Figure 23) et cherché des corrélations entre les répartitions mensuelles des variables temporelles suivantes[46] :

- (VT1) nombre des tweets
- (VT2) nombre des utilisateurs
- (VT3) nombre des utilisateurs à une période d'activité d'un mois
- (VT4) nombre des utilisateurs à un seul tweet, répartis par mois
- (VT5) nombre des *Localités*

Afin de pouvoir comparer avec ces variables, les analyses ont été faites sur des valeurs non corrigées.

Figure 23 : Répartition temporelle des localités Twitter



4.3.2.1 Résultats

Les analyses montrent une très forte corrélation entre chaque paire des 5 variables (0.945 – 0.991). L'augmentation du nombre des tweets est donc plus ou moins parallèle à

démontrable grâce à la fréquence *d'hashtags* qui y font référence, comme #1August et #SwissNationalDay. De l'autre côté les attentats de Barcelone ont également entraîné une augmentation du nombre des tweets, avec la forte utilisation du *hashtag* Barcelona entre le 17 et le 19 juillet. Bien que les deux tombent sur des mois d'été, il ne s'agit pas de relation cause à effet. Une étude des textes et des *hashtags* des messages serait souhaitable pour un examen approfondi de l'hypothèse HTL2, mais une analyse qualitative avec textmining dépasserait les limites de cette étude.

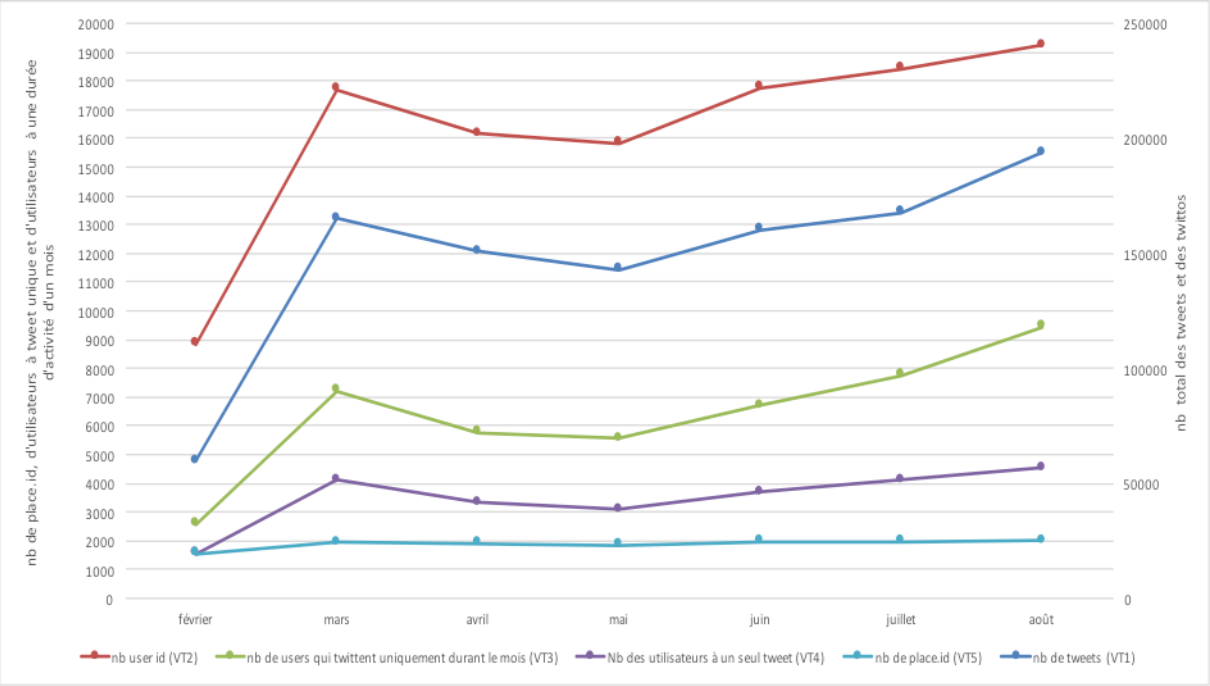
[46] D'autres variables, tout aussi prometteuses, n'ont pas été testées par manque de temps : origine des comptes, time zone, etc.

celle du nombre des utilisateurs, du nombre des utilisateurs à une *Période ou durée d'activité* d'un mois, des utilisateurs à un tweet unique, et des *place id* (Tableau 3). On peut l'également observer sur les courbes de répartition des cinq variables qui montrent également un degré important de parallélisme (Figure 24).

Tableau 3 : Corrélation entre les variables temporelles

Corrélation entre	**V1**	**V2**	**V3**	**V4**	**V5**
V1	1	0.991	0.961	0.980	0.984
V2	0.991	1	0.945	0.979	0.990
V3	0.961	0.945	1	0.979	0.917
V4	0.980	0.979	0.979	1	0.961
V5	0.984	0.990	0.917	0.961	1

Figure 24 : Courbes de la répartition mensuelle du nombre et de la moyenne des tweets, du nombre des utilisateurs, du nombre des utilisateurs à une durée d'activité d'un mois, et du nombre des localités Twitter



Les résultats des analyses sont donc positifs et soutiennent l'hypothèse HLT2. L'augmentation du nombre des utilisateurs durant les périodes des vacances et notamment de ceux dont la durée d'activité ne dépasse pas un mois, ainsi que la variation de leur répartition spatiale et linguistique durant les mois de vacances peut constituer un indice en ce sens. Toutefois pour le confirmer ces résultats nécessitent d'être confrontés d'une part avec le contenu des tweets, et d'autre part avec les statistiques officielles sur le nombre, la provenance et les destinations des touristes qui dépassent les limites de ce projet.

5. Réflexion sur la notion et la définition de "tweet suisse" dans le cadre de l'archivage des tweets

5.1 Définition d'un tweet suisse à partir des données de notre échantillon : question de recherche 3

5.1.1 Représentativité de l'échantillon

L'échantillon n'ayant pas été construit de manière probabiliste, la représentativité de l'échantillon est presque nulle. Les tweets géolocalisés sont une infime partie des tweets, les *Tweet privés* et les retweets en sont exclus (voir chapitres 3.1.3.2 et 3.1.3.2) et nous ne connaissons pas la taille exacte de notre population 2, vu qu'elle n'est pas définie. De gros biais sont donc possibles. Nous allons tout de même, à partir des données dont nous disposons et de la littérature, tenter de définir ce que pourrait être un « tweet suisse ».

5.1.2 Critères pour définir la nationalité

Sur internet, comment définir la nationalité d'un contenu ? Que cela soit un site internet ou un tweet, sur quels éléments se baser ?

La Bibliothèque nationale définit ainsi les documents qu'elle a pour mission d'acquérir et de conserver au nom de la Suisse :

> *« par Helvetica, on entend les écrits suisses[47]; toutes les œuvres parues à l'étranger concernant la Suisse et ses habitants; et finalement, les œuvres d'auteurs suisses parues à l'étranger, traductions comprises »*[48]

Elle archive également les « e-Helvetica », c'est-à-dire « *les informations […] stockées sur d'autres supports que le papier, ayant un lien avec la Suisse […] comme par exemple les e-books, les e-journals et les sites web* »[49]

Au vu de ces définitions et des études faites dans d'autres pays (voir chapitre 2.3.1), nous allons tester ici différents critères pour évaluer la 'suissitude' d'un tweet.

5.1.2.1 Langue des comptes ou des messages

Au contraire d'autres pays qui ont une langue unique et spécifique (suédois par exemple), la Suisse est multilingue : elle est composée de quatre langues nationales, dont le romanche qui n'est pas reconnu par Twitter (voir le chapitre 4.2) et dont les trois autres sont également celles des pays limitrophes. D'autre part, une grande utilisation de l'anglais et de multiples étrangers, touristes ou résidents compliquent l'identification des utilisateurs 'suisses' au travers de la métadonnée *user.lang.*

Le champ *source.lang*, au vu des tests réalisés sur la population 1, n'est pas fiable.

[47] « tout ce que les éditeurs suisses publient, toutes matières et langues confondues. La production éditoriale suisse ne se limite pas aux livres d'auteurs suisses, elle comprend énormément de publications d'auteurs étrangers soit dans leurs langues originales soit traduites. Les publications à compte d'auteurs, celles des organisations gouvernementales ou non gouvernementales ayant leur siège en Suisse, les catalogues des musées et des galeries d'art sont également recueillis. » https://www.nb.admin.ch/snl/fr/home/collections/helvetica/la-production-editoriale-suisse.html

[48] https://www.nb.admin.ch/snl/fr/home/collections/helvetica.html

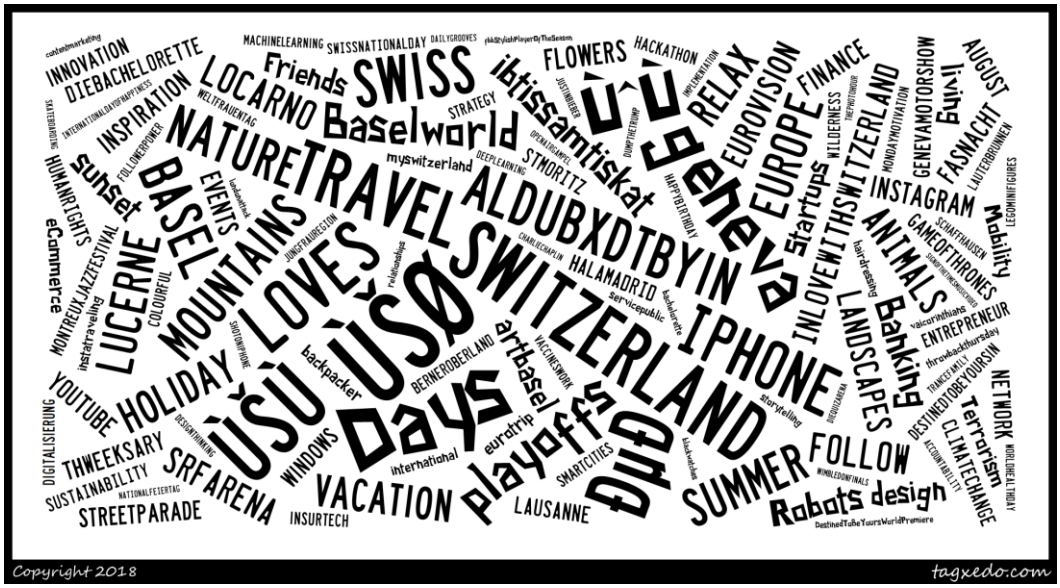
[49] https://www.nb.admin.ch/snl/fr/home/bn-professionnel/e-helvetica.html

La langue n'étant pas un critère pour la Bibliothèque nationale (puisqu'elle prend également les traductions), cela ne nous permet donc pas d'évaluer un tweet suisse.

5.1.2.2 Sujets des messages

Les *Hashtags* sont écrit sur une base volontaire et ne sont pas formalisés par un thésaurus. Il y a donc autant de *hashtag* que de twittos, particulièrement dans un pays multilingue. Voici un aperçu des mots les plus utilisés parmi nos tweets :

Figure 25 : Nuage des *hashtags* les plus utilisés



Le nom du pays sort en bonne place, ainsi que de nombreux autres lieux, personnalités, événements, … Afin de cibler des tweets suisses, une liste de mots pourrait être établie en partant des *hashtags* existants (Morstatter, Liu, 2017), en cherchant des mots ou localités suisses (Potting, 2016) ou en utilisant des *Trending topics* suisses. Les résultats pourraient être exploités directement ou en identifiant et rapatriant les twittos ayant mentionné ces mots (Marchetti, Ceccobelli, 2016 ; Ausserhofer, Maireder, 2013).

5.1.2.3 Localisation

Les twittos mettent à disposition, plus ou moins consciemment, plusieurs possibilités pour connaître leur localisation (voir aussi Annexe 1).

5.1.2.3.1 Géolocalisation fine *: données GPS*

Les twittos permettant ce type de localisation sont peu nombreux. Si d'après la littérature, seuls 1 à 2% des twittos indiquent une localisation, ceux donnant leurs coordonnées géographiques précises sont encore moins nombreux. Parmi notre échantillon, seuls 230'864 tweets (22,2%) écrits par 26'557 twittos (40,7%) les ont autorisées. Cette donnée fiable (puisqu'elle est fournie automatiquement par l'appareil de l'utilisateur) est très précieuse pour beaucoup de champs de recherche (détections de maladie, informations sur le trafic, aide en cas de catastrophe, …), et également pour définir un tweet envoyé depuis la Suisse.

5.1.2.3.2 Géolocalisation manuelle

Basée sur une déclaration du *twittos*, voire sur des automatismes de Twitter, cette localisation a moins de valeur que la précédente. Lors de nos tests du 10 mars 2017 depuis

le bâtiment B de la HEG à Carouge, nous avons déclaré twitter depuis des pays étrangers, tout comme des personnes à des milliers de kilomètres de la Suisse peuvent déclarer se trouver physiquement dans notre pays. Ces indications nous permettent de retrouver ces tweets géolocalisés par la moulinette de Twitter (voir Annexe 1) dans notre échantillon, et donc de multiplier par cinq les tweets localisés par GPS.

5.1.2.3.3 Localisation du profil

889'621 tweets (85,5% de notre échantillon) indiquent un « Emplacement du profil » sur leur compte. Ils ont été écrits par 24'681 utilisateurs (37,8%). Ceux-ci peuvent être très précis en mentionnant leur ville, voire leur quartier, tandis que d'autres seront plus évasifs (Suisse, Europe), voire carrément faux ou poétiques (« dans la lune »)

Figure 26 : Nuage des localisations déclarées

Etant donné que ce champ est complété par seulement un tiers des utilisateurs et qu'il contient des informations très diverses et non certifiées, cette technique n'est pas satisfaisante pour déterminer quels sont les tweets suisses, même si elle pourrait être utilisée pour 'rapatrier' des utilisateurs et donc des tweets suisses (Leetaru et al., 2013 ; Weidemann, Swift, 2013 ; Minot et al., 2015).

5.1.2.3.4 Localisation par fuseau horaire

Sur la base du projet australien TrISMA (voir chapitre 2.3.1.2.1), les indications données par les fuseaux horaires de notre échantillon ont été étudiées. 71% des tweets récoltés (envoyés par 43'252 twittos, soit 66,3%) contiennent un fuseau horaire (le champ *timezone* qui n'est pas obligatoire). Ils indiquent 184 lieux différents, dans le monde entier.

Figure 27 : Distribution par fuseaux horaires

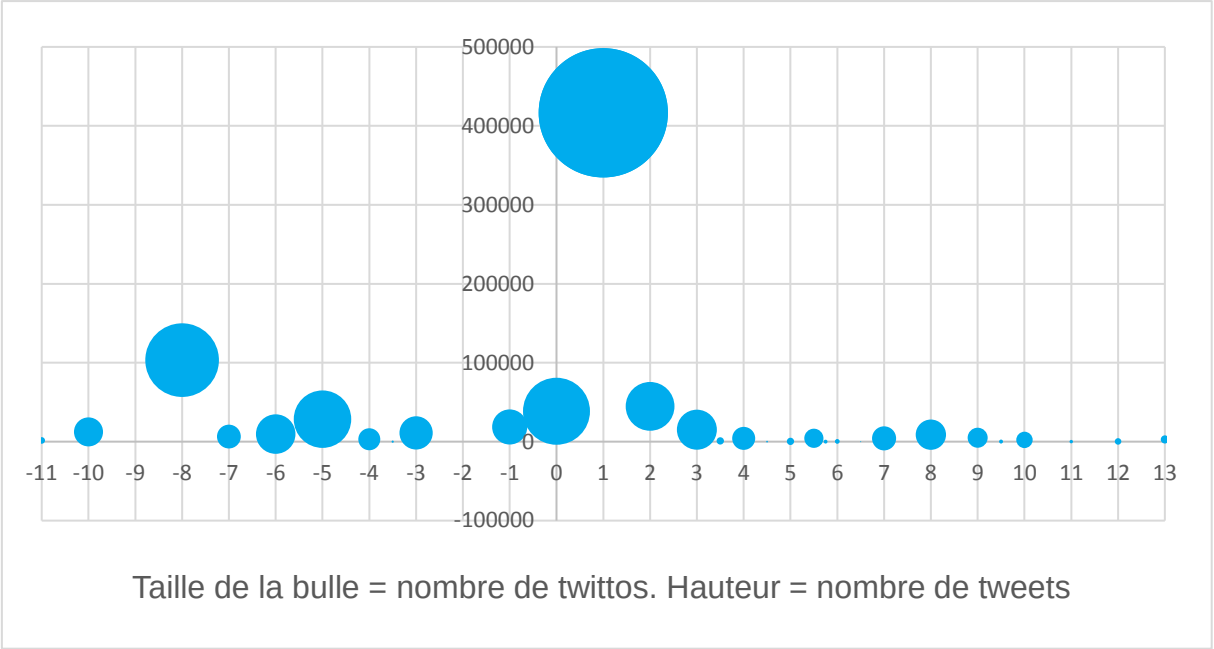


Tableau 4 : Distribution par continent

	Nombre de tweets		Nombre de twittos	
Amérique du Nord	156'539	21%	11'162	26%
Amérique centrale	603	0%	60	0%
Amérique du Sud	21'537	3%	2'222	5%
Europe	510'741	69%	24'122	56%
dont la Suisse	*240'835*	*32%*	*6'094*	*14%*
Afrique	7'785	1%	761	2%
Asie	43'174	6%	4'556	11%
Océanie	2'803	0%	362	1%
Non classables	20	0%	7	0%
Total général	**743'202**		**43'252**	

D'après nos analyses, les robots ne représenteraient que 9% des tweets envoyés avec d'autres fuseaux horaires, tandis que ceux envoyés depuis des applications tierces (Instagram par exemple) représentent 14%. Deux explications nous paraissent possibles pour expliquer cette diversité : soit ce sont des touristes qui ne changent pas leurs informations quand ils sont en Suisse, soit les twittos indiquent des informations pas toujours exactes dans ce champ.

Si les Australiens font du fuseau horaire une condition obligatoire pour être reconnu comme twittos australien, de notre côté nous ne pouvons être aussi éliminatoires. Ce champ peut être utile, mais si nous l'avions pris comme base pour échantillonner, nous n'aurions eu que 14% des tweets.

5.1.2.4 Tweets manuels - automatisés

Suite à la revue de littérature, nous nous demandions si des robots se géolocalisaient et si nous en aurions dans notre échantillon. En utilisant le champ *source_keyword* et en appliquant les méthodes décrites notamment par Tsou (2017), nous avons pu identifier 388

sources comme étant des robots (sur 504). Ceux qui étaient reconnaissables ont été classifiés plus finement. Notre analyse montre des tendances ; cependant elle mériterait d'être affinée, voire confrontée au Botometer (voir note 11).

Tableau 5 : Tweets manuels ou automatisés

	Nombre de tweets		**Nombre de twittos**	
Humain	915'374	88.03%	70'028	94.61%
Robot	107'074	10.30%	1'568	2.12%
dont robot-news (informations générales)	*33'349*	*3.21%*	*37*	*0.05%*
dont robot-météo (station météo ou info tremblement de terre	*15'757*	*1.52%*	*34*	*0.05%*
dont robot-pub (spam)	*9'381*	*0.90%*	*301*	*0.41%*
dont robot-transports (info contrôleurs en Suisse romande)	*7'075*	*0.68%*	*6*	*0.01%*
dont robot-jobs (emplois)	*4'475*	*0.43%*	*46*	*0.06%*
dont robot-porno	*1'034*	*0.10%*	*331*	*0.45%*
dont inclassables	*36'003*	*3.46%*	*813*	*1.10%*
Indéterminé	17'371	1.67%	2'425	3.28%
Total général	**1'039'819**	**100.00%**	[50] **74'021**	**100.00%**

Deux de nos plus grands twittos sont des robots : un twitte les *Trending topics* suisses (26'719 tweets sur la période concernée, le pic de tweets du canton d'Obwald, voir la Figure 5) et l'autre est une station météo qui envoie des données météorologiques toutes les 30 minutes.

Il est donc possible de séparer les robots des humains. 2% de robots envoient plus de 10% des tweets. Mais est-ce que ces robots peuvent être considérés comme 'suisses' ?

5.1.2.5 Comptes des *Twittos*

Les chercheurs australiens, autrichiens et suédois de notre revue de littérature (voir chapitre 2.3.1) ont tous répertoriés des comptes de twittos afin d'obtenir les tweets de leurs pays respectifs. En Suisse, les personnes ayant créé le projet Pegasus Data (voir chapitre 2.3.1.1), actualisent encore 12 listes établies de twittos romands (totalisant 5'742 personnes au 11 janvier 2018), et politiques (593 personnes)[51]. Cela pourrait être une base intéressante si les comptes suisses devaient être répertoriés, et permettrait de séparer les comptes des touristes. Mais se pose ensuite la question : si un *twittos* déménage à l'étranger, est-ce qu'il écrit encore des tweets suisses ? Les Australiens excluent de tels comptes de leurs données.

Deux signes permettraient d'identifier des touristes parmi les twittos :

- Le temps passé sur place croisé avec la durée du séjour : 5434 twittos (8,3%) sont signalés un mois sur place, mais ont circulé dans au moins deux cantons différents ;

[50] N = 74'021, car les 65'221 comptes twittos peuvent utiliser plusieurs moyens pour twitter : ordinateur, appareil mobile, application tierce, … cela concerne 7'044 comptes humains, 142 comptes robots et 11 indéterminés

[51] https://twitter.com/PegasusData/lists?lang=fr

- L'unicité des tweets : 24'320 twittos (37% de notre échantillon) n'a twitté qu'une fois (voir chapitre 4.3.1.2.3). Cela indique certainement le fait qu'ils étaient de passage en Suisse, voire qu'ils ont twitté depuis une application-tierce (comme Instagram) qui les géolocalise d'office.

Pour identifier de manière certaine les comptes suisses, il faudrait donc analyser plus en détail leur activité.

5.2 Pistes de réflexion pour la définition d'un « tweet suisse »

Après avoir étudié les critères mentionnés plus haut, il est encore difficile de définir un « tweet suisse ». Il nous semble qu'il ne saurait être défini par un seul critère, mais que serait la combinaison de plusieurs d'entre eux.

Parmi les critères définis pour les Helvetica, celui qui semble le plus évident à propos des tweets est le lien avec la Suisse.

Mais de quel type doit être le lien :

- les sujets en liens avec la Suisse ? ils peuvent être relativement faciles à trouver, via les *hashtags* pour autant que ceux-ci soient bien choisis et traduits dans les différentes langues. Le nom Suisse existe en plus de 10 variantes !
- des comptes twittos suiveurs d'institutions suisses : CFF, personnalités politiques, sportives ou locales, … ? est-ce que les 11'182'928 personnes qui suivent Roger Federer ont un « lien avec la Suisse » ?

Est-ce que les tweets envoyés depuis la Suisse, sont suisses ?

- les géo-localisables peuvent être techniquement connus et récupérés comme dans le cadre de notre étude. Mais comment récupérer les autres ?
- Et à contrario, si un twittos suisse part en vacances à l'étranger, est-ce que ses tweets sont suisses ?
- Est-ce que le fuseau horaire indiqué sur le compte est à considérer, même si cela ne concerne pas tous les twittos ou que les déclarations ne peuvent pas être vérifiées ?

Est-ce que le twittos doit être Suisse ?

- Comment définir un compte 'suisse ? est-ce la nationalité (mais comment la connaître), le fait d'habiter en Suisse, de manière permanente ou pas ?
- Est-ce que les tweets envoyés par des robots sont suisses ? certains sont géolocalisés en Suisse, mais font partie d'une entreprise internationale.
- Est-ce que le twittos doit parler une langue nationale ou twitter dans une de ces trois[52] langues ?

On le voit, de nombreux critères peuvent être étudiés sans qu'une réponse unique ne puisse être donnée. Nous avons étudié notre échantillon sous l'angle quantitatif ; peut-être qu'une analyse qualitative pourrait aider à mieux comprendre les tweets suisses et proposer une définition plus précise ? Peut-être que dans le monde sans frontières d'internet, il est impossible d'avoir une définition unique de la nationalité ?

[52] Le romanche n'étant pas reconnu

5.3 Archivage des tweets suisses

Définir un tweet suisse n'est pas évident. Définir ce qui doit être sauvegardé pour les générations futures non plus. La Bibliothèque du Congrès n'archive plus l'intégralité des tweets depuis janvier 2018 (voir chapitre 2.3.2.1) et les autres bibliothèques archivent seulement partiellement les réseaux sociaux, dont Twitter (Meikle, 2013). Les Australiens archivent les tweets reliés à des comptes australiens, mais également les liens entre les comptes (*followers*, *followees*) et les interactions entre les tweets : *retweets*, favoris, *likes* (voir chapitre 2.3.1.2.1).

Comme les projets d'archivage des tweets actuellement connus ne conservent que les messages bruts, de nombreux chercheurs (Acker, Kriesberg, 2017 ; Gayo-Avello, 2016 ; Leetaru, 2017 ; Thomson, 2017 ; Zimmer, 2015) plaident en faveur d'un archivage non seulement des fichiers textes ou .json, mais également de l'entièreté de Twitter :

> *« (1) the cultural artifact that is Twitter, with (1a) its look and feel and technical affordances over the course of time, and (1b) the broader societal context into which Twitter is embedded, including user numbers, demographics and usage practices, and (2) the Twitter data consisting of (2a) the complete collection of all user-generated content, including non-textual information and hyperlinks, and (2b) contextual information like collections of hashtags for important events or lists of usernames for important groups of users. » (Bruns, Weller, 2016, p. 185)*

En Suisse, rien n'est encore fait, même si la réflexion a débuté (voir chapitre 2.3.1.1). D'après Barbara Signori de la Bibliothèque nationale, l'archivage des réseaux sociaux pourrait reprendre les critères établis pour les Helvetica (voir chapitre 5.1.2). Cependant, l'exhaustivité étant difficile à atteindre, l'archivage pourrait être sélectif, comme pour les sites internet : la collection e-Helvetica de la BN se construit sur une stratégie sélective qui exclut, pour l'instant, une collecte systématique du domaine .ch[53]. Pour échantillonner, une possibilité serait d'archiver les publications officielles[54], c'est-à-dire les comptes Twitter de la Confédération, des cantons, voire d'organismes fédéraux (sur le modèle du web canadien, la Bibliothèque et Archives Canada archive depuis 2005 les sites internet du gouvernement du Canada)[55], ou encore, comme pour les e-Helvetica, de mandater les bibliothèques cantonales pour désigner les comptes à enregistrer.

Le Consortium pour la préservation d'internet (International Internet Preservation Consortium)[56] regroupe 45 pays, dont la Suisse. Il organise chaque année un colloque international sur l'archivage du web, mène des projets collaboratifs et met à disposition des informations qui peuvent être utiles également pour les réseaux sociaux : établissement de liens permanents, accès aux contenus effacés, textmining, etc.

Après avoir défini un tweet et les critères d'archivage, se poseront encore les questions d'accès, comme pour la Bibliothèque du Congrès …

Entre intérêt historique et protection des personnes, défis techniques et complétude, de nombreux problèmes sont encore à régler. Ce travail n'a pas pour ambition de les résoudre.

[53] https://www.nb.admin.ch/snl/fr/home/bn-professionnel/e-helvetica/infos-pour-les-fournisseurs/sites-web-_-archives-web-suisse/faq-sur-l_archivage-web.html.
[54] https://www.nb.admin.ch/snl/fr/home/collections/helvetica/publications-officielles.html
[55] http://www.collectionscanada.gc.ca/archivesweb/index-f.html. Les autres sites ne sont actuellement pas archivés, les questions de droit d'auteur n'ayant pas été réglées.
[56] http://netpreserve.org

6. Conclusions

Cette étude exploratoire sur un million de tweets géolocalisés en Suisse est à notre connaissance la première à être réalisée. Les résultats ont clairement confirmé l'intérêt, mais également montré les limites de l'utilisation des données fournies par l'API de Twitter dans des recherches sur la Suisse, dans les domaines de la sociologie des données et des sciences de l'information. Nous avons donc structuré nos conclusions en deux temps. Dans ce chapitre, nous allons mettre en avant les résultats de l'étude et répondre aux trois questions de recherche, tandis que dans le suivant nous formulerons nos recommandations dans le but de poursuivre des recherches sur les tweets géolocalisés suisses.

6.1 Comment peut-on exploiter la géolocalisation de Twitter dans le contexte suisse ? (QR1)

La base de données utilisée par Twitter pour assigner des *place.id* à la géolocalisation des tweets est sujette à caution. Sa composition et sa maintenance sont opaques. Notre échantillon a confirmé qu'elle comporte des données erronées et obsolètes. Pour pouvoir l'exploiter il est absolument indispensable de croiser et corriger ces données par des données officielles et mises à jour. Après avoir effectué ce travail grâce à la création d'une table de concordance, nous avons pu dégager des résultats intéressants sur la répartition spatiale des tweets et des twittos au niveau des cantons.

Le premier constat important est d'une part la répartition très inégale des tweets et des comptes actifs parmi les cantons, Genève, Zurich et Vaud dépassant très largement les autres. D'autre part, cette répartition virtuelle ne reflète pas la répartition réelle de la population, les cantons romands étant surreprésentés tout aussi bien en proportion des tweets que des twittos par rapport à celle de leur population. Une des hypothèses que nous avons avancées pour expliquer ce fait, sans avoir eu l'occasion de la tester, est l'utilisation moins fréquente de la géolocalisation chez les twittos alémaniques par rapport aux twittos romands. Un indice en ce sens pourrait toutefois être le fait que la proportion des comptes germanophones de notre échantillon (15%) est à peine plus élevée que celle des comptes francophones (13 %), et pourtant dans la littérature rien n'indique que les Alémaniques ou les Allemands auraient en moyenne moins de comptes Twitter que les autres communautés linguistiques.

6.2 Dans quelle mesure le système d'identification automatique des langues par Twitter permet-il d'obtenir une image réelle de la diversité linguistique de la Suisse ? (QR2)

Les tests réalisés sur notre échantillon ont permis de confirmer la grande marge d'erreur de la détection automatique des langues des tweets et ce tout aussi bien pour les grandes langues européennes que pour les langues « exotiques ». Afin d'obtenir une image plus réelle des pratiques linguistiques virtuelles de la Suisse, et pour pallier à cette incertitude, nous avons pris le parti de croiser les données fournies par Twitter sur les *source.lang* avec celles fournies par les utilisateurs eux-mêmes sur les *user.lang*. Les résultats de cette méthode nous semblent prometteurs.

Nous avons ainsi confirmé la première place de l'anglais dans la Suisse virtuelle, cette langue arrivant en première position à la fois comme *source.lang* et comme *user.lang*, et ce tout aussi bien au niveau du pays que dans la majorité des cantons. Il est toutefois bien plus présent dans les cantons alémaniques que dans les cantons romands et au Tessin, où, pour la plupart, c'est une langue nationale qui arrive en première position, à l'exception notable de Genève. Les analyses croisées ont clairement démontré que ce fait ne s'explique pas uniquement par le grand nombre des comptes anglophones du pays, mais également par la propension importante des Alémaniques à twitter en anglais. Et comme mentionné plus haut, l'évitement de la géolocalisation des Alémaniques peut également avoir une incidence sur ces résultats.

L'analyse de la répartition temporelle des tweets et des twittos a également permis d'affiner nos connaissances sur les pratiques linguistiques virtuelles de la Suisse. Nous avons d'abord constaté une forte augmentation du nombre des messages et des comptes actifs pendant les mois d'été et particulièrement au mois d'août. L'augmentation du nombre des twittos n'ayant twitté qu'une seule fois durant la période de collecte, de même que ceux dont la *Période ou durée d'activité* se limite à un seul mois, sont les plus significatives pendant l'été. Tandis que le nombre des *source.lang* et des *user.lang* reste stable tout au long de la période, la variation de leur composition est remarquable. D'une part la proportion des trois langues nationales et de l'anglais stagne ou même baisse en été tout aussi bien comme *source.lang* que comme *user.lang*. D'autre part, les deux langues « typiques » de touristes que nous avons testées, l'arabe et le japonais, augmentent exponentiellement dans les deux champs. Le nombre des localités où les tweets sont géolocalisés monte également pendant ces périodes. La corrélation entre les cinq variables testées s'est révélée significative (0.945-0.991) et nous a permis de confirmer notre hypothèse sur l'effet positif de l'augmentation du temps libre (congés), les expériences inédites (vacances, loisirs), et l'apport extérieur des touristes intérieurs et extérieurs sur l'augmentation du nombre des tweets.

6.3 Est-il possible de définir un tweet suisse à partir des données de notre échantillon ? (QR3)

Nous n'avons pas de réponse définitive à donner vu la complexité du sujet, mais nous espérons que les critères ébauchés pourront servir comme base de réflexion. S'il existe des règles établies pour tout ce qui est production 'papier', il est difficile de les appliquer à du contenu électronique, encore plus au contenu des réseaux sociaux, matériel plus volatile, moins cadré que des sites internet et rédigé par une multitude de personnes dans de nombreuses langues. La Suisse, pays quadrilingue mais sans langue spécifique, dont les particularités ne sont pas reconnues par Twitter, est un cas encore plus complexe que d'autres pays. Les critères devront être définis et affinés : nationalités, langues, lieux ...

Plusieurs comportements typiques permettant d'identifier les touristes et les robots ont été proposés. Il reste cependant encore à déterminer si les tweets envoyés par des touristes et des robots sont suisses.

7. Recommandations

Nos analyses ont soulevé plusieurs points qui mériteraient une investigation plus poussée, que les limites de cette étude n'ont pas permise, et d'autres que nous n'avons pas eu la possibilité d'aborder dans le cadre d'une recherche purement quantitative. Selon les connaissances acquises au travers de ce projet de recherche, nous proposons :

- de comparer les analyses entre les localisations fines et manuelles
- d'étendre les analyses à la localisation déterminée par le textmining sur le contenu des messages ou des localisations des comptes. D'après la littérature, en exploitant ces informations le taux de localisation des tweets peut augmenter à 70% (pour autant que ce soit compatible avec l'actuelle et la future loi sur la protection des données).
- d'approfondir la granularité des analyses au niveau des communes au lieu des cantons
- d'examiner les raisons des différences de l'utilisation de Twitter entre les différentes parties linguistiques
- de développer des méthodes pour reconnaître les tweets en romanche
- de mener des analyses de type textmining sur les messages ou les localisations afin d'approfondir les critères de 'suissitude' des tweets
- de lister les twittos suisses en partant de la base du projet Pegasus Data. Cela permettra de récupérer ensuite leurs tweets et les métadonnées afin de mener des analyses multicritères, sur le modèle du projet TrISMA australien. Ces chercheurs pourraient être contactés afin de profiter de leurs connaissances en infrastructure et de leurs '*best practices*'
- de poursuivre les analyses sur les robots : les estimations faites dans ce travail mériteraient d'être affinées et confrontées à d'autres approches.
- de compléter le travail d'analyse des statistiques OFS sur le nombre, la provenance et la répartition des touristes

Après avoir travaillé sur ce projet pendant près d'un an, nous sommes persuadées que les tweets suisses – avec ou sans géolocalisation – n'ont pas encore livré tous leurs secrets….

Bibliographie

ACKER, Amelia et KRIESBERG, Adam, 2017. Tweets may be archived: Civic engagement, digital preservation and obama white house social media data. In : *Proceedings of the Association for Information Science and Technology* [en ligne]. 24 octobre 2017. Vol. 54, n° 1, p. 1‑9. [Consulté le 28 novembre 2017]. Disponible à l'adresse : http://onlinelibrary.wiley.com/doi/10.1002/pra2.2017.14505401001/abstract.

AGENCE FRANCE PRESSE, 2017. Des campagnes d'influence ont été postées sur Twitter depuis la Russie. In : *LeTemps.ch* [en ligne]. Genève, 29 septembre 2017. [Consulté le 3 janvier 2018]. Disponible à l'adresse : https://www.letemps.ch/monde/2017/09/29/campagnes-dinfluence-ont-postees-twitter-russie.

AMINEDIGIREP, 2011. 2,5 millions d'utilisateurs de Twitter en France, 9,5% protègent leur compte. In : *Digital Reputation Blog* [en ligne]. 16 mars 2011. [Consulté le 29 décembre 2017]. Disponible à l'adresse : http://digitalreputationblog.com/2011/03/16/2-5-millions-dutilisateurs-de-twitter-en-france/.

AUSSERHOFER, Julian et MAIREDER, Axel, 2013. National politics on Twitter: Structures and topics of a networked public sphere. In : *Information, Communication & Society* [en ligne]. avril 2013. Vol. 16, n° 3, p. 291‑314. [Consulté le 29 décembre 2017]. Disponible à l'adresse : http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.756050.

BANFI, Elisa et BÉGUELIN, Fanny, 2016. *GEoTweet: exploration des tweets géolocalisés à Genève* [en ligne]. Genève : Haute école de gestion de Genève HEG-GE. [Consulté le 28 février 2017]. Disponible à l'adresse : http://doc.rero.ch/record/258990. 1072061407

BEEVOLVE INC., 2012. An Exhaustive Study of Twitter Users Across the World. In : *http://www.beevolve.com* [en ligne]. 10 octobre 2012. [Consulté le 29 décembre 2017]. Disponible à l'adresse : http://www.beevolve.com/twitter-statistics/.

BROWN, Brendan, 2017. Trump Twitter Archive. In : [en ligne]. 2 décembre 2017. [Consulté le 28 novembre 2017]. Disponible à l'adresse : http://www.trumptwitterarchive.com.

BRUNS, Axel, 2017. Australian Twitter is more diverse than you think. In : *The Conversation* [en ligne]. 3 mai 2017. [Consulté le 28 novembre 2017]. Disponible à l'adresse : http://theconversation.com/australian-twitter-is-more-diverse-than-you-think-76864.

BRUNS, Axel, BURGESS, Jean et HIGHFIELD, Tim, 2014. A 'big data'approach to mapping the Australian Twittersphere. In : *Advancing Digital Humanities* [en ligne]. S.l. : Springer. p. 113–129. [Consulté le 28 novembre 2017]. Disponible à l'adresse : https://eprints.qut.edu.au/82986/1/A%20Big%20Data%20Approach%20to%20Mapping%20the%20Australian%20Twittersphere.pdf.

BRUNS, Axel, MOON, Brenda, MÜNCH, Felix et SADKOWSKY, Troy, 2017. The Australian Twittersphere in 2016: Mapping the Follower/Followee Network. In : *Social Media+ Society*. 13 décembre 2017. Vol. 3, n° 4. DOI 10.1177/2056305117748162.

BRUNS, Axel et WELLER, Katrin, 2016. Twitter As a First Draft of the Present: And the Challenges of Preserving It for the Future. In : *Proceedings of the 8th ACM Conference on Web Science* [en ligne]. New York, NY, USA : ACM. 2016. p. 183–189. [Consulté le 28 novembre 2017]. Disponible à l'adresse : http://doi.acm.org/10.1145/2908131.2908174.

CASEY, Dylan, 2010. Replay it: Google search across the Twitter archive. In : *Official Google Blog* [en ligne]. 14 avril 2010. [Consulté le 28 novembre 2017]. Disponible à l'adresse : https://googleblog.blogspot.com/2010/04/replay-it-google-search-across-twitter.html.

CHU, Zi, GIANVECCHIO, Steven, WANG, Haining et JAJODIA, Sushil, 2012. Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg? In : *IEEE Transactions*

on Dependable and Secure Computing [en ligne]. novembre 2012. Vol. 9, n° 6, p. 811‑824. [Consulté le 29 octobre 2017]. Disponible à l'adresse : http://ieeexplore.ieee.org/document/6280553.

COËFFÉ, Thomas, 2017. Chiffres Twitter - 2017. In : *Blog du Modérateur* [en ligne]. 18 septembre 2017. [Consulté le 6 novembre 2017]. Disponible à l'adresse : https://www.blogdumoderateur.com/chiffres-twitter/.

CORTHÉSY, Matthieu, 2015. Le nombre d'utilisateurs de Twitter en Suisse en 2015. In : *PME WEB* [en ligne]. 26 janvier 2015. [Consulté le 22 avril 2017]. Disponible à l'adresse : https://www.pme-web.com/nombre-utilisateurs-twitter-suisse-2015/.

DAVIS, Clayton Allen, VAROL, Onur, FERRARA, Emilio, FLAMMINI, Alessandro et MENCZER, Filippo, 2016. Botornot: A system to evaluate social bots. In : *Proceedings of the 25th International Conference Companion on World Wide Web*. Geneva : International World Wide Web Conferences Steering Committee. 2016. p. 273–274. arXiv.org : 1602.00975

DIAS CARDOSO, Pedro Miguel et ROY, Anindya, 2016. Language Identification for Social Media: Short Messages and Transliteration. In : *Proceedings of the 25th International Conference Companion on World Wide Web*. New-York : ACM Press. 2016. p. 611‑614.

EXTRADIGITAL, 2016. Social Media in Germany. In : *ExtraDigital* [en ligne]. mai 2016. [Consulté le 21 décembre 2017]. Disponible à l'adresse : https://www.extradigital.co.uk/articles/social-media/social-media-germany.html.

GATES, Sara, 2013. Are You Inadvertently Tweeting Your Location? Many Are, Study Finds. In : *Huffington Post* [en ligne]. 4 septembre 2013. [Consulté le 3 janvier 2018]. Disponible à l'adresse : https://www.huffingtonpost.com/2013/09/04/twitter-users-reveal-location-tweets_n_3867930.html.

GAUDINAT, Arnaud, 2016. GEoTweet: Les tweets geolocalisées de Genève. In : *hesge.ch* [en ligne]. 9 mai 2016. [Consulté le 30 mars 2017]. Disponible à l'adresse : http://geotweet.hesge.ch.

GAYO-AVELLO, Daniel, 2016. *How I Stopped Worrying about the Twitter Archive at the Library of Congress and Learned to Build a Little One for Myself* [en ligne]. 24 novembre 2016. S.l. : s.n. [Consulté le 28 novembre 2017]. Disponible à l'adresse : http://arxiv.org/abs/1611.08144. arXiv.org

GILANI, Zafar, FARAHBAKHSH, Reza, TYSON, Gareth, WANG, Liang et CROWCROFT, Jon, 2017a. An in-depth characterisation of Bots and Humans on Twitter. In : *arXiv:1704.01508 [cs]* [en ligne]. 5 avril 2017. [Consulté le 29 octobre 2017]. Disponible à l'adresse : http://arxiv.org/abs/1704.01508.

GILANI, Zafar, FARAHBAKHSH, Reza, TYSON, Gareth, WANG, Liang et CROWCROFT, Jon, 2017b. Of Bots and Humans (on Twitter). In : *ASONAM '17 , July 31 - August 03, 2017* [en ligne]. Sydney : IEEE/ACM. 2017. [Consulté le 29 octobre 2017]. Disponible à l'adresse : https://www.cl.cam.ac.uk/~szuhg2/docs/papers/ASONAM17_8501_65_1.pdf.

GRAHAM, Mark, HALE, Scott A. et GAFFNEY, Devin, 2014. Where in the World Are You? Geolocation and Language Identification in Twitter. In : *The Professional Geographer* [en ligne]. 2 octobre 2014. Vol. 66, n° 4, p. 568‑578. [Consulté le 18 mars 2017]. Disponible à l'adresse : http://dx.doi.org/10.1080/00330124.2014.907699.

HARRIS INTERACTIVE, 2017. Social Life 2017 - Baromètre des usages des réseaux sociaux. In : *Harris interactive* [en ligne]. 30 mars 2017. [Consulté le 22 avril 2017]. Disponible à l'adresse : http://harris-interactive.fr/newsfeeds/social-life-2017-barometre-annuel-des-usages-des-reseaux-sociaux-en-france/.

JEANNERET, Philippe, 2015. *L'exploration du Big Data par sa visualisation – Application au projet GEoTweet* [en ligne]. Genève : HEG-GE. [Consulté le 24 mars 2017]. Disponible à l'adresse : http://doc.rero.ch/record/258631/files/Travail_Bachelor_-_Philippe_Jeanneret.pdf. TDIG 129

KINDER-KURLANDA, Katharina, WELLER, Katrin, ZENK-MÖLTGEN, Wolfgang, PFEFFER, Jürgen et MORSTATTER, Fred, 2017. Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. In : *Big Data & Society* [en ligne]. 1 décembre 2017. Vol. 4, n° 2. [Consulté le 8 janvier 2018]. Disponible à l'adresse : https://doi.org/10.1177/2053951717736336.

KUMAR, Abhinav, SINGH, Jyoti Prakash et RANA, Nripendra P., 2017. Authenticity of Geo-Location and Place Name in Tweets. In : *AMCIS 2017 Proceedings* [en ligne]. Atlanta : Association for Information Systems. 2017. [Consulté le 26 décembre 2017]. Disponible à l'adresse : https://www.researchgate.net/publication/317013784. AMCIS-0493-2017.R1

LATZER, Michael, BÜCHI, Moritz, FESTIC, Noemi et JUST, Natascha, 2017. *Internetanwendungen und deren Nutzung in der Schweiz 2017 : Themenbericht aus dem World Internet Project – Switzerland 2017* [en ligne]. Zürich. Universität Zürich. [Consulté le 2 janvier 2018]. Disponible à l'adresse : http://mediachange.ch/media/pdf/publications/Anwendungen_Nutzung_2017.pdf.

LEETARU, Kalev, 2017. Why We Need To Archive The Web In Order To Preserve Twitter. In : *Forbes* [en ligne]. 18 juillet 2017. [Consulté le 28 novembre 2017]. Disponible à l'adresse : https://www.forbes.com/sites/kalevleetaru/2017/07/18/why-we-need-to-archive-the-web-in-order-to-preserve-twitter/.

LEETARU, Kalev, WANG, Shaowen, CAO, Guofeng, PADMANABHAN, Anand et SHOOK, Eric, 2013. Mapping the global Twitter heartbeat: The geography of Twitter. In : *First Monday* [en ligne]. 22 avril 2013. Vol. 18, n° 5. [Consulté le 28 octobre 2017]. Disponible à l'adresse : http://firstmonday.org/ojs/index.php/fm/article/view/4366.

LIU, Yabing, KLIMAN-SILVER, Chloe et MISLOVE, Alan, 2014. The Tweets They Are a-Changin: Evolution of Twitter Users and Behavior. In : *Proceedings of the Eighth International Conference on Weblogs and Social Media* [en ligne]. Palo Alto, California : AAAI press. 2014. [Consulté le 29 décembre 2017]. Disponible à l'adresse : http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8043. AAAI Digital Library

MARCHETTI, Rita et CECCOBELLI, Diego, 2016. Twitter and Television in a Hybrid Media System. In : *Journalism Practice* [en ligne]. 3 juillet 2016. Vol. 10, n° 5, p. 626‑644. [Consulté le 2 janvier 2018]. Disponible à l'adresse : https://doi.org/10.1080/17512786.2015.1040051.

MARI, Marcello, 2013. Twitter: 4 milioni di utenti attivi in Italia. In : *Tech Economy* [en ligne]. 27 février 2013. [Consulté le 2 janvier 2018]. Disponible à l'adresse : http://www.techeconomy.it/2013/02/27/twitter-4-milioni-di-utenti-attivi-in-italia/.

MATÉRNE, Hanna, 2017. *Gender Inference on Twitter in Swedish Contexts : Master's thesis* [en ligne]. Gothenburg : Chalmers University of Technology. [Consulté le 29 décembre 2017]. Disponible à l'adresse : publications.lib.chalmers.se/records/fulltext/250449/250449.pdf.

MCGILL, Andrew, 2016. Can Twitter Fit Inside the Library of Congress? In : *The Atlantic* [en ligne]. 4 août 2016. [Consulté le 17 octobre 2017]. Disponible à l'adresse : https://www.theatlantic.com/technology/archive/2016/08/can-twitter-fit-inside-the-library-of-congress/494339/.

MEEDER, Brendan, TAM, Jennifer, KELLEY, Patrick Gage et CRANOR, Lorrie Faith, 2010. RT@ IWantPrivacy: Widespread violation of privacy settings in the Twitter social network. In : *W2SP 2010: Web 2.0 Security and Privacy 2010* [en ligne]. S.l. : s.n. 20 mai 2010. p. 1–2. [Consulté le 27 novembre 2017]. Disponible à l'adresse : http://www.w2spconf.com/2010/papers/p28.pdf.

MEIKLE, James, 2013. British Library adds billions of webpages and tweets to archive. In : *The Guardian* [en ligne]. London, 4 avril 2013. [Consulté le 2 janvier 2018]. Disponible à l'adresse : http://www.theguardian.com/technology/2013/apr/05/british-library-archive-webpages-tweets.

MINOT, Ariana S., HEIER, Andrew, KING, Davis, SIMEK, Olga et STANISHA, Nicholas, 2015. Searching for Twitter Posts by Location. In : *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* [en ligne]. Boston : ACM Press. 2015. p. 357‑360. [Consulté le 30 décembre 2017]. Disponible à l'adresse : http://dl.acm.org/citation.cfm?doid=2808194.2809480.

MOON, Brenda, 2017. Identifying Bots in the Australian Twittersphere. In : *#SMSociety'17, July 28-30, 2017* [en ligne]. Toronto, Canada : ACM Press. 2017. [Consulté le 29 octobre 2017]. Disponible à l'adresse : http://dl.acm.org/citation.cfm?doid=3097286.3097335.

MORSTATTER, Fred et LIU, Huan, 2017. Discovering, assessing, and mitigating data bias in social media. In : *Online Social Networks and Media* [en ligne]. juin 2017. Vol. 1, p. 1–13. [Consulté le 29 décembre 2017]. Disponible à l'adresse : http://www.sciencedirect.com/science/article/pii/S2468696416300040.

MORSTATTER, Fred, PFEFFER, Jürgen, LIU, Huan et CARLEY, Kathleen M., 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In : *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media* [en ligne]. Boston : AAAI press. 2013. [Consulté le 29 octobre 2017]. Disponible à l'adresse : http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/download/6071/6379.

MOSCA, Giuditta, 2016. Social media in Italia: crollano Google+ e Twitter, esplode Snapchat. In : *Wired* [en ligne]. 4 avril 2016. [Consulté le 10 janvier 2018]. Disponible à l'adresse : https://www.wired.it/internet/social-network/2016/04/04/social-media-italia-crollo-twitter-esplode-snapchat/.

NGUESSAN, Noel, 2016. Twitter crée la Timeline des Tweets géolocalisés avec Foursquare. In : *Arobasenet.com* [en ligne]. juin 2016. [Consulté le 3 janvier 2018]. Disponible à l'adresse : http://www.arobasenet.com/2016/06/twitter-tweets-geolocalises-avec-foursquare-3102.html.

OFFICE FÉDÉRAL DE LA STATISTIQUE, 2017a. Liste historisée des communes de la Suisse [fichier TXT]. In : *Office fédéral de la statistique* [en ligne]. 23 mars 2017. [Consulté le 11 janvier 2018]. Disponible à l'adresse : https://www.bfs.admin.ch/bfs/fr/home/bases-statistiques/repertoire-officiel-communes-suisse/liste-historisee-communes.assetdetail.2245010.html.

OFFICE FÉDÉRAL DE LA STATISTIQUE, 2017b. Population résidante permanente selon l'âge, par canton, district et commune, 2010-2016 [fichier excel]. In : *Office fédéral de la statistique* [en ligne]. 30 août 2017. [Consulté le 11 janvier 2018]. Disponible à l'adresse : https://www.bfs.admin.ch/bfs/fr/home/statistiques/population/effectif-evolution/population.assetdetail.3222001.html.

OFFICE FÉDÉRAL DE LA STATISTIQUE, 2017c. Répertoire officiel des communes de Suisse [Version MS-Excel]. In : *Office fédéral de la statistique* [en ligne]. 23 mars 2017. [Consulté le 11 janvier 2018]. Disponible à l'adresse : https://www.bfs.admin.ch/bfs/fr/home/bases-statistiques/repertoire-officiel-communes-suisse.assetdetail.2245009.html.

OSTERBERG, Gayle, 2013. Update on the Twitter Archive at the Library of Congress. In : *Library of Congress Blog* [en ligne]. 4 janvier 2013. [Consulté le 28 novembre 2017]. Disponible à l'adresse : //blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/.

OSTERBERG, Gayle, 2017. Update on the Twitter Archive at the Library of Congress. In : *Library of Congress Blog* [en ligne]. 26 décembre 2017. [Consulté le 3 janvier 2018]. Disponible à l'adresse : http://blogs.loc.gov/loc/2017/12/update-on-the-twitter-archive-at-the-library-of-congress-2.

PEGASUSDATA, 2013. Suisses romands sur Twitter : combien sont-ils ? In : *Pegasus Data Project* [en ligne]. 9 septembre 2013. [Consulté le 22 avril 2017]. Disponible à l'adresse : https://pegasusdata.com/2013/09/09/suisses-romands-twitter/.

PLEWNIA, Lukas, 2017. Wie Twitter die Schweiz bewegt. In : *PPC Insider* [en ligne]. 5 novembre 2017. [Consulté le 21 décembre 2017]. Disponible à l'adresse : https://ppc-insider.ch/wie-twitter-die-schweiz-bewegt/.

POTTING, J., 2016. *Completeness in Twitter datasets : A critical review on Twitter research methodologies* [en ligne]. Master thesis. Utrecht : Utrecht University. [Consulté le 28 novembre 2017]. Disponible à l'adresse : http://dspace.library.uu.nl/handle/1874/338956. Utrecht University Repository

RAJABIFARD, Abbas, LAYLAVI, Farhad et KALANTARI, Mohsen, 2016. Hide your location on Twitter? We can still find you and that's not a bad thing in an emergency. In : *The Conversation* [en ligne]. 23 mai 2016. [Consulté le 5 janvier 2018]. Disponible à l'adresse : http://theconversation.com/hide-your-location-on-twitter-we-can-still-find-you-and-thats-not-a-bad-thing-in-an-emergency-58649.

RAUCHFLEISCH, Adrian et METAG, Julia, 2016. The special case of Switzerland: Swiss politicians on Twitter. In : *new media & society* [en ligne]. 2016. Vol. 18, n° 10, p. 2413–2431. [Consulté le 29 décembre 2017]. Disponible à l'adresse : http://journals.sagepub.com/doi/pdf/10.1177/1461444815586982.

RAYMOND, Matt, 2010. How Tweet It Is!: Library Acquires Entire Twitter Archive. In : *Library of Congress Blog* [en ligne]. 14 avril 2010. [Consulté le 28 novembre 2017]. Disponible à l'adresse : //blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/.

RICKMANN, Andreas, 2017. Wie viele Nutzer Instagram, Facebook, Xing und Co. in Deutschland haben. In : *Andreas Rickmann* [en ligne]. 10 septembre 2017. [Consulté le 21 décembre 2017]. Disponible à l'adresse : http://andreasrickmann.de/2017/09/10/wie-viele-nutzer-instagram-facebook-xing-und-co-in-deutschland-haben/.

SLOAN, Luke et MORGAN, Jeffrey, 2015. Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. In : *PLoS ONE* [en ligne]. 6 novembre 2015. Vol. 10, n° 11. [Consulté le 5 janvier 2018]. Disponible à l'adresse : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4636345/.

STONE, Biz, 2010. Tweet Preservation. In : *blog.twitter.com* [en ligne]. 14 avril 2010. [Consulté le 28 novembre 2017]. Disponible à l'adresse : https://blog.twitter.com/official/en_us/a/2010/tweet-preservation.html.

SUBRAHMANIAN, V. S., AZARIA, Amos, DURST, Skylar, KAGAN, Vadim, GALSTYAN, Aram, LERMAN, Kristina, ZHU, Linhong, FERRARA, Emilio, FLAMMINI, Alessandro et MENCZER, Filippo, 2016. The DARPA Twitter Bot Challenge. In : *Computer*. juin 2016. Vol. 49, n° 6, p. 38‑46. DOI 10.1109/MC.2016.183.

SULLIVAN, Danny, 2011. As Deal With Twitter Expires, Google Realtime Search Goes Offline. In : *Search Engine Land* [en ligne]. 4 juillet 2011. [Consulté le 29 décembre 2017]. Disponible à l'adresse : https://searchengineland.com/as-deal-with-twitter-expires-google-realtime-search-goes-offline-84175.

SWISSTOPO, 2017. Répertoire officiel des localités : LV03 / MN03 [fichier CSV]. In : *cadastre.ch* [en ligne]. 2017. [Consulté le 11 janvier 2018]. Disponible à l'adresse : https://www.cadastre.ch/content/cadastre-internet/fr/services/service/plz.html.

THOMSON, Sara Day, 2017. Preserving Social Media: applying principles of digital preservation to social media archiving. In : *IIPC Web Archiving Conference (WAC) 14-16 June 2017* [en ligne]. London : International Internet Preservation Consortium. juin 2017. [Consulté le 2 janvier 2018]. Disponible à l'adresse :

https://archivedweb.blogs.sas.ac.uk/files/2017/06/RESAW2017-Thomson-applying_principles_of_digital_preservation_to_social_media_archiving.pdf.

@TM, 2015. Evaluating language identification performance. In : *blog.twitter.com/engineering* [en ligne]. 16 novembre 2015. [Consulté le 22 décembre 2017]. Disponible à l'adresse : https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance.html.

TROMBLE, Rebekah, STORZ, Andreas et STOCKMANN, Daniela, 2017. We Don't Know What We Don't Know: When and How the Use of Twitter's Public APIs Biases Scientific Inference. In : *ssrn.com* [en ligne]. 29 novembre 2017. [Consulté le 27 décembre 2017]. Disponible à l'adresse : https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3079927_code2459643.pdf.

TSOU, Ming-Hsiang, ZHANG, Hao et JUNG, Chin-Te, 2017. Identifying Data Noises, User Biases, and System Errors in Geo-tagged Twitter Messages (Tweets). In : *arxiv.org* [en ligne]. 6 décembre 2017. [Consulté le 27 décembre 2017]. Disponible à l'adresse : https://arxiv.org/ftp/arxiv/papers/1712/1712.02433.pdf.

TWITTER INC., 2017a. Developer Agreement and Policy. In : *developer.twitter.com* [en ligne]. 3 novembre 2017. [Consulté le 8 janvier 2018]. Disponible à l'adresse : https://developer.twitter.com/en/developer-terms/agreement-and-policy.

TWITTER INC., 2017b. Selected Company Metrics and Financials. In : *Twitter Investor Relations* [en ligne]. Third quarter 2017. [Consulté le 6 novembre 2017]. Disponible à l'adresse : https://investor.twitterinc.com/results.cfm.

VAROL, Onur, FERRARA, Emilio, DAVIS, Clayton A., MENCZER, Filippo et FLAMMINI, Alessandro, 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In : *arXiv:1703.03107 [cs]* [en ligne]. 8 mars 2017. [Consulté le 28 octobre 2017]. Disponible à l'adresse : http://arxiv.org/abs/1703.03107.

WEIDEMANN, C. et SWIFT, Jennifer N., 2013. Social media location intelligence: The next privacy battle - An ArcGIS add-in and analysis of geospatial data collected from Twitter.com. In : *International Journal of Geoinformatics* [en ligne]. juin 2013. Vol. 9, n° no 2, p. 21‑27. [Consulté le 29 décembre 2017]. Disponible à l'adresse : https://www.researchgate.net/publication/288777678.

ZIMMER, Michael, 2015. The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. In : *First Monday* [en ligne]. 6 juillet 2015. Vol. 20, n° 7. [Consulté le 17 octobre 2017]. DOI 10.5210/fm.v20i7.5619. Disponible à l'adresse : http://firstmonday.org/ojs/index.php/fm/article/view/5619.

Annexe 1 : Techniques de localisation

La localisation géographique sur Twitter se fait de deux manières différentes : soit l'utilisateur active le GPS de son appareil, soit il choisit manuellement un lieu parmi ceux proposés. Ce lieu est ensuite encodé selon les coordonnées géographiques.

La géolocalisation n'est donc pas activée par défaut, comme l'indique clairement Twitter dans ses conditions générales[57] :

> *« […] vous pouvez choisir de publier votre localisation dans vos Tweets et sur votre profil Twitter. Vous pouvez aussi nous communiquer votre localisation actuelle en la renseignant sur Twitter.com. Nous pouvons également déterminer votre localisation en utilisant les autres données de votre appareil, telles que les informations de localisation précises de votre GPS, les informations concernant les réseaux sans fil ou les antennes-relais à proximité de votre appareil mobile, ou votre adresse IP. »*

L'action de localiser ses tweets est donc de la responsabilité de la personne qui twitte, généralement appelée « twittos ». Il peut choisir de le faire pour toutes ses publications, ou seulement pour certaines[58]. Cependant, certains procédés automatiques peuvent également le faire sans que l'utilisateur en soit complétement conscient : « […] *une fois que vous avez publié un Tweet avec localisation, vos Tweets suivants incluent automatiquement une localisation générale.* […] [59]»

La localisation déclarée par le twittos peut être complétée par des données de localisation précises via le GPS de l'appareil utilisé :

> *« Si vous appuyez sur l'icône de localisation lorsque vous composez votre Tweet et que vous activez l'option permettant d'indiquer votre localisation précise,* ***votre Tweet inclura à la fois l'étiquette de localisation de votre choix et la localisation précise de votre appareil (latitude et longitude),*** *qui peut être trouvée via l'API.*
>
> *Si vous choisissez d'activer le bouton Partager la localisation exacte (disponible sur Twitter pour iOS version 6.26 ou ultérieure et sur Twitter pour Android version 5.55 ou ultérieure), votre localisation précise (latitude et longitude) sera associée au Tweet et pourra être trouvée via l'API.*
>
> *Si vous tweetez depuis une version antérieure de Twitter pour iOS ou Android, tous vos Tweets géolocalisés contiendront la localisation précise de votre appareil (latitude et longitude), qui peut être trouvée via l'API. »*[60]

Twitter indique clairement, pour le twittos attentif qui lira les conditions générales et les FAQ mises à disposition, les dangers de la localisation et comment l'effacer :

> *« Vous pouvez* ***effacer vos données de localisation de vos Tweets passés*** *en une seule étape (consultez cet article pour des instructions détaillées).*
>
> *Soyez attentif à la quantité d'informations que vous partagez en ligne. Vous pouvez souhaiter partager vos informations de localisation pour certaines actualités ("Le défilé commence" ou "Un camion vient de renverser plein de friandises sur la route !"), mais pas pour d'autres.* ***Vous ne souhaitez pas forcément communiquer l'adresse de***

[57] https://twitter.com/privacy?lang=fr

[58] De plus, cela peut dépendre de l'appareil utilisé : un même twittos peut être localisé sur un smartphone, mais pas depuis un ordinateur.

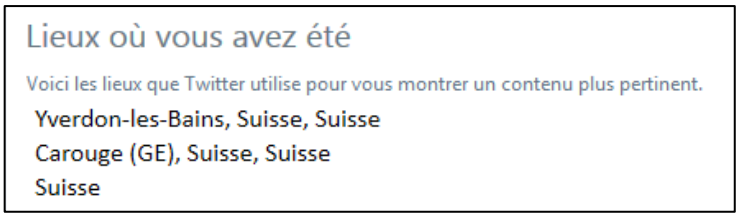
[59] https://support.twitter.com/articles/264388#

[60] Cf note 59

votre domicile. De la même façon, faites attention lorsque vous tweetez depuis des lieux que vous ne désirez pas rendre publics. »[61] [la mise en gras est d'origine]

En plus de ces localisations qui doivent être faites de manière volontaire, Twitter enregistre des paramètres de pays sur les comptes utilisateurs, indiqués au niveau de la ville :

Figure 28 : Localisation d'un compte Twitter



« Votre compte Twitter est associé au pays où vous vivez. Votre pays nous aide à personnaliser votre expérience Twitter et peut affecter le contenu que nous sommes en mesure d'afficher. Vous pouvez voir et modifier votre paramètre de pays dans vos paramètres de compte sur twitter.com, iOS et Android. »

Les informations sont tirées « de l'adresse IP, des données de localisation GPS précises ou des informations sur les réseaux sans fil ou les antennes relais se trouvant à proximité de l'appareil mobile. Cependant, à la différence de la localisation du [sic] votre profil qui apparaît dans le profil de compte public, et qui est entièrement facultative, le paramètre de pays n'est pas une information publique. Il sert à « personnaliser l'expérience Twitter » et peut affecter le contenu affiché »[62].

[61] Cf note 59
[62] https://help.twitter.com/fr/managing-your-account/how-to-change-country-settings

Annexe 2 : Exemples de tweets

Contenu

Le texte des tweets récoltés est très divers : des personnes échangent des informations personnelles (exemples 3, 4, 6 et 8), d'autres sont plutôt actives professionnellement (exemple 5). Des robots envoient également régulièrement des messages : météo (exemple 7), '*Trending topics*' (exemple 2), radars, ... Les langues ne sont pas toujours bien reconnues (exemples 1, 2, 3 et 7).

	Langue	Localité	Tweet
1	Polonais	Kloten	LSZH 172350Z VRB01KT 1500 0700S R14/0900N R16/0650U R28/P2000N R34/P2000N BCFG FEW001 SCT050 M00/M00 Q1027 TEMPO 0300
2	Tchèque	Engelberg	1. #srfarena 2. #TPMPTouteLaVerite 3. Problem 4. #SCBern 5. #EHCBGSHC
3	Gallois	Zurich	I'm at Abaton in Zürich, ZH https://t.co/Gyxirm85L3b
4	Français	Chaux-de-Fonds	Jspr vrmt que c une blague
5	Anglais	Berne	Resilience is the most trustworthy skill, only it can save you from your dangerous self. #leadership #Entrepreneur
6	Anglais	Genève	Omg 😂😍❤❤❤❤❤
7	Indonésien	Rupperswil	00:46 Temp. 4.5°C, Hum. 82%, Dewp. 0.8°C, Bar. 1027 hpa, Wind 206° 1.0 km/h
8	Français	Pully	Le retour à la normale sur la ligne #cff régionale Lausanne-Pully est prévue pour lundi matin #éboulement

Quelques exemples choisis parmi les tweets récoltés

Quelques tweets envoyés pour tester la localisation et la fiabilité des langues

9	Anglais	Vernier, Suisse	Home sweet home, WP
10	Anglais	Zurich, Suisse	Hund, chien. Dog
11	Anglais	Zurich, Suisse	pain chocolat
12	Français	Zurich, Suisse	Pain au chocolat.
13	Danois	Carouge (GE), Suisse	kgjfkld

Annexe 3 : Métadonnées retenues ou exclues

Tableau 6 : 30 champs retenus

_index _source.created_at _source.source _source.geo.coordinates _source.id_str _source.text _source.place.country_code _source.place.country _source.place.full_name _source.place.bounding_box.coordinates _source.place.place_type _source.place.name _source.place.id _source.lang _source.coordinates.coordinates	_source.entities.hashtags _source.location.lon _source.location.lat _source.user.utc_offset _source.user.friends_count _source.user.created_at _source.user.protected _source.user.screen_name _source.user.id_str _source.user.lang _source.user.verified _source.user.statuses_count _source.user.followers_count _source.user.name _source.user.location

Tableau 7 : Champs exclus

_id _score _source.extended_entities.media _source.in_reply_to_status_id_str _source.in_reply_to_status_id _source.in_reply_to_user_id_str _source.retweet_count _source.retweeted _source.geo.type _source.filter_level _source.in_reply_to_screen_name _source.is_quote_status _source.tmp _source.in_reply_to_user_id _source.@version _source.favorite_count _source.id _source.place.bounding_box.type _source.place.attributes _source.place.url _source.favorited _source.possibly_sensitive _source.coordinates.type _source.truncated _source.timestamp_ms _source.@timestamp _source.entities.urls _source.entities.media _source.entities.user_mentions	_source.entities.symbols _source.display_text_range _source.contributors _source.user.profile_image_url_https _source.user.listed_count _source.user.profile_background_image_url _source.user.default_profile_image _source.user.favourites_count _source.user.description _source.user.is_translator _source.user.profile_background_image_url_https _source.user.profile_link_color _source.user.id _source.user.geo_enabled _source.user.profile_background_color _source.user.profile_sidebar_border_color _source.user.profile_text_color _source.user.profile_image_url _source.user.time_zone _source.user.url _source.user.contributors_enabled _source.user.profile_background_tile _source.user.follow_request_sent _source.user.profile_use_background_image _source.user.default_profile _source.user.following _source.user.profile_sidebar_fill_color _source.user.notifications

Parmi les champs exclus, nous en avons cependant utilisés certains depuis Kibana : source.id, source.user.id, source_user.geo_enabled,, et source.user-time_zone.

Annexe 4 : Fichier de concordance

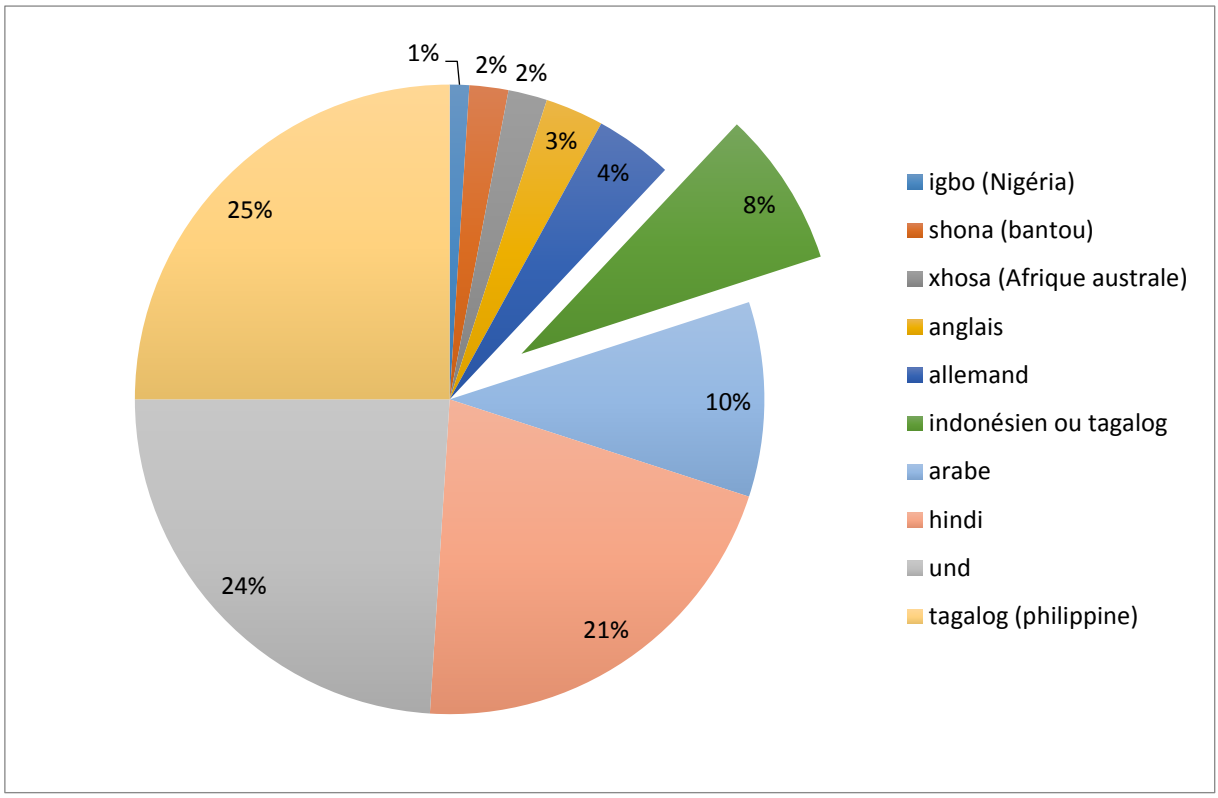
place.full_name.keyword: Descending	place.id.keyword: Descending	Localité (manuelle) en bleu les différences de langue	Canton	no OFS communes liste historisée	No OFS communes actuelles	Nom communes actuelles	Population résidente 2016	code postal Cadastre.ch	coordonnée Est	coordonnée Nord
Aadorf, Schweiz	d29b524faece2f23	Aadorf	Turgovie	4551	4551	Aadorf	8865	8355	709949	261089
Aarau, Schweiz	0e0b7f0796a22e3a	Aarau	Argovie	4001	4001	Aarau	21036	5000	646060	248866
Aarberg, Schweiz	9cee1db6d75d6bdf	Aarberg	Berne	301	301	Aarberg	4527	3270	587587	210292
Aarburg, Schweiz	79248c3b36c1d8ca	Aarburg	Argovie	4271	4271	Aarburg	7854	4663	634825	241201
Aarwangen, Schweiz	1e1380d77054dc92	Aarwangen	Berne	321	321	Aarwangen	4518	4912	624590	232268
Abtwil, Schweiz	bd4bb780492064c5	Abtwil	Argovie	4221	4221	Abtwil	997	5646	669471	225299
Aclens, Suisse	fb185824833c049a	Aclens	Vaud	5621	5621	Aclens	521	1123	528792	157797
Acquarossa, Svizzera	c3f8269d28ee180b	Acquarossa	Tessin	5048	5048	Acquarossa	1850	6716	715114	146055
Adelboden, Switzerland	e22405b0993ab18d	Adelboden	Berne	561	561	Adelboden	3370	3715	609227	149073
Adligenswil, Schweiz	1a4ff3d6a1b96e68	Adligenswil	Lucerne	1051	1051	Adligenswil	5352	6043	670369	213699
Adlikon, Schweiz	d5d97a95855eaff0	Adlikon	Zurich	21	21	Adlikon	666	8452	694409	270774
Adliswil, Schweiz	04c60b5de694cc1f	Adliswil	Zurich	131	131	Adliswil	18742	8134	681938	240784
Aedermannsdorf, Schweiz	3f1b7a8ca1cc7f1e	Aedermannsdorf	Soleure	2421	2421	Aedermannsdorf	556	4714	612829	239205
Aefligen, Schweiz	8f022a7eb6e80ed5	Aefligen	Berne	401	401	Aefligen	1089	3426	608588	216085
Aegerten, Schweiz	2fa825a85251e634	Aegerten	Berne	731	731	Aegerten	2011	2558	588846	218917
Aesch (BL), Schweiz	1067fd7acdcc5620	Aesch (BL)	Bâle-Campagne	2761	2761	Aesch (BL)	10184	4147	611908	257353
Aesch (LU), Schweiz	34368e9fd53b5398	Aesch (LU)	Lucerne	1021	1021	Aesch (LU)	1141	6287	660592	234064
Aesch (ZH), Schweiz	4375b927b81ee644	Aesch (ZH)	Zurich	241	241	Aesch (ZH)	1275	8904	675546	243425
Aeschi (SO), Schweiz	f4ba724a1a8b4ab5	Aeschi (SO)	Soleure	2511	2511	Aeschi (SO)	1213	4556	617223	225346
Aeschi bei Spiez, Schweiz	4aa96e844c378b2e	Aeschi bei Spiez	Berne	562	562	Aeschi bei Spiez	2232	3703	619578	167663
Aetigkofen, Schweiz	adff40d6c258c4dc	Aetigkofen	Soleure	2441	2465	Buchegg	2532	4583	602128	219137

Annexe 5 : Tests de langues

Tableau 8 : Langues détectées par Twitter dans l'échantillon du test (TL1)

Langue détectée par twitter	Nombre de tweets	Nombre de fausse détection	Proportion de fausse détection
Nl : hollandais	25	25	100%
In : indonésien	13	13	100%
Ht : créol	4	4	100%
Cy : gallois	3	3	100%
Et : estonien	3	3	100%
Da : danois	2	2	100%
Fi : finnois	2	2	100%
Pl : polonais	2	2	100%
Cs : tchèque	1	1	100%
Tl : tagalog	3	1	33%
Sl : slovène	4	1	25%
En : anglais	293	47	16%
De : allemand	82	12	15%
Es : espagnol	50	5	10%
Fr : français	132	11	8%
Pt : portugais	60	5	8%
Und : indéterminé	130	7	5%
It : Italien	66	3	5%
Ar : arabe	91	0	0%
Tr : turque	18	0	0%
Ja : japonais	5	0	0%
Pa : panjabi	3	0	0%
Ru : russe	3	0	0%
Sv : suédois	2	0	0%
Th : thailandais	2	0	0%
Ta : tamoul	1	0	0%

Figure 29 : Réattribution des langues "in" (TL2)



Annexe 6 : Listes des *user.lang* et des *source.lang* de notre échantillon

User.lang	Proportion des comptes	Nb de comptes
en	50,81%	33’218
de	14,86%	9’718
fr	13,15%	8’597
es	5,53%	3’618
it	4,59%	3’001
pt	2,01%	1’317
ar	1,73%	1’128
tr	1,29%	845
en-gb	1,25%	814
nl	1,06%	694
ja	0,92%	603
ru	0,79%	519
id	0,26%	171
sv	0,20%	129
ca	0,18%	117
pl	0,16%	106
th	0,16%	105
fi	0,13%	88
zh-cn	0,13%	85
ko	0,11%	72
cs	0,10%	67
da	0,07%	43
el	0,06%	40
hu	0,06%	38
no	0,05%	32
sr	0,04%	29
ro	0,04%	25
zh-tw	0,03%	21
hr	0,03%	20
he	0,03%	17
zh-Hans	0,02%	14
uk	0,02%	13
sk	0,02%	11
bg	0,01%	8
fa	0,01%	7

Source.lang	proportion des tweets	Nb des tweets
en	35,909%	373’385
fr	15,958%	165’932
de	13,923%	144’775
und	11,399%	118’524
es	4,210%	43’772
pt	3,282%	34’131
it	2,747%	28’560
ar	2,544%	26’448
tr	2,124%	22’086
ja	1,277%	13’276
in	1,145%	11’906
nl	0,806%	8’382
ru	0,711%	7’396
tl	0,495%	5’145
th	0,432%	4’488
sv	0,425%	4’420
et	0,305%	3’168
ht	0,298%	3’097
fi	0,256%	2’657
da	0,255%	2’649
ko	0,198%	2’059
pl	0,196%	2’036
ro	0,159%	1’654
sl	0,134%	1’390
no	0,122%	1’269
cy	0,109%	1’137
cs	0,078%	807
hi	0,068%	704
lt	0,058%	601
eu	0,051%	527
hu	0,048%	503
is	0,040%	419
zh	0,040%	411
el	0,037%	388
lv	0,035%	362

gl	0,01%	7
pt-PT	0,01%	7
lv	0,01%	4
msa	0,01%	4
fil	0,005%	3
es-MX	0,005%	3
fr-CA	0,003%	2
xx-lc	0,003%	2
vi	0,003%	2
eu	0,003%	2
de-CH	0,002%	1
sq	0,002%	1
bn	0,002%	1
sr-Latn	0,002%	1
nb	0,002%	1
en-AU	0,002%	1
fr-CH	0,002%	1
ta	0,002%	1
hi	0,002%	1
mk	0,002%	1
gsw	0,002%	1
ga	0,002%	1

fa	0,033%	339
sr	0,019%	194
uk	0,013%	136
iw	0,013%	132
vi	0,012%	129
ne	0,012%	121
bg	0,011%	117
ur	0,008%	82
ta	0,004%	46
si	0,002%	19
dv	0,001%	8
ps	0,001%	7
bo	0,0005%	5
am	0,0005%	5
pa	0,0005%	5
ka	0,0003%	3
mr	0,0002%	2
ml	0,0001%	1
ckb	0,0001%	1
or	0,0001%	1
my	0,0001%	1
bn	0,0001%	1

Annexe 7 : Répartition des tweets et twittos par canton

Tableau 9 : Répartition de la population résidente de plus de 15 ans (2016) par canton, comparée à la proportion des tweets et des twittos

Cantons	**% population résidente**	**% twittos**	**% tweets**
AG, Argovie	7.83%	3.10%	3.21%
AI, Appenzell-Intérieur	0.19%	0.20%	0.12%
AR, Appenzell-Extérieur	0.65%	0.20%	0.14%
BE, Berne	12.26%	11.93%	8.92%
BL, Bâle-Campagne	3.41%	1.39%	1.20%
BS, Bâle-Ville	2.33%	2.95%	4.06%
FR, Fribourg	3.62%	3.37%	2.03%
GE, Genève	5.75%	13.96%	19.65%
GL, Glaris	0.48%	0.26%	0.07%
GR, Grisons	2.39%	3.74%	1.80%
JU, Jura	0.86%	0.40%	0.28%
LU, Lucerne	4.77%	4.00%	3.52%
NE, Neuchâtel	2.10%	1.36%	1.88%
NW, Nidwald	0.51%	0.69%	0.37%
OW, Obwald	0.44%	1.96%	3.22%
SG, St-Gall	5.94%	2.53%	2.54%
SH, Schaffhouse	0.97%	0.63%	0.55%
SO, Soleure	3.22%	0.97%	1.50%
Sz, Schwytz	1.85%	1.51%	0.44%
TG, Turgovie	3.20%	1.14%	0.79%
TI, Tessin	4.27%	5.22%	3.62%
UR, Uri	0.43%	0.63%	0.17%
VD, Vaud	9.18%	11.50%	13.25%
VS, Valais	4.03%	5.66%	4.70%
ZG, Zoug	1.46%	1.01%	1.04%
ZH, Zurich	17.85%	18.30%	18.69%

Annexe 8 : Tableaux des répartitions mensuelles des tweets, sans et avec correction des valeurs

Tableau 10 : Répartition temporelle des tweets par mois sans correction des valeurs

	Période	**Février**	**Mars**	**Avril**	**Mai**	**Juin**	**Juillet**	**Août**
jours	192	11	31	30	28	30	31	31
moyenne	5416	5’414	5328	5023	5104	5335	5411	6248
médiane	5386	5’471	5375	4981	5277	5325	5423	6385
min	800	4’836	800*	4536	2289	4656	4728	4577
max	7721	5’577	6’848	5700	6391	6057	6291	7721
écart type	742	270	999	310	870	351	415	664
écart moyen	494	196	536	255	560	282	329	529
Total	1039819	59’551	165180	150676	142919	160062	167739	193692

Tableau 11 : Répartition temporelle des tweets par mois après correction des valeurs

	Période	**Février**	**Mars**	**Avril**	**Mai**	**Juin**	**Juillet**	**Août**
jours	195	11	31	30	31	30	31	31
moyenne	5471	5414	5479	5023	5334	5335	5411	6248
médiane	5386	5471	5401	4981	5217	5325	5423	6385
min	4307	4836	4307	4536	4779	4656	4728	4577
max	7721	5577	6848	5700	6391	6057	6291	7721
écart type	577	269.9	539	310	378	351	415	664
écart moyen	438	195.9	391	255	273	282	329	529
total	1066929	59551	169862	150676	165347	160062	167739	193692

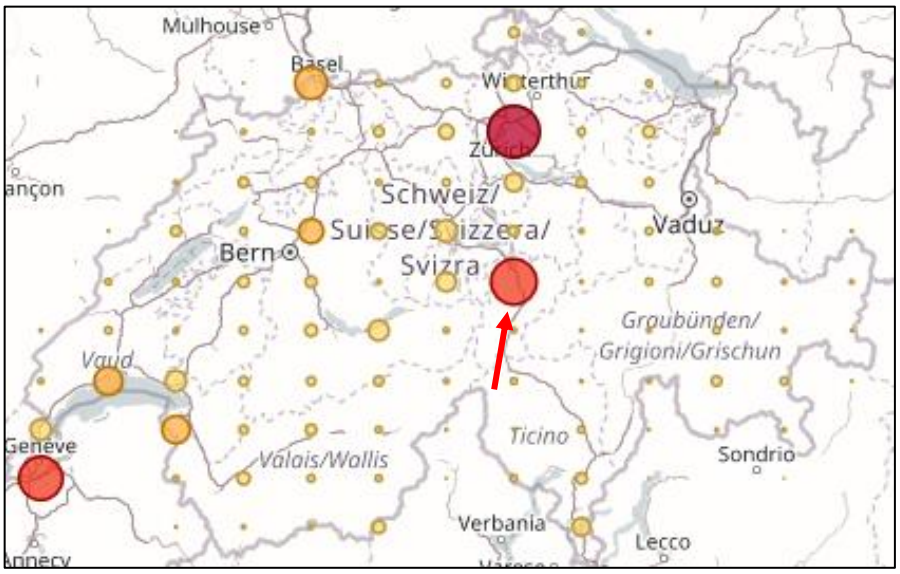
Petites questions et réponses

Durant notre recherche, nous avons trouvé des anomalies, ou cherché la réponse à certaines questions. Comme les résultats sont intéressants mais n'ont pas été intégrés dans le rapport, nous les avons indiqués ici.

Point central Suisse

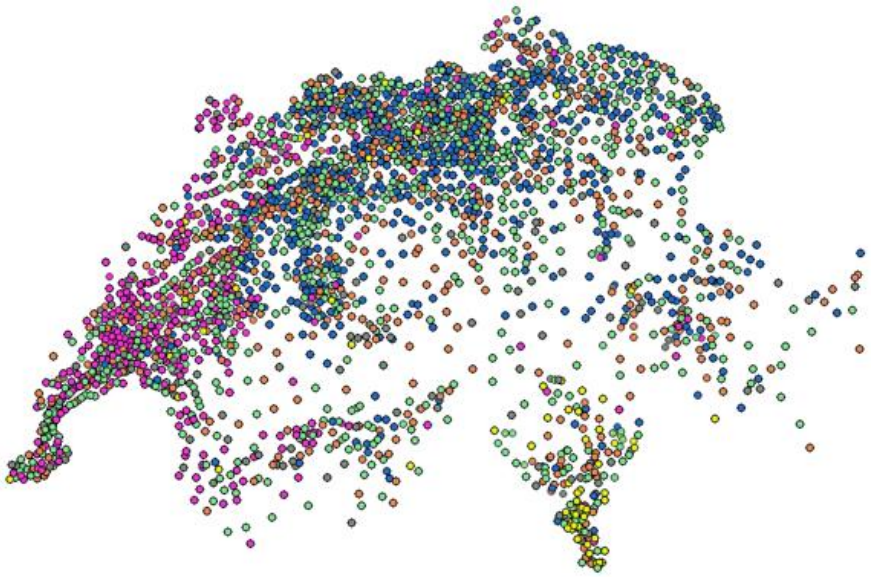
Lors des premières explications sur Kibana, nous avions constaté qu'un gros point, signifiant de nombreux messages, apparaissait au centre de la Suisse. Ce point était presque aussi grand que celui de Genève.

Figure 30 : Proportion de tweets envoyés en Suisse



Par contre, sur la carte des simples points géographiques réalisée pour le poster, ce point central n'apparaissait pas :

Figure 31 : Points géographiques des tweets



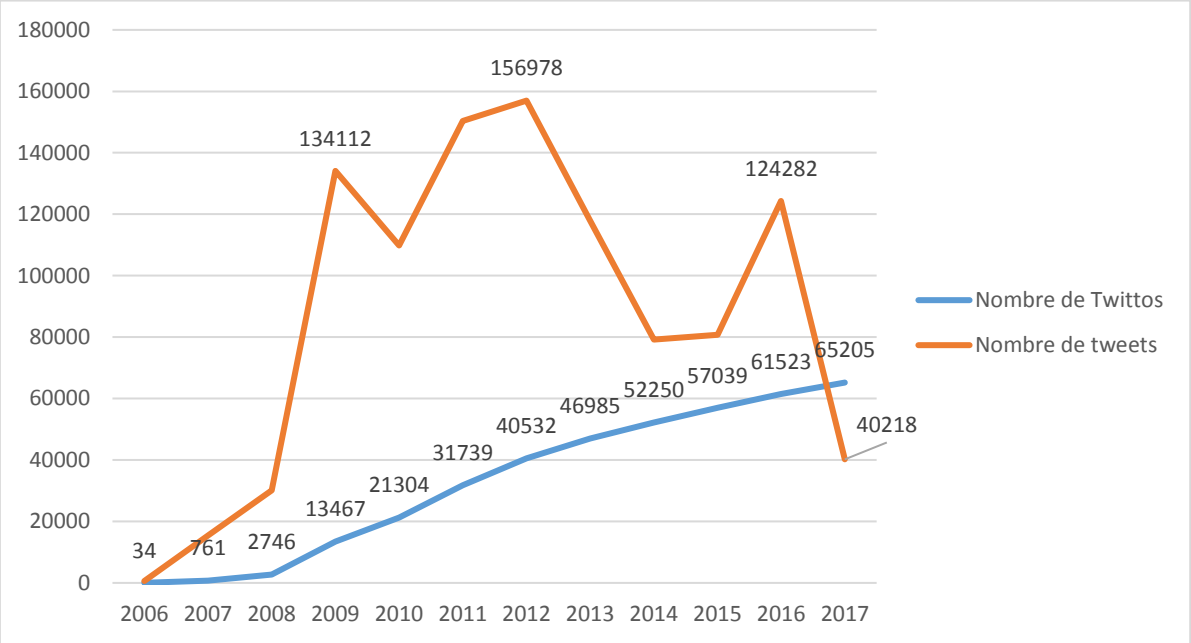
Carte créée avec le logiciel QGIS 2.18

Après investigation, il s'avère que ce point représente bien un grand nombre de messages, envoyés non par des touristes, ou des personnes ayant simplement localisé « Suisse », mais par trois robots, dont notre plus grands twittos (26'719 messages à lui tout seul). Les coordonnées géographiques étaient toujours les mêmes, c'est pourquoi ils n'apparaissent pas sur la deuxième carte !

Dates de création des comptes et tweets relatifs

Twitter existe depuis 2006, et nous avons vérifié l'ancienneté de nos twittos pour voir si cela avait une signification par rapport au nombre de tweets envoyés.

Figure 32 : Activité des twittos selon leur ancienneté



Note : malgré plusieurs vérifications, le nombre de twittos extraits pour cette vérification était de 65'205 et non 65'221.

Parmi notre échantillon de messages géolocalisés, les 'anciens' *twittos* inscrits en 2009, 2011 et 2012 sont encore très actifs. Par contre, les inscrits 2017 ont encore peu twitté.

Notre robot grand *twittos* représente 21% des tweets des comptes 2016.

Quels sont les jours de la semaine les plus propices à Twitter?

Les jours manquants ou partiels de collecte fausseraient les analyses concernant la répartition des tweets par jours de la semaine nous avons donc effectué les analyses de deux manières. D'abord sur les valeurs réelles mais excluant les mois de février et de mai de l'échantillon (Méthode 1). Ensuite sur les valeurs corrigées en substituant les moyennes journalières du mois aux valeurs manquantes ou tronquées (voir chapitre 3.1.2.2).

Le résultat global est identique pour les deux méthodes : le milieu de la semaine semble plus propice à Twitter que la fin de la semaine. Les moyennes de mercredi sont les plus élevées,

suivies par celles de mardi et de jeudi. Celles de samedi sont les plus basses précédées par celles du vendredi.

Figure 33 : Répartition des tweets par jour de la semaine (Méthode 1)

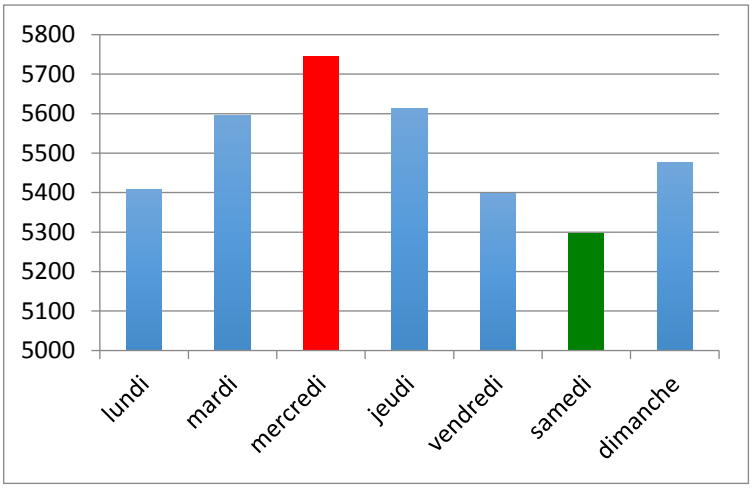


Figure 34 : Répartition des tweets par jour de la semaine (Méthode 2)