h e g

# GEoTweet : exploration des tweets géolocalisés à Genève

**Mémoire de recherche réalisé par :**

**Elisa BANFI**

**Fanny BÉGUELIN**

Sous la direction de :

**Arnaud GAUDINAT, adjoint scientifique HES**

**Carouge, le 18 janvier 2016**

**Master en Sciences de l'information**

**Haute École de Gestion de Genève (HEG-GE)**

Hes·so GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

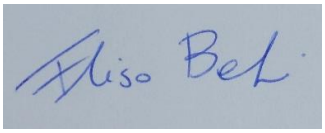
# Déclaration

Ce mémoire de recherche est réalisé dans le cadre du Master en Sciences de l'information de la Haute école de gestion de Genève. Les étudiants acceptent, 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 des auteurs, ni celle de l'encadrant.

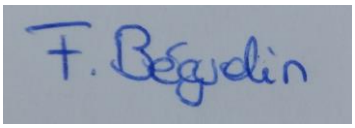
« Nous attestons avoir réalisé le présent travail sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Carouge, le 18 janvier 2016

Elisa Banfi



Fanny Béguelin



# Remerciements

Nous adressons nos chaleureux remerciements à toutes les personnes qui nous ont aidées à mener à bien ce travail, à savoir :

Romaine Kaufmann, qui a fait partie intégrante de l'équipe GEoTweet pour la préparation de la bibliographie commentée durant les cinq premiers mois du projet,

Nadine Sharaf et Céline Piccand, membres du LTI (Laboratoire des technologies et de l'information), dont les conseils ont été très précieux pour l'élaboration de notre poster scientifique,

et bien entendu Arnaud Gaudinat, sans qui ce projet n'existerait pas, que nous souhaitons remercier pour sa disponibilité, son expertise, sa gentillesse et son humour, qui ont fait de ce projet une aventure très plaisante.

# Résumé

Que découvre-t-on sur la société genevoise quand on modélise des frontières au sein d'un réseau social tel que Twitter ?

Comment la visualisation des tweets géolocalisés nous permet-elle de saisir la variété des communautés linguistiques à Genève ?

Le projet GEoTweet propose des réponses à ces questions de recherche, à travers la mise en place d'une infrastructure permettant de récolter les tweets, de les explorer et de les interroger.

Il implémente des modalités techniques d'accès, de stockage et d'exploitation des tweets géolocalisés à partir du streaming public mis à disposition par Twitter via son API, explore les tweets géolocalisés dans un contexte spatio-temporel défini (Genève et ses alentours), et teste des formes de visualisation des données géolocalisées en considérant leur potentiel analytique.

Le premier axe de recherche concerne les frontières linguistiques virtuelles. Les tweets géolocalisés donnent des informations sur la population résidente et, en même temps, sur les flux volatiles d'autres catégories d'utilisateurs de l'espace urbain genevois.

Le deuxième axe de recherche concerne les convergences et les divergences entre les frontières géopolitiques et les frontières virtuelles, et met en évidence leur perméabilité et leur influence réciproque.

Le troisième axe de recherche concerne la qualité de données récoltées et leur fiabilité, à travers l'analyse de l'intégration entre différents media sociaux géolocalisés, les procédures d'élimination des robots et l'importance des sauvegardes pour le stockage.

Le projet GEoTweet ouvre des perspectives additionnelles d'exploration et d'analyse des données, et est le point de départ d'une série d'événements grand public visant à la vulgarisation du big data.

**Mots-clés :** Twitter, GEoTweet, big data, data visualisation, exploration des données, géolocalisation, Genève, Lausanne, frontières virtuelles, frontières linguistiques, CartoDB, Solr, données géolocalisées

# Table des matières

# Liste des tableaux



# Liste des figures

# 1. Introduction

Le projet GEoTweet se propose d'explorer les possibilités d'analyse et d'exploitation scientifique des tweets géolocalisés à Genève. Les professeurs Arnaud Gaudinat et Jean-Philippe Trabichet en qualité de membres du groupe BiTeM[1] ont contribué avec ce projet à « Mapping, l'événement HES » organisé par la HES-SO Genève.

Le projet GEoTweet développe la thématique « frontières et urbanité » qui est au cœur de l'évènement large public « Mapping HES ». Suivant les directives données par les organisateurs de cet événement, l'étude, intitulée GEoTweet, propose une réflexion sur la notion de frontière et de limite à l'aide de Twitter. Le projet de recherche implémente et analyse de façon critique les processus d'exploitation des tweets géolocalisés de la phase de la capture jusqu'à celle de la visualisation.

Ce travail de recherche intègre pleinement les objectifs du projet GEoTweet : d'une part, l'exigence analytique et théorique d'explorer les tweets géolocalisés, d'autre part, le souci de visualisation et de vulgarisation des données récoltées auprès du large public.

En fait, la visualisation n'est que la dernière étape du processus de manipulation des tweets. Les formes de visualisation dépendent non seulement de la phase de capture, mais également de celle du stockage et de l'exploration.

Figure 1 : Etapes précédant la visualisation des tweets géolocalisés


(Banfi 2016)

Or, le projet se concentre sur la valorisation du potentiel analytique des tweets géolocalisés à travers les outils de visualisation. Cette étude approfondit également les enjeux théoriques et empiriques liés à la capture, au stockage et à l'exploitation des données géolocalisées.

Eléments clés du projet :

- Durée : 14 mois (de février 2015 à avril 2016)
- Poster : décembre 2015
- Mémoire de recherche : janvier 2016
- Evénement HES : avril 2016

[1] Bibliomics and Text Mining Group, http://bitem.hesge.ch

# 2. Hypothèses

Le projet GEoTweet répond à différentes questions de recherche élaborées par les membres du BiTeM à l'aide de l'analyse des tweets récoltés.

Les deux premières questions concernent la modélisation des frontières au sein d'un réseau social tel que Twitter et leur nature virtuelle et physique :

1. Que découvre-t-on sur la société genevoise quand on modélise des frontières au sein d'un réseau social tel que Twitter ?

2. Comment modéliser ces frontières à la fois virtuelles et physiques ?

La troisième question concerne la collecte des informations et leurs visualisations :

3. Quelles sont les caractéristiques techniques de la collecte des tweets géolocalisés et de leurs visualisations ?

Elle résume principalement la préoccupation théorique et technique de vérifier la fiabilité des données obtenues.

Ce travail de recherche se structure autour neuf hypothèses qui découlent des trois questions de recherche susmentionnées.

Le premier groupe d'hypothèses reprend la première question en s'interrogeant sur les frontières linguistiques virtuelles, les communautés linguistiques géolocalisées et leur variation dans le temps et dans l'espace.

Le deuxième groupe d'hypothèses développe la deuxième question en réfléchissant aux convergences et divergences entre les frontières géopolitiques et les frontières virtuelles en analysant la distribution des langues dans le canton de Genève et entre certaines villes du Grand Genève et l'espace urbain interne au canton.

Le troisième groupe d'hypothèses concerne la qualité de données récoltées et leur fiabilité. Il regroupe des hypothèses qui s'intéressent à la présence de tweets émis par des robots et les problématiques liées au stockage et à l'intégration des tweets provenant d'autres médias sociaux.

## 2.1 Frontières linguistiques

La première hypothèse concernant les frontières linguistiques postule la diversité entre l'identité linguistique officielle de la ville et la variété des communautés virtuelles linguistiques qui y opèrent. Ces dernières demeurent néanmoins perméables en s'intégrant de façon structurelle à l'identité linguistique francophone de la ville.

H 1.1a : « 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 ».

A travers une deuxième hypothèse, nous suggérons la présence de communautés virtuelles linguistiques qui varient dans l'espace et dans le temps (horaires de la journée et jours fériés). Nous suggérons donc qu'à Genève, les facteurs spatio-temporels influencent les frontières linguistiques virtuelles de Twitter.

H 1.1b : « 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 ».

La visualisation interactive des données Twitter permet d'observer la variété linguistique de la ville de Genève et les modalités d'utilisation des espaces publics par les différentes communautés virtuelles linguistiques.

H 1.1c : « Les différentes communautés linguistiques virtuelles se différencient dans l'utilisation des espace publics d'où elles émettent des tweets ».

Afin de confronter la diversité linguistique de tweets géolocalisés à Genève avec un autre corpus, nous avons sélectionné un corpus de tweets géolocalisés à Lausanne et nous avons comparé les résultats des analyses.

H 1.1d : « Les différentes communautés linguistiques virtuelles géolocalisées dans la ville de Genève sont comparables à celles d'une ville avec le même profil sociodémographique comme la ville de Lausanne »

## 2.2 Frontières virtuelles et géopolitiques

Dans ce groupe d'hypothèses, les frontières virtuelles et géopolitiques genevoises sont comparées et leurs interactions analysées.

La frontière physique genevoise est très étroite, cependant l'influence économique, scientifique et culturelle de la ville rayonne à l'échelon régional. C'est pourquoi nous postulons que la résonance de Genève via Twitter dépasse ses frontières géopolitiques. Par exemple, la frontière virtuelle de Genève déborde en France en raison des travailleurs qui gravitent autour de l'économie du canton.

H 1.2a : « La variété de langues dans le Grand Genève diffère de celle du canton, mais, en même temps, est influencée par la nature internationale de celui-ci ».

Les frontières et les caractéristiques urbaines des quartiers de Genève influencent la densité et la distribution linguistique des tweets qui y sont produits. Nous supposons que, à l'échelle des quartiers, les tweets géolocalisés donnent l'opportunité de visualiser des frontières citoyennes en mouvement, notamment celles liées aux centres d'intérêts culturels et/ou professionnels des internautes.

H 1.2b : « Selon les quartiers, la densité et la diversité des langues des tweets varient ».

## 2.3 Qualité des données

Nous supposons que la qualité du corpus de tweets est influencée par différents facteurs. Premièrement, la qualité des données obtenues est dépendante du système de capture et de stockage.

H 1.3a : « La quantité de tweets dépend des paramètres de géolocalisation de Twitter et de la gestion des serveurs où les tweets sont stockés. La perte d'information liée à ces deux facteurs a des conséquences non anodines sur la fiabilité et la significativité de l'information obtenue ».

Les tweets provenant d'autres média sociaux doivent être considérés lors de l'analyse des caractéristiques comportementales des utilisateurs de Twitter.

H 1.3b : « Les tweets natifs de Twitter représentent seulement une partie des tweets géolocalisés récoltés »

La présence de robots produisant des tweets géolocalisés doit être analysée afin de nettoyer le corpus de données géolocalisées exploitables pour des recherches scientifiques portant sur les comportements et les attitudes des utilisateurs humains.

H 1.3c : « La présence de robots altère la quantité et la représentativité des tweets géolocalisés produits par des utilisateurs humains ».

# 3. Méthodologie

La méthodologie que nous avons employée pour ce projet est celle de la capture des tweets via streaming. Elle peut être résumée en cinq grandes étapes :

- Les utilisateurs envoient des tweets, géolocalisés ou non selon le paramétrage de leur application
- Twitter met à disposition ces données via son streaming public, auquel nous avons appliqué un filtre pour qu'il ne retourne que le flux de tweets liés à Genève (coordonnées géographiques, ville définie par l'utilisateur, zone géographique)
- Les tweets ainsi récupérés sont stockés et dupliqués sur trois serveurs, afin de garantir l'accès aux données et leur sauvegarde, et de pouvoir les comparer
- Une instance Solr par serveur est créée, afin d'interroger les données. Les requêtes effectuées dans cette interface permettent par exemple d'obtenir uniquement les tweets compris dans un rayon de 20 km autour de Genève, ou de supprimer les robots
- Les données sont visualisées et analysées à l'aide de différents outils : Banana, CartoDB et Microsoft Excel

## 3.1 Capture des tweets

Selon la littérature, les tweets géolocalisés représentent 1% des échanges (Hawelka et al. 2013). Nous entendons par « tweet géolocalisé » un tweet auquel est liée une coordonnée géographique (latitude, longitude). Dans le cadre de notre projet, le corpus de tweets genevois récoltés entre le 28 avril et le 26 octobre 2015 compte 2'303'297 tweets, parmi lesquels 48'160 sont géolocalisés, soit 2.09%.

Le streaming public à partir duquel nous récoltons les tweets est mis à disposition par Twitter via son API (Application Programming Interface). Cependant, l'intégralité des données de Twitter n'est pas accessible pour le grand public : officiellement, seul 1% du trafic de tweets est mis à disposition (Hawelka et al. 2013, Morstatter et al. 2013).

Cet échantillon de tweets sous forme de flux est émis selon des critères définis par l'utilisateur (par exemple la géolocalisation, ou les mots-clés). L'API de Twitter est utilisé par beaucoup de chercheurs pour analyser les données et les informations circulant sur Twitter, mais présente un défaut : en effet, il n'existe pratiquement pas de documentation concernant la quantité et la nature des données obtenues via l'API. Il est donc difficile de savoir si l'échantillon est représentatif de tout Twitter ou non.

Il existe cependant un flux contenant la totalité des tweets publics, le Firehose. Ce flux n'est pas accessible au grand public : seuls quelques partenaires de Twitter y ont accès, et cet accès est très coûteux tant en termes de prix qu'en termes d'infrastructures nécessaires (serveurs, espace disque, réseau).

La capture des tweets peut donc se faire avec deux outils différents : le streaming API, gratuit mais limité, et le Firehose, exhaustif mais très cher.

Une étude parue en 2013 (Morstatter et al., 2013) compare les données obtenues avec le streaming de l'API et celles obtenues avec le Firehose. Cette étude démontre que le streaming de l'API couvre en réalité bien souvent plus de tweets que les 1% annoncés, et qu'il est

possible d’augmenter le nombre de résultats obtenus en affinant les critères de recherche. De plus, il en ressort que les échantillons de tweets géolocalisés sont souvent exhaustifs même dans l’API, grâce au fait que seul un faible pourcentage des tweets est géolocalisé. C'est le cas pour notre étude, car la zone que nous couvrons (Culoz - Fribourg), est relativement petite, et ne génère donc pas une masse de tweets allant au-delà de la taille autorisée par Twitter pour la récolte. Même limitées par le 1% de l'API, nous obtenons donc un corpus exhaustif de tous les tweets genevois.

Les tweets récupérés contiennent non seulement le texte du tweet et le compte d'utilisateur lié, mais également un grand nombre d'autres métadonnées. 84 champs différents sont visualisables, tels que l'heure et la date de création du tweet, ses coordonnées géographiques, l'application source, le pays du compte utilisateur, le nombre de retweets, l'identifiant de l'utilisateur, la langue dans laquelle le compte Twitter est paramétré, ou la langue du texte lui-même.

C'est sur ce dernier paramètre que nous nous sommes basées pour l'analyse de la répartition linguistique des tweets. La langue est détectée automatiquement par Twitter, qui renseigne donc cette métadonnée dans son streaming public. Nous n'avons pas pu définir quantitativement la fiabilité de cette détection automatique, mais en observant le contenu des tweets nous avons constaté que dans la plupart des cas, la langue était correctement détectée. Lorsque le tweet ne contient que des émoticônes ou des abréviations non reconnaissables, la langue détectée est « und », soit « indéterminé ». Les tweets comprenant un mélange de différentes langues ou uniquement des URLs sont la plupart du temps répertoriés sous « ht », soit le créole haïtien. Ces inexactitudes dans la reconnaissance des langues peuvent créer un biais dans les statistiques concernant la variété linguistique des tweets.

## 3.2 Stockage

Les tweets obtenus via l'API étant sous la forme d'un flux continu de données, nous avons mis en place un système permettant d'enregistrer ces données sur des serveurs afin de constituer un corpus de tweets suffisant pour pouvoir les étudier.

Les données sont dupliquées sur trois serveurs : chaque serveur reçoit donc exactement les mêmes données, ce qui permet de les comparer et détecter les éventuelles irrégularités.

Au départ, nous ne travaillions que sur un serveur (A), hébergé à la HEG. Après environ quatre mois de projet, nous avons décidé de mettre en place un second serveur (B), hébergé sur un serveur externe, Kimsufi, afin de ne pas perdre les données en cas de panne globale sur l'un des serveurs. Puis environ un mois après, nous avons décidé d'ajouter un troisième serveur (C), car des coupures avaient eu lieu à plusieurs reprises sur les deux serveurs de base. Nous avons donc préféré avoir une troisième sauvegarde, afin de pouvoir combler les éventuels trous de données dans les autres serveurs. Finalement, le corpus de 2'303'297 de tweets sur lequel nous travaillons est extrait du serveur A, et nous n'avons pas eu besoin de combler les trous avec des données provenant d'autres serveurs. (Le chapitre 9. Problèmes techniques détaille ces problèmes de trous dans la capture). Les trois serveurs sont toujours actifs, car nous continuons à constituer des corpus de tweets genevois en vue d'analyses ultérieures.

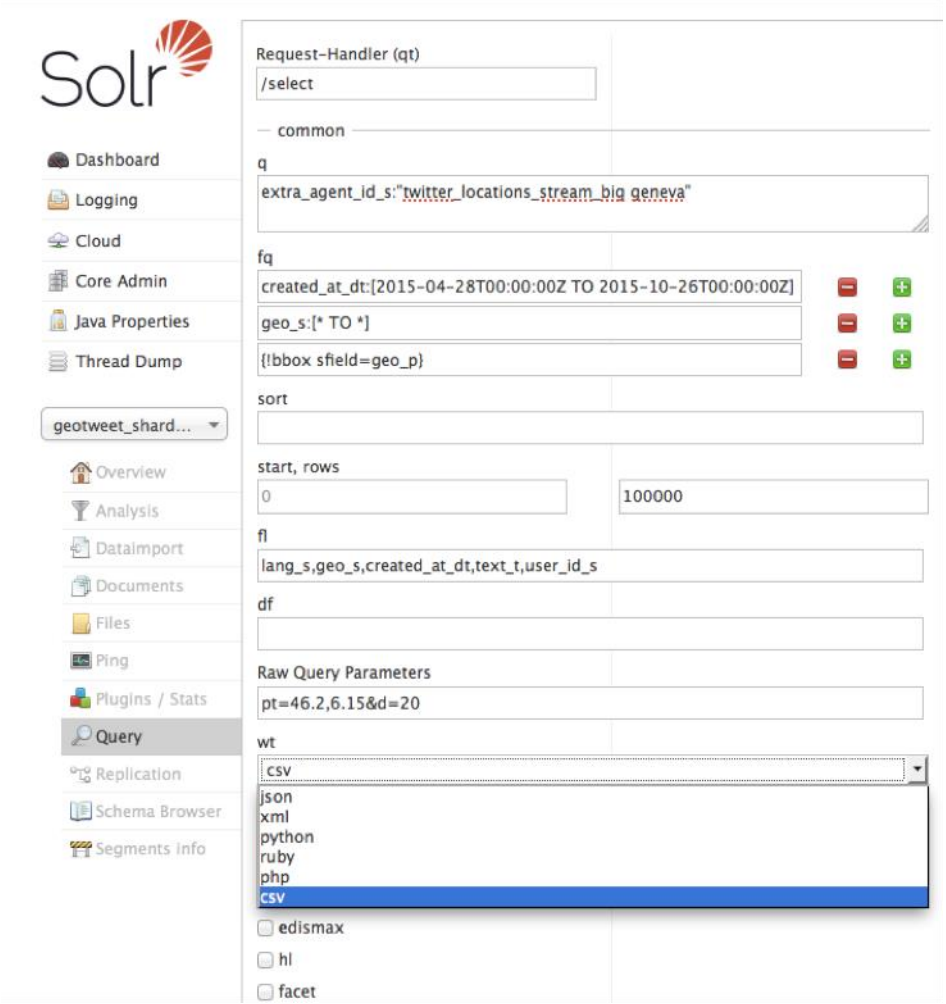
## 3.3 Accès aux données

L'interrogation des données présentes sur nos serveurs se fait avec Solr.

Apache Solr est une plateforme logicielle de moteur de recherche open source distribuée sous licence libre, créée par la Fondation Apache. Solr s'appuie sur la bibliothèque Lucene, écrite en Java, qui permet d'indexer et de rechercher du texte. Ce moteur de recherche permet d'exécuter des requêtes dynamiques précises, en temps réel. Il est utilisé par de nombreuses entreprises et institutions, de tous types.

Nous avons utilisé la requête suivante afin d'obtenir les tweets géolocalisés dans un rayon de 20 km autour du point central de Genève (entre la Place du Bourg-de-Four et la Promenade Saint-Antoine) :

Figure 2 : Interface de requête Solr



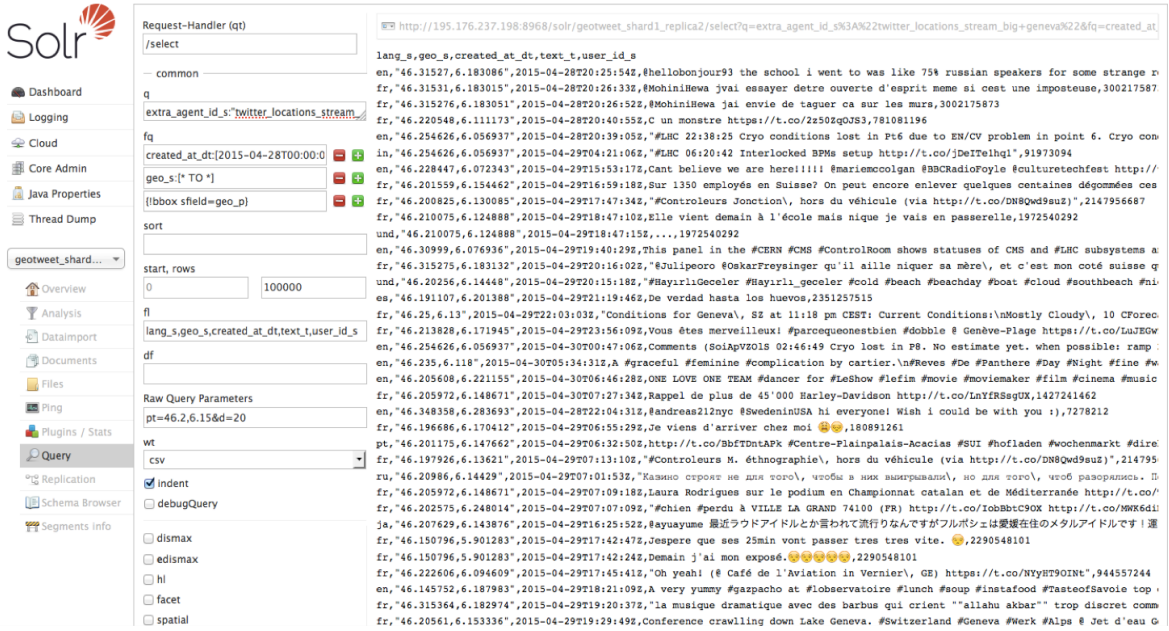

(Béguelin 2015)

- extra_agent_id_s:"twitter_location_stream_big_geneva" correspond à l'agent de surveillance défini dans le streaming de l'API, selon lequel nous filtrons les données pour les récupérer
- created_at_dt définit la plage temporelle que nous souhaitons (ici du 28 avril au 26 octobre 2015)
- geo_s:[* TO *] indique que nous souhaitons tous les tweets géolocalisés

- {!bbox sfield=geo_p} est complété par le Raw Query Parameters pt=46.2,6.15&d=20 et sert à indiquer le point géographique et le rayon dans lequel nous souhaitons récupérer les tweets
- les différents champs lang_s, geo_s, created_at_dt, text_t et user_id_s correspondent aux métadonnées que l'on souhaite voir affichées dans nos résultats
- le format choisi pour cet exemple est .csv, mais plusieurs formats d'exports sont proposés.

Le résultat s'affiche sous forme de tableau, avec un tweet par ligne et les métadonnées séparées en colonnes par une virgule. Nous pouvons alors exporter ce fichier pour analyser les données.

Figure 3 : Résultats d'une requête Solr




(Béguelin 2015)

## 3.4 Visualisation et analyse

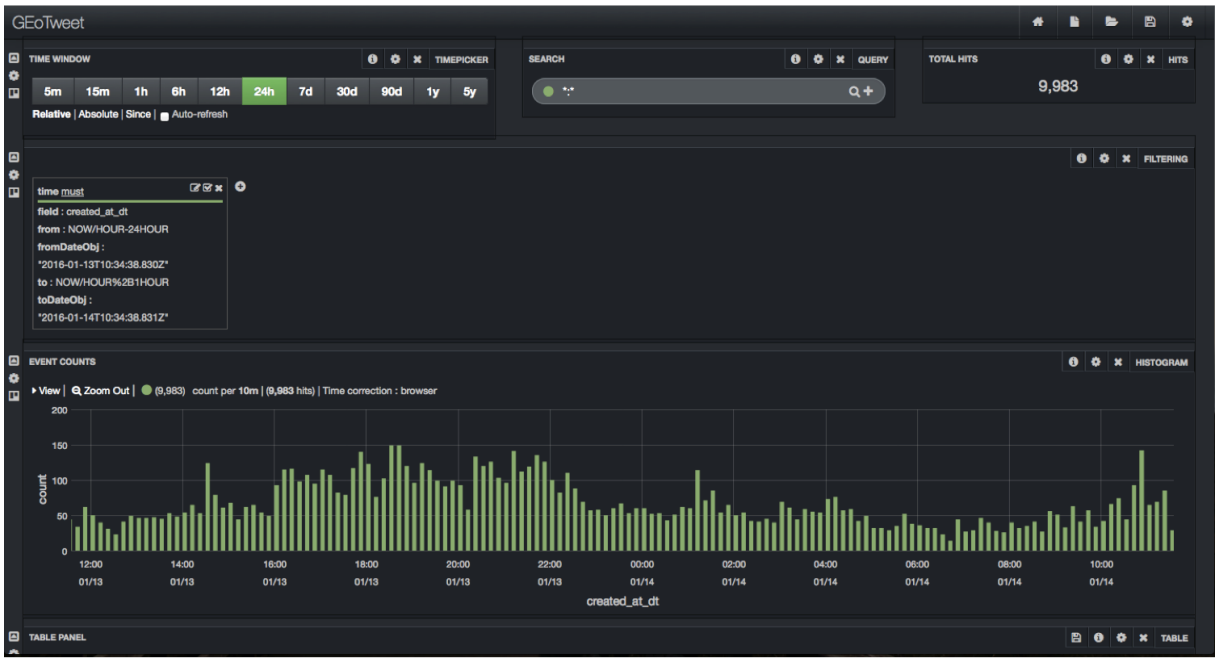
Nous avons utilisé différents outils pour visualiser et analyser les données.

### 3.4.1 Banana

Le tableau de bord Banana, branche de Kibana, est développé par Apache et basé sur un navigateur de recherche et d'analyse pour Elasticsearch, un moteur de recherche open source. Banana utilise les capacités de configuration de tableau de bord de Kibana, et propose des fonctionnalités supplémentaires. Il s'agit d'une interface utilisateur flexible et riche, qui permet également de développer facilement des applications tirant parti de la puissance de Solr. Il permet de visualiser les données sous formes de graphiques, de sélectionner les métadonnées à afficher, de trier selon les critères de notre choix. Avant que notre corpus ne devienne assez grand pour que l'on puisse générer des visualisations parlantes, le tableau de bord Banana nous a été très utile pour explorer les premières données. Nous l'avons utilisé ensuite tout au long du projet, car grâce à la combinaison des requêtes Solr avec les visualisations sur le tableau de bord, Banana est un outil très efficace.

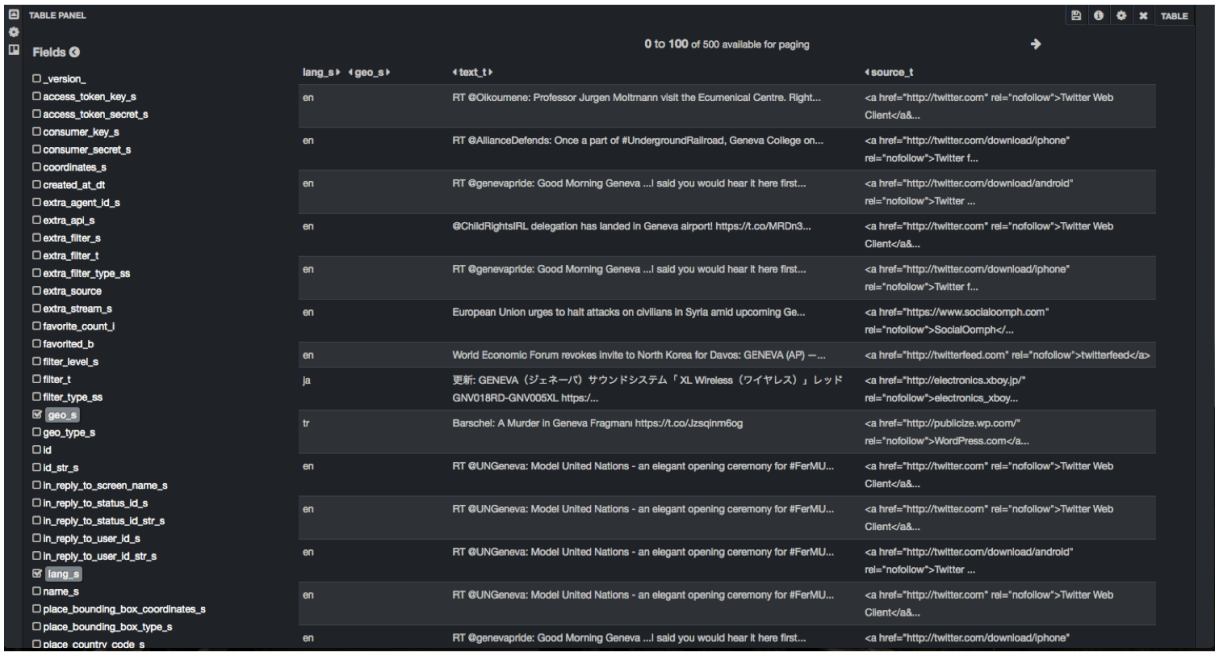
Voici un exemple de visualisation dans Banana des tweets émis sur les dernières 24h :

Figure 4 : Visualisation sous forme graphique dans Banana des tweets émis au cours des dernières 24h




(Béguelin 2015)

Figure 5 : Visualisation sous forme de liste dans Banana des tweets émis au cours des dernières 24h et de leurs métadonnées




(Béguelin 2015)

### 3.4.2 Tableur

Une fois les données exportées en .csv depuis Solr, nous avons pu les étudier dans le logiciel de tableur Microsoft Excel. Les fonctions de filtre, tri, recherche et les tableaux croisés nous ont permis de faire parler nos données. Nous avons entre autres pu définir la répartition linguistique en pourcentages, le nombre d'utilisateurs uniques, et le nombre de tweets par

utilisateur. L'analyse de cette métrique combinée à la position géographique depuis laquelle les tweets sont émis, nous a permis de détecter des robots. En effet, certains utilisateurs envoient un grand nombre de tweets, et ce depuis un point géographique dont les coordonnées GPS ne varient pas. Il s'agit par exemple de comptes Twitter publicitaires, ou de bornes météo qui envoient sous forme de tweets les mesures détectées par leurs capteurs.

Après avoir identifié les comptes d'utilisateurs robots, nous avons réajusté notre requête Solr en rajoutant ce paramètre : élimination des tweets provenant des comptes concernés. Le processus est donc itératif : après un premier export de données puis analyse de celles-ci, les éventuels problèmes sont détectés, menant à un réajustement de la requête pour un nouvel export, une nouvelle analyse des données, et ainsi de suite.

### 3.4.3 CartoDB

Le logiciel que nous avons principalement utilisé pour la visualisation de nos données est CartoDB[2]. Cette plateforme en SaaS propose des cartes provenant d'applications web telles que OpenStreetMap ou GoogleMaps, et permet la visualisation de données géolocalisées. CartoDB permet aux utilisateurs avec des compétences en programmation très diverses de créer des visualisations à partir de big data et de les diffuser sur le web. Logiciel open source, il peut également être personnalisé par les utilisateurs qui souhaiteraient développer leur propre version.

[2] CartoDB https://cartodb.com/

# 4. Frontières linguistiques – Résultats

Les résultats confirment l'hypothèse H1.1a. Nous retrouvons une importante variété linguistique avec 40 langues détectées par Twitter parmi les tweets géolocalisés. La polarisation entre le français (38.4%) et l'anglais (33.9%) est évidente, bien que les langues latines (espagnol, portugais et italien) s'élèvent également à 8%. Les 32 autres langues représentent le 10.7% de la totalité de tweets.

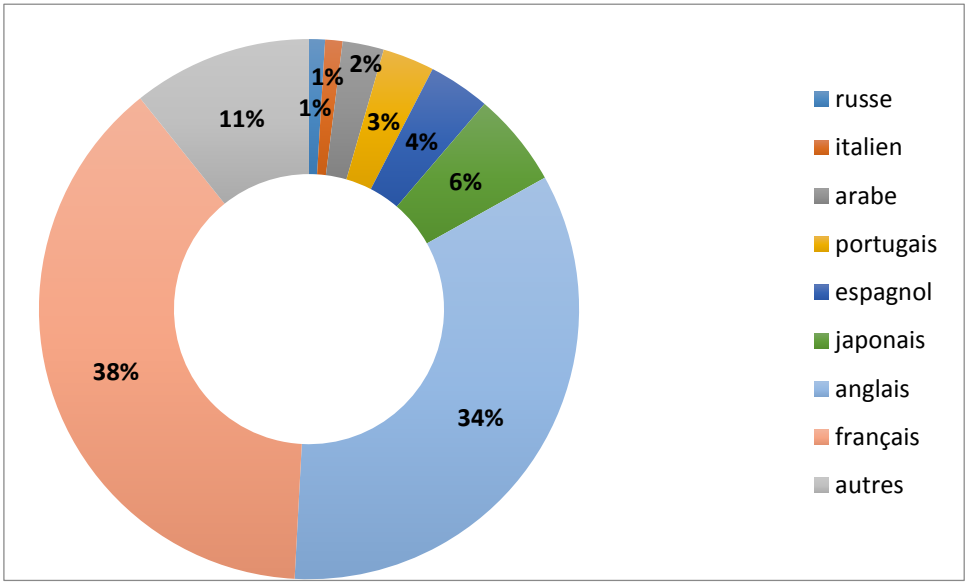
Tableau 1 : 8 premières langues détectées parmi les tweets géolocalisés à Genève

| | | |
|---|---|---|
| français | 14174 | 38.4 |
| anglais | 12525 | 33.9 |
| autres | 3965 | 10.7 |
| japonais | 2084 | 5.6 |
| espagnol | 1362 | 3.7 |
| portugais | 1145 | 3.1 |
| arabe | 913 | 2.5 |
| italien | 384 | 1.0 |
| russe | 352 | 1.0 |
| Total | 36904 | 100.0 |

(Banfi 2015)

L'étonnante présence du japonais est due à la suractivité d'une seule utilisatrice qui tweete le 4.3% de la totalité de tweets (voir chapitre 8. Utilisateurs uniques et profils).

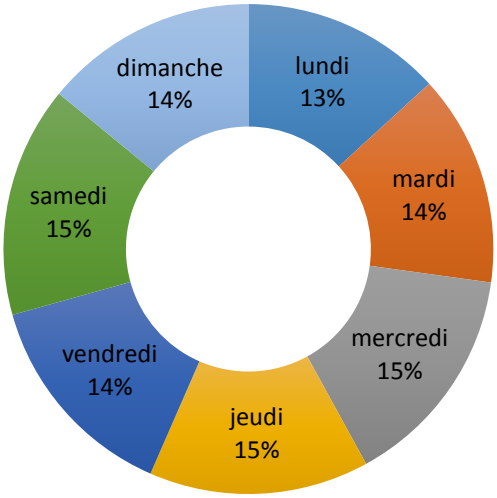
Figure 6 : Variété des langues parmi les tweets géolocalisés à Genève




(Banfi 2015)

L'hypothèse H1.1b est seulement partiellement confirmée par les résultats car il n'y a pas de variations significatives de la quantité de tweets entre les jours de la semaine.

Figure 7 : Distribution des tweets géolocalisés à Genève par jours de la semaine

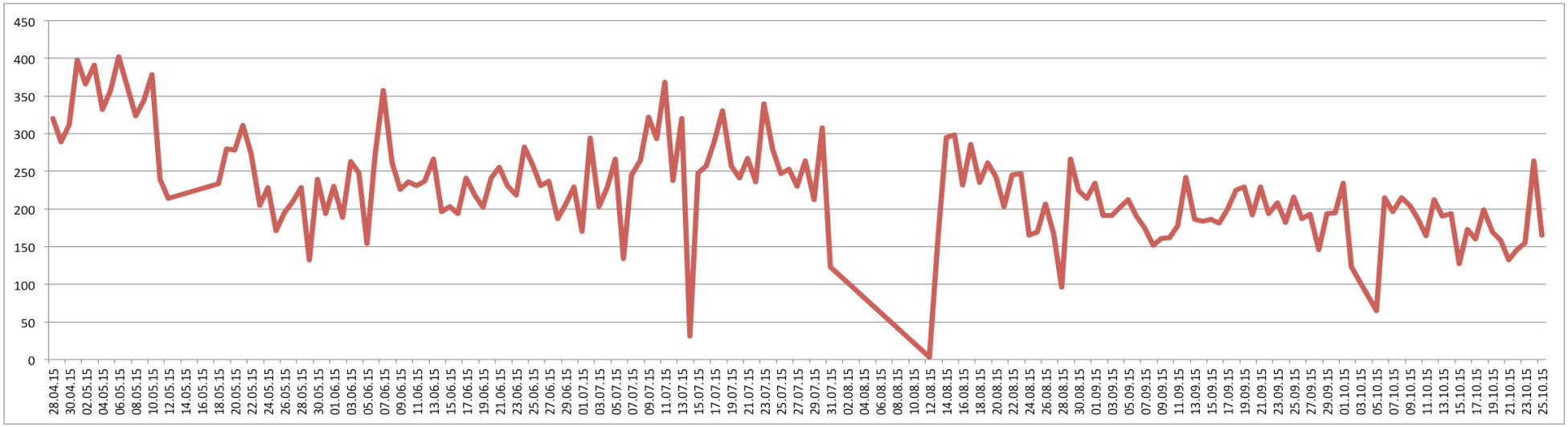



(Banfi 2015)

La distribution totale de tweets géolocalisés par semaine est homogène et s'élève à 15% par jour de la semaine avec une légère baisse le lundi (13%) et une variation à la hausse les mercredis, les jeudis et les samedis de 15%.

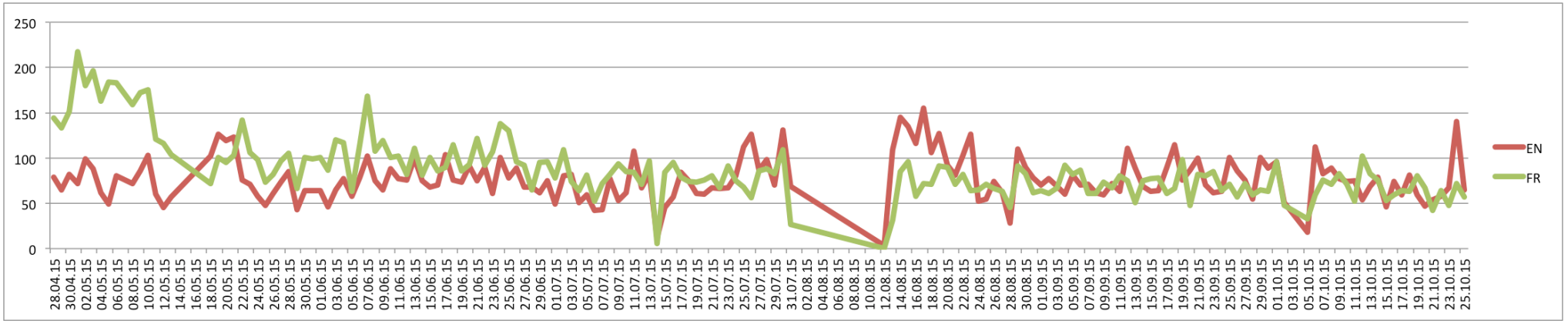
Par contre, certains vendredis, samedis et dimanches en été, le nombre de tweets augmente au-dessus de 300 tweets par jour. Nous observons également que certaines périodes accusent une baisse de tweets à cause de problèmes techniques expliqués dans le chapitre 9.Problèmes techniques.

Figure 8 : Nombre de tweets par jour




(Banfi 2015)

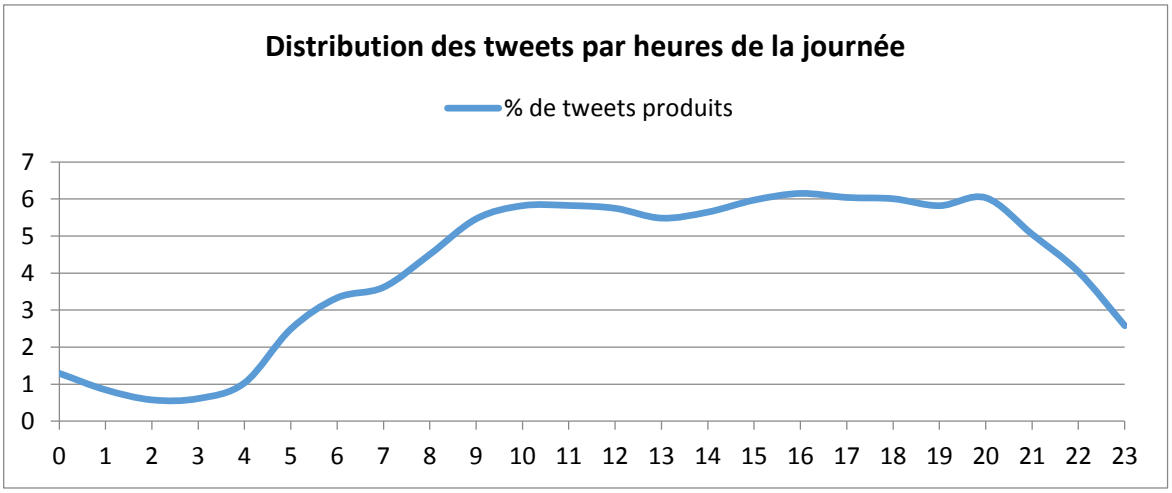
Figure 9 : Nombre de tweets en anglais et en français par jour




(Banfi 2015)

Nous observons également une variation significative de la production de tweets entre les périodes diurnes et nocturnes.

Figure 10 : Production de tweets géolocalisés par heures de la journée



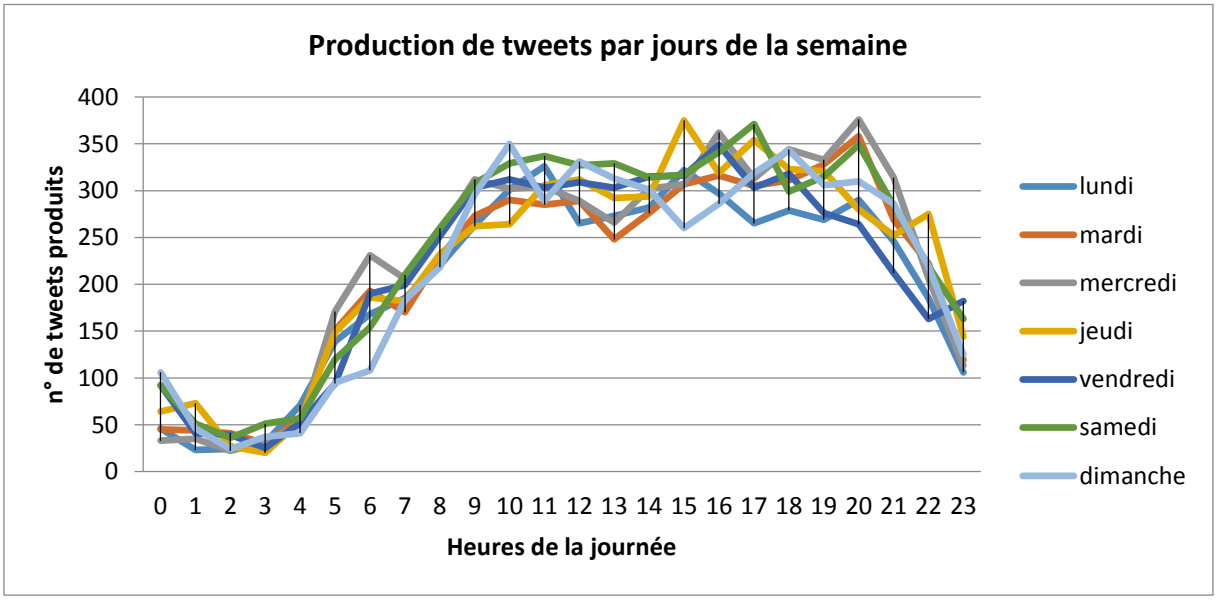

(Banfi 2015)

La production de tweets suit les phases de la vie ordinaire d'un citoyen genevois. Le pourcentage de tweets augmente de 0.5% à 1.5% entre minuit et 04h00 du matin pour progresser de 2% à 4% dans la phase du réveil, entre 5h00 et 6h00. Un pourcentage qui augmente encore entre 8h00 et 10h00 quand la journée de travail démarre pour se stabiliser autour de 6 % pendant toute la journée de travail jusqu'à 20h00.

Entre 20h00 et 23h00, une baisse de nombre de tweets s'enregistre. A 23h00, le pourcentage de tweets produits s'élève à 2,5%.

La production des tweets est plus importante pendant les heures de travail que pendant les moments de repos. La majorité de la population tweete pendant les heures diurnes au travail ou à l'école.

Figure 11 : Production de tweets par jours de la semaine



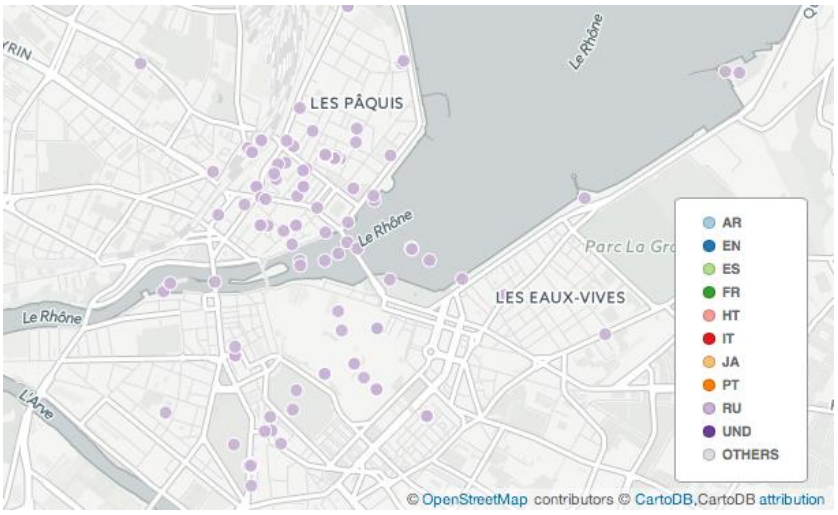

(Banfi 2015)

Toutefois, selon les jours de la semaine, la distribution par heures des tweets peut enregistrer des variations plus au moins accentuées.

Entre 2h00 et 4h00, la différence est exiguë selon les jours de la semaine. Par contre, nous trouvons une variation majeure notamment entre le mercredi et le mardi à 6h00, entre le dimanche et le jeudi à 15h00, entre le vendredi et le dimanche à 13h00, entre le mercredi et le vendredi à 20h00, entre le lundi et le samedi à 17h00 (pour détails voir tableau en annexe 1).

Pour ce qui concerne la troisième hypothèse H1.1c, elle est vérifiée par le biais des explorations visuelles dans CartoDB.

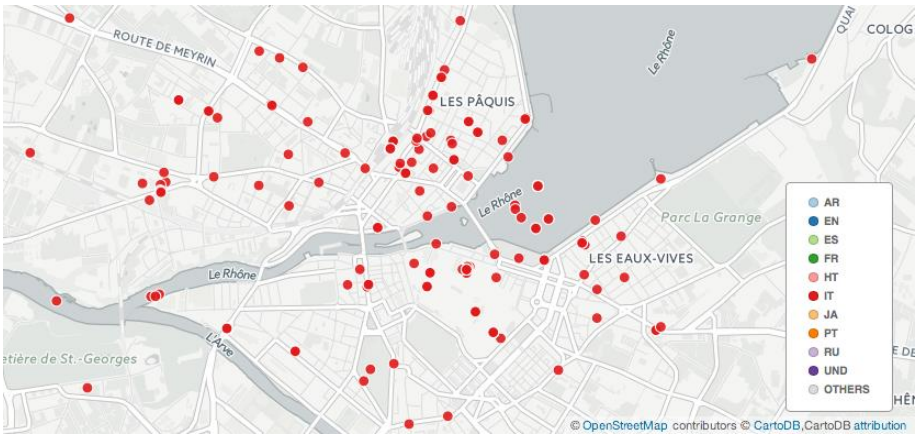
Figure 12 : Tweets en russe




(Banfi 2015, réalisé avec CartoDB)

Les tweets en russes se concentrent entre la gare et le Pont de la machine, en passant par l'île Rousseau et les Pâquis. Nous en retrouvons encore quelques-uns au rond-point de Plainpalais et au Jet d'eau.

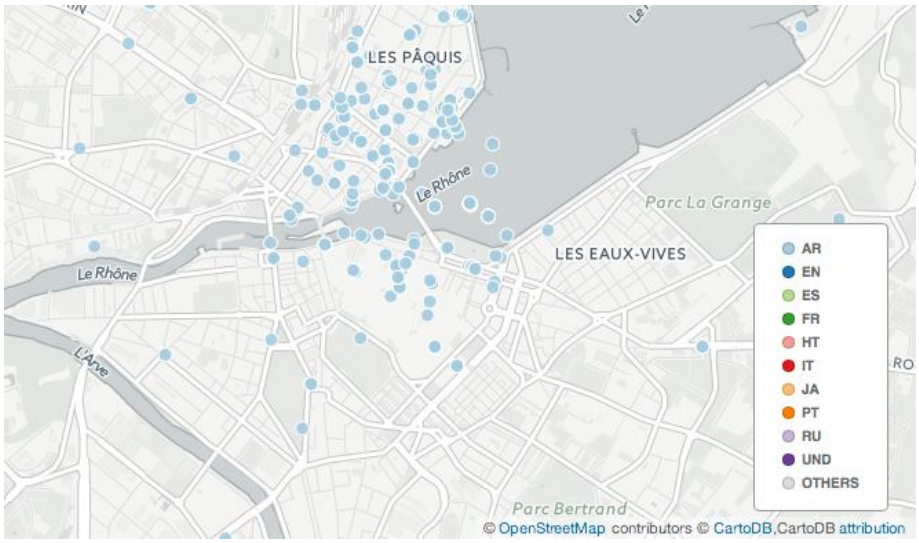
Figure 13 : Tweets en italien




(Banfi 2015, réalisé avec CartoDB)

Les tweets en italien se concentrent à la gare, le long de la rue de Lausanne et dans les restaurants-pizzeria (la pizzeria Al Molino au Molard, des pizzerias aux Charmilles) et au Jet d'eau.

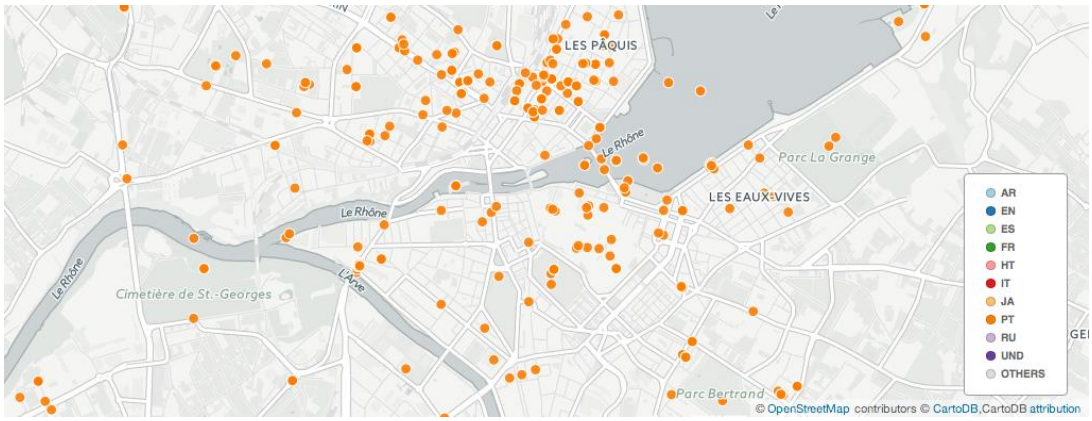
Figure 14 : Tweets en arabe




(Banfi 2015, réalisé avec CartoDB)

Les tweets en arabe proviennent en majorité du square du Mont Blanc, de la place de la Fusterie et de la place du Molard, de la rue du Mont Blanc, du Jet d'eau et du quai du Mont Blanc.

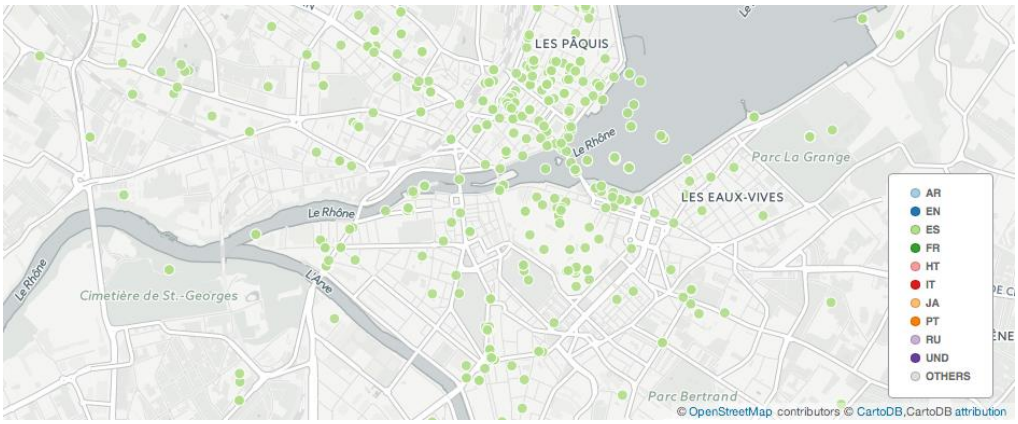
Figure 15 : Tweets en portugais




(Banfi 2015, réalisé avec CartoDB)

Les tweets en portugais se concentrent dans les parcs derrière la gare, à la Jonction, au parc de Bastions et au Jet d'eau, dans des cafés le long de la rue de la Servette et aux Charmilles.

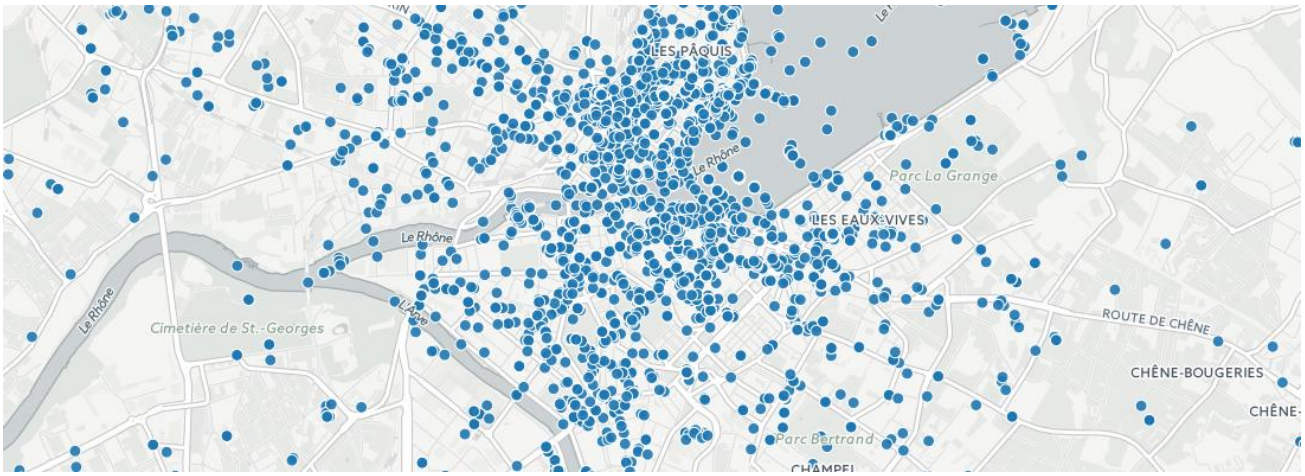
Figure 16 : Tweets en espagnol




(Banfi 2015, réalisé avec CartoDB)

Les tweets en espagnol se concentrent au Jardin Anglais, aux Acacias, à la Jonction, aux Pâquis et sur le pont du Mont Blanc.

Figure 17 : Tweets en anglais




(Banfi 2015, réalisé avec CartoDB)

Les tweets en anglais se concentrent à l'aéroport, au parc de l'Ariana, à la Place des Nations, le long du lac, aux Pâquis, à l'Organisation mondiale du commerce, à la Perle du lac, à l'Hôtel intercontinental Genève, à Palexpo, au CERN, dans la vieille ville (Place de la Fusterie et Place du Molard, le long de la rue du Mont Blanc et du quai du Mont Blanc).

Figure 18 : Tweets en français


(Banfi 2015, réalisé avec CartoDB)

Les tweets en français se concentrent dans les écoles, les supermarchés, aux Pâquis, le long du bord du lac, à Carouge et aux arrêts de bus, à la gare et à l'aéroport.

Selon les données de l'Office cantonal de la statistique de Genève, le français joue un rôle majoritaire au-delà de 60% dans toute la ville suivi par l'italien, l'espagnol, l'anglais, l'allemand. Nos données géolocalisées révèlent une divergence entre les données statistiques officielles du canton (qui ne prennent pas en compte les fonctionnaires internationaux et les ambassadeurs) et la pratiques linguistiques sur Twitter des résidents genevois.

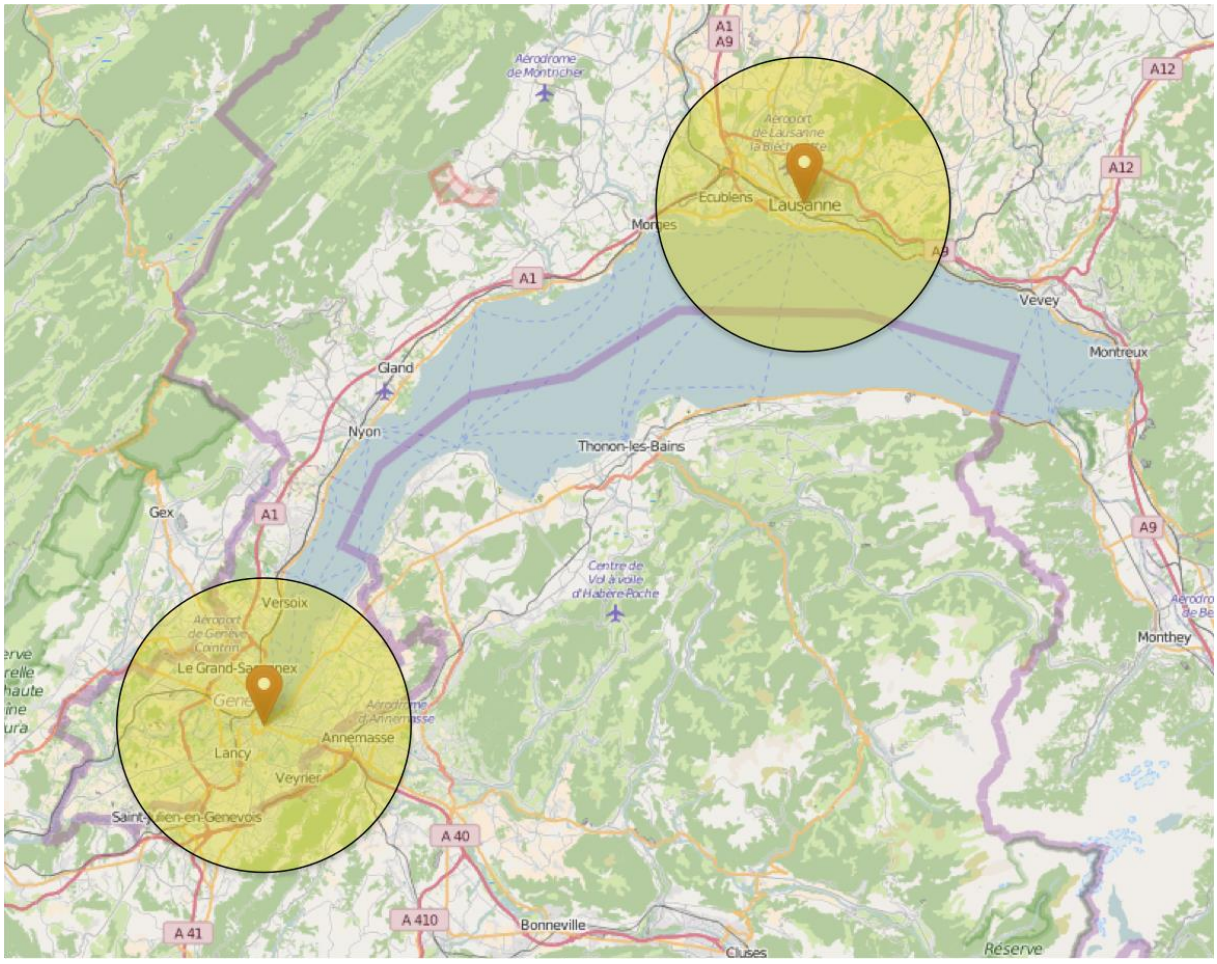
# 5. Comparaison entre Genève et Lausanne

Le projet GEoTweet se concentrant sur les tweets genevois, nos analyses ont donc porté sur un corpus géolocalisé à Genève. Toutefois, nous considérons qu'il est intéressant de comparer ces résultats avec une autre ville au profil comparable, Lausanne. En effet, Genève et Lausanne sont situées dans la même région politique et économique, et partagent des caractéristiques telles que l'internationalité et la proximité avec la frontière française (travailleurs frontaliers venant de l'autre rive du lac Léman).

Afin de comparer ces deux villes, nous avons travaillé sur deux corpus de tweets récoltés sur six mois (28 avril - 26 octobre 2015) dans un rayon de 10 km depuis le point central de chacune:

- 46°12'00.0" de latitude Nord et 6°09'00.0" de longitude Est pour Genève, ce qui correspond à un point situé entre la Place du Bourg-de-Four et la Promenade Saint-Antoine
- 46°31'15.6" de latitude Nord et 6°37'51.6" de longitude Est pour Lausanne, ce qui correspond à un point situé entre la Place Centrale et la Place de l'Europe

Figure 19 : Zones de capture des tweets pour la comparaison entre Genève et Lausanne (rayon de 10 km)




(Béguelin 2016, réalisé avec OpenStreetMap)

## 5.1 Nombre de tweets

Une fois les robots éliminés[3], nous avons obtenu 33'715 tweets pour le corpus de Genève, contre 9'826 pour celui de Lausanne.

Pour le même rayon et la même fourchette temporelle, nous avons donc environ trois fois plus de tweets à Genève qu'à Lausanne. Cette différence n'est pas justifiée par la densité d'habitants de Genève par rapport à Lausanne.[4]

## 5.2 Langues

La comparaison entre ces deux corpus révèle une plus grande diversité linguistique à Genève, où l'on dénombre 40 langues différentes contre 37 à Lausanne. Les langues détectées à Genève mais absentes du corpus lausannois sont les suivantes : l'hindi (2 tweets), le serbe (9 tweets), et l'ourdou[5] (1 tweet).

Sans surprise, les deux langues les plus importantes sont le français et l'anglais, à Genève comme à Lausanne. Les autres langues les plus importantes varient toutefois entre les deux villes :

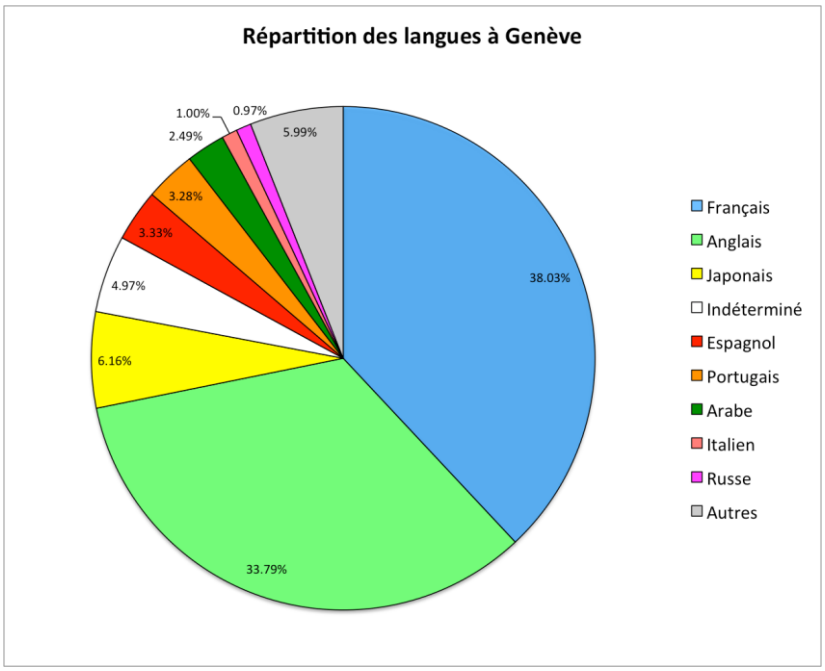
Tableau 2 : Liste des langues les plus fréquentes dans les tweets

| | **Genève** | **Lausanne** |
|---|---|---|
| **1** | Français | Français |
| **2** | Anglais | Anglais |
| **3** | Japonais | Portugais |
| **4** | Indéterminé | Indéterminé |
| **5** | Espagnol | Espagnol |
| **6** | Portugais | Allemand |
| **7** | Arabe | Italien |
| **8** | Italien | Arabe |
| **9** | Russe | Estonien |
| **10** | Autres | Autres |

(Béguelin 2016)

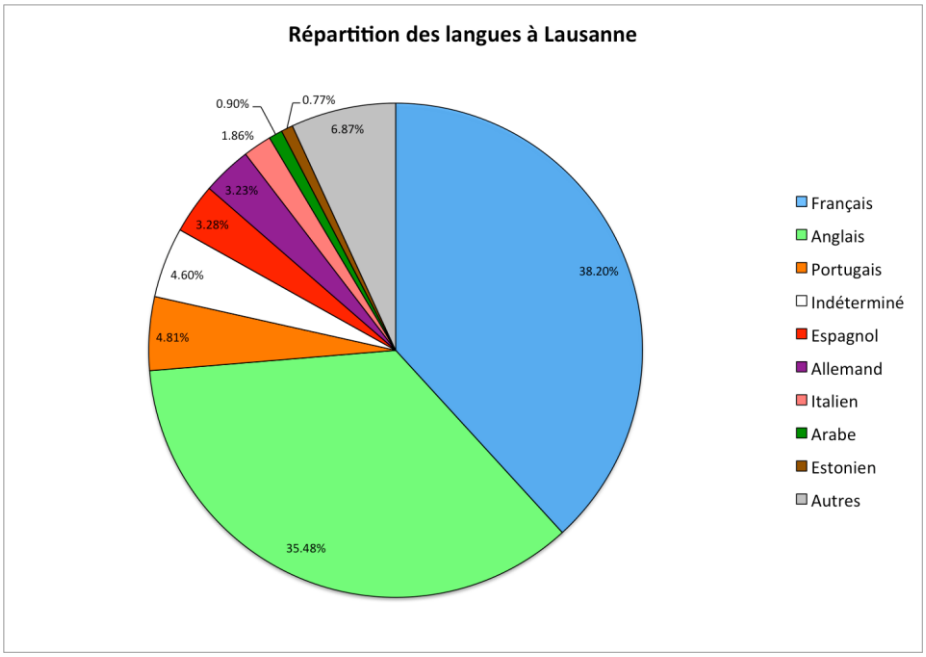
[3] Voir chapitre 3. Méthodologie, pour l'élimination des robots
[4] Selon les chiffres 2000 de l'Office fédéral de la statistique (voir annexe 2)
[5] Langue écrite en caractères arabes, utilisée au Pakistan

Figure 20 : Répartition des langues à Genève




(Béguelin 2016)

Figure 21 : Répartition des langues à Lausanne




(Béguelin 2016)

On constate que le japonais[6] et le russe sont présents dans les dix premières langues pour Genève, mais apparaissent plus tard pour Lausanne.

L'allemand et l'estonien en revanche, apparaissent dans les premiers résultats du classement lausannois, mais plus bas dans celui de Genève.

[6] Voir chapitre 8. Utilisateurs uniques et profils, pour comprendre pourquoi le japonais arrive en troisième position à Genève

La langue "Indéterminé" correspond aux tweets comprenant uniquement des émoticônes ou des signes n'appartenant pas à une langue précise.

En comparant nos résultats avec les données statistiques de l'Office fédéral de la statistique (recensement 2000) de la langue principale de la population résidante dans la ville de Genève et la ville de Lausanne, nous constatons que l'anglais est surreprésenté dans les langues détectées par Twitter. En revanche, l'allemand est sous-représenté dans les tweets genevois, mais parfaitement représentatif dans les tweets lausannois (voir annexe 2).

## 5.3 Source des tweets

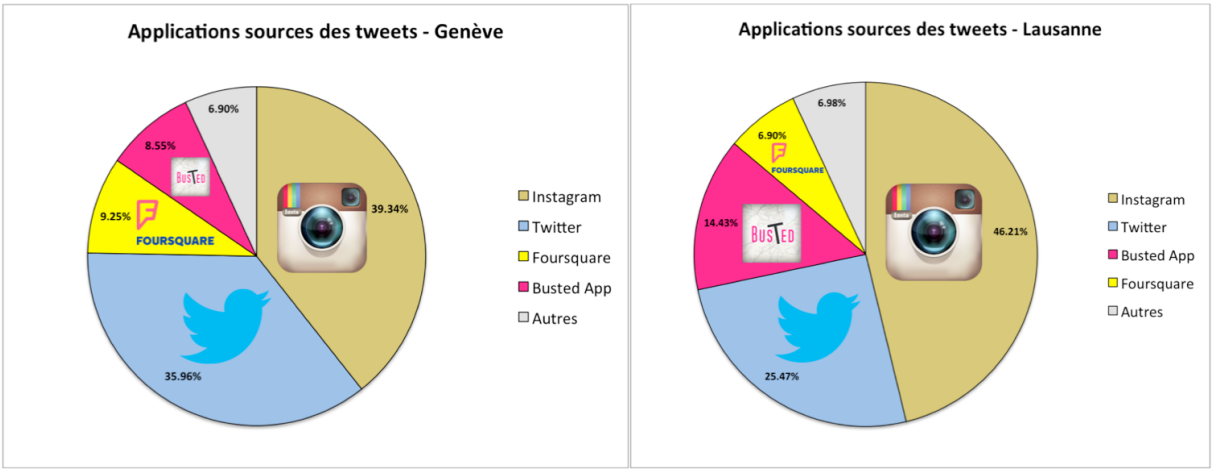
Les applications sources des tweets sont également plus variées à Genève qu'à Lausanne : le corpus de tweets genevois compte 11 sources de plus que le corpus lausannois.

La liste complète est disponible en annexe 3.

Tout comme pour Genève dans notre analyse des corpus obtenus aussi bien sur un rayon de 20 km[7] que sur un rayon de 10 km, Instagram est la première source de tweets à Lausanne.

Si dans le cas de Genève la différence entre les tweets natifs de Twitter et d'Instagram est faible (35.96% contre 39.34%), elle est en revanche très marquée à Lausanne (25.47% pour Twitter, 46.21% pour Instagram). Cela signifie donc que les trois quarts des tweets de notre corpus lausannois proviennent d'applications tierces.

Figure 22 : Applications sources des tweets - Genève et Lausanne




(Béguelin 2016)

Les autres principales applications sources sont Foursquare et BustedApp. La différence n'est pas très marquée à Genève (9.25% pour Foursquare, 8.55% pour BustedApp), mais est beaucoup plus forte à Lausanne, où BustedApp (14.43%) prend la troisième position devant Foursquare (6.90%).

Les twittos[8] lausannois utilisent donc plus ce réseau social comme plateforme pour faire transiter leurs contenus d'autres applications, alors qu'à Genève ils sont un plus grand nombre à utiliser directement Twitter.

[7] Voir chapitre 7. Source des tweets
[8] Utilisateur de Twitter

## 5.4 Utilisateurs uniques

Le taux d'utilisateurs uniques à Genève et Lausanne correspond au taux de tweets émis dans ces deux villes.

En effet, le corpus de Genève compte 5988 utilisateurs uniques et celui de Lausanne en compte 2085, soit environ trois fois moins.

Environ la moitié des utilisateurs n'a envoyé qu'un seul tweet sur les six mois de notre capture, et environ un tiers entre 2 et 5 tweets, tant à Lausanne qu'à Genève.

Les tableaux suivants présentent le détail de ces résultats :

Tableau 3 : Tweets envoyés par utilisateur Genève

| **Genève** | |
|---|---|
| 1 seul tweet envoyé | 51.42% des utilisateurs |
| 2 tweets envoyés | 17.54% des utilisateurs |
| 3 tweets envoyés | 9.19% des utilisateurs |
| 4 tweets envoyés | 4.98% des utilisateurs |
| 5 tweets envoyés | 3.26% des utilisateurs |
| Entre 6 et 10 tweets envoyés | 6.75% des utilisateurs |
| Entre 11 et 100 tweets envoyés | 6.53% des utilisateurs |
| Entre 101 et 500 tweets envoyés | 0.33% des utilisateurs |
| Entre 500 et 1000 tweets envoyés | 0.07% des utilisateurs |
| Plus de 1000 tweets envoyés | 0.03% des utilisateurs |

(Béguelin 2016)

Tableau 4 : Tweets envoyés par utilisateur Lausanne

| **Lausanne** | |
|---|---|
| 1 seul tweet envoyé | 57.12% des utilisateurs |
| 2 tweets envoyés | 15.92% des utilisateurs |
| 3 tweets envoyés | 7.58% des utilisateurs |
| 4 tweets envoyés | 4.56% des utilisateurs |

| | |
|---|---|
| 5 tweets envoyés | 2.88% des utilisateurs |
| Entre 6 et 10 tweets envoyés | 5.90% des utilisateurs |
| Entre 11 et 100 tweets envoyés | 5.71% des utilisateurs |
| Entre 101 et 500 tweets envoyés | 0.29% des utilisateurs |
| Entre 500 et 1000 tweets envoyés | 0% des utilisateurs |
| Plus de 1000 tweets envoyés | 0.05% des utilisateurs |

(Béguelin 2016)

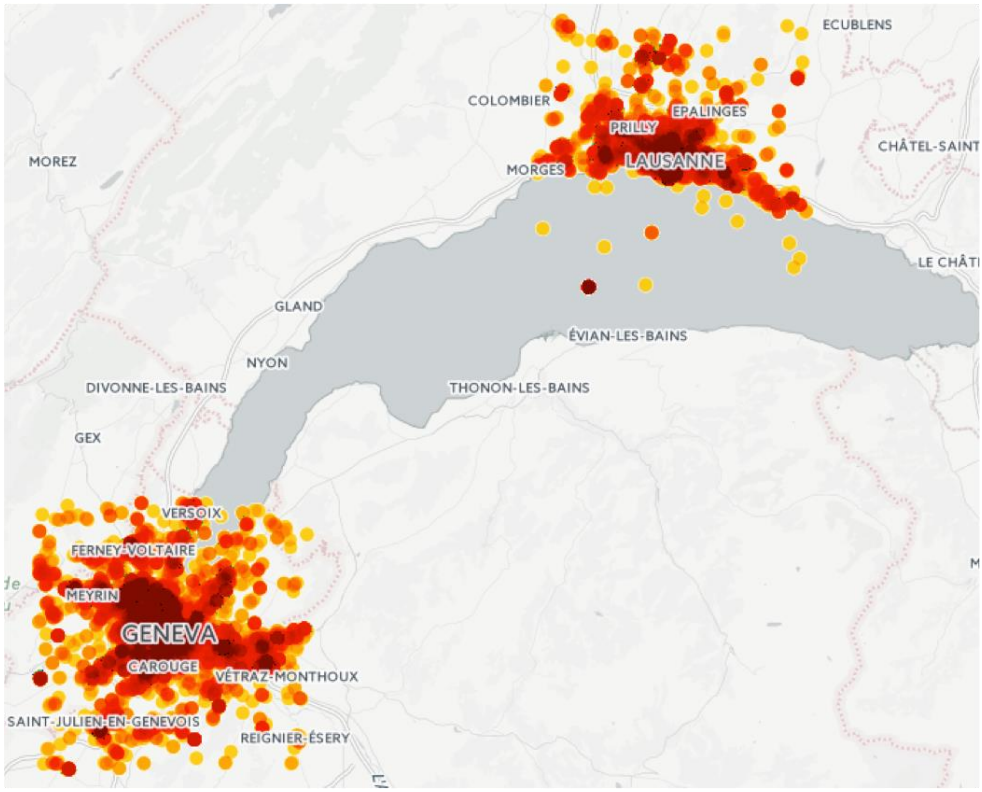
Genève compte plus de "gros" utilisateurs que Lausanne. En effet, six comptes Twitter genevois ont envoyé plus de 500 tweets en six mois, contre un seul à Lausanne.

Le profil de ces utilisateurs qui tweetent le plus est détaillé au chapitre 8 pour Genève. Concernant Lausanne, le seul compte Twitter ayant envoyé plus de 500 tweets est celui de Busted Lausanne. Il ne s'agit donc pas d'un utilisateur unique, mais bien de plusieurs utilisateurs regroupés via l'application Busted.

Le corpus de Lausanne ne comprend donc pas de twittos intensifs, contrairement à celui de Genève (voir détails au chapitre 8. Utilisateurs uniques et profils).

## 5.5 Visualisation comparée

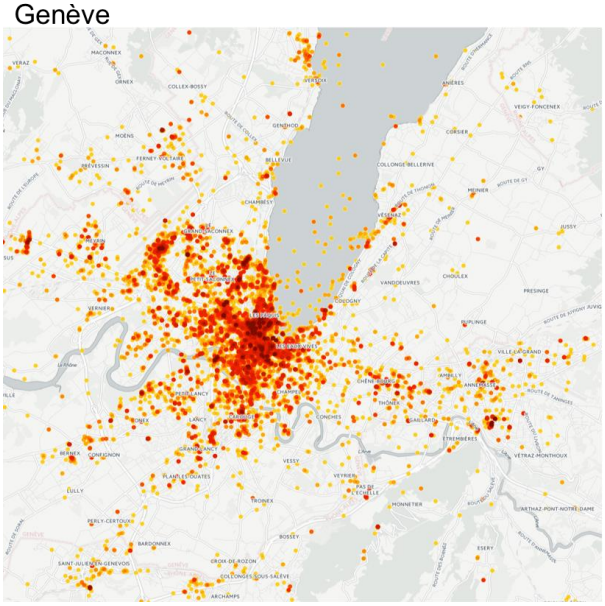
Figure 23 : Comparaison des tweets entre Genève et Lausanne - vue générale



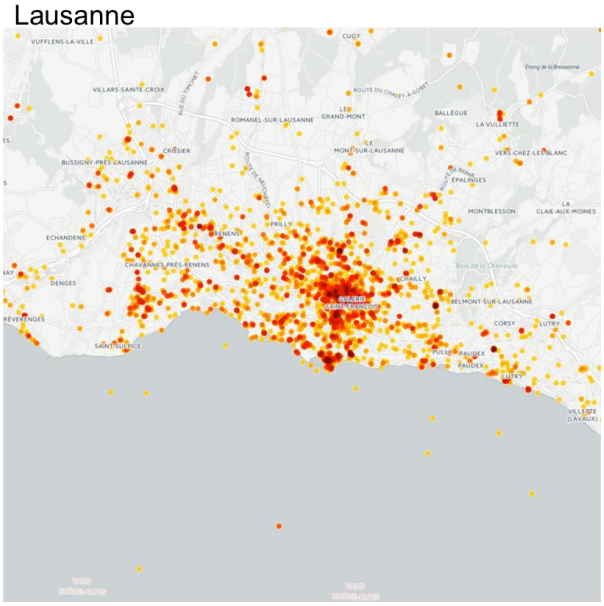

(Béguelin 2016)

Figure 24 : Comparaison des tweets entre Genève et Lausanne - détails

Genève



Lausanne



(Béguelin 2016)

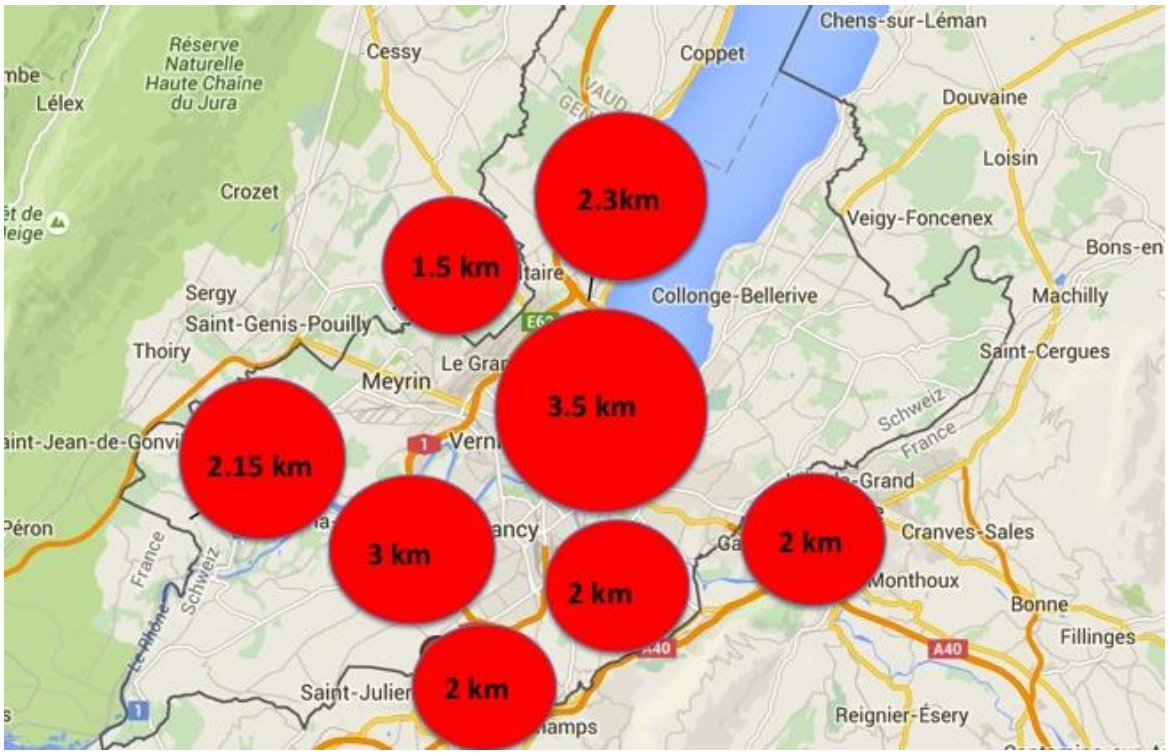
Ces visualisations effectuées à l'aide de CartoDB, représentent les deux corpus selon leur intensité. Plus la couleur est foncée, plus il y de tweets.

La visualisation reflète bien le fait que le corpus genevois comprend trois fois plus de tweets que Lausanne, mais nous pouvons également constater que les tweets genevois sont plus concentrés sur le centre-ville, alors que les tweets lausannois sont plus répartis.

# 6. Frontières virtuelles et géopolitiques – Résultats

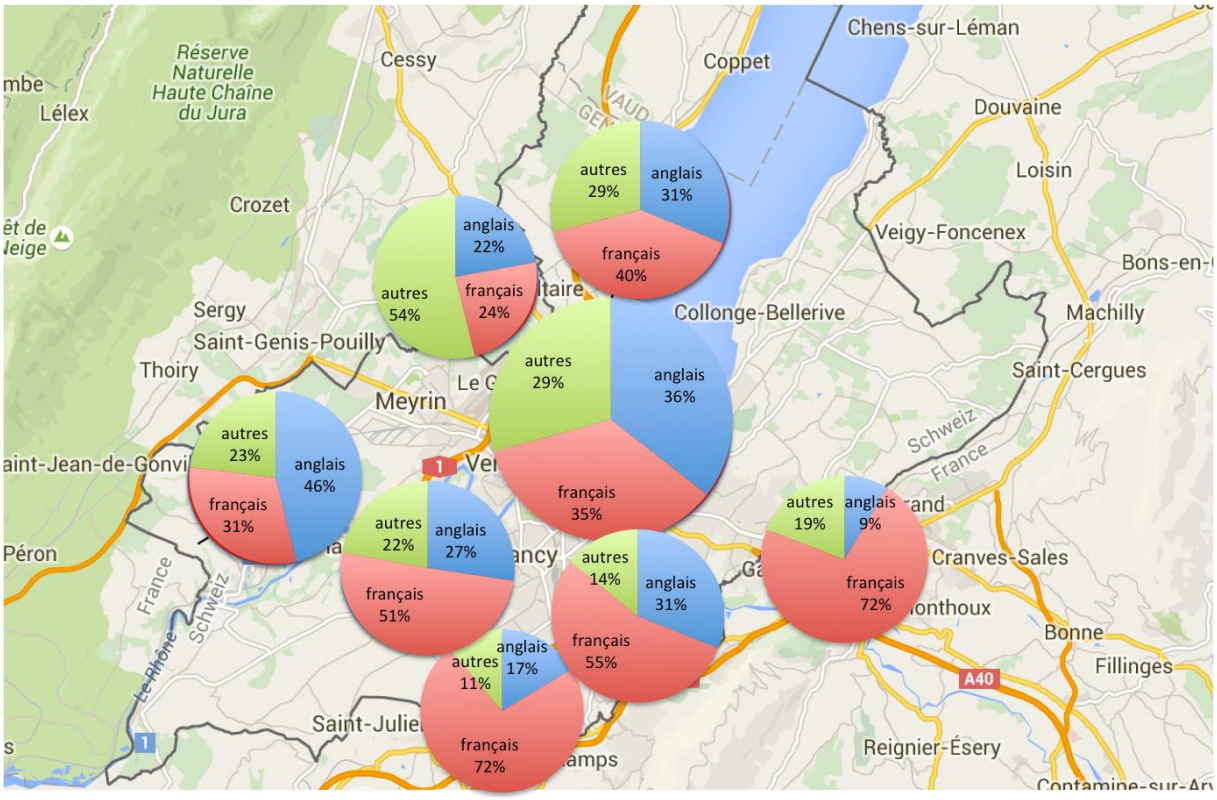
Pour ce qui concerne l'hypothèse H1.2a, les langues des tweets varient selon la proximité entre le centre de la ville et la frontière française. Pour l'analyse de tweets, 8 zones étaient analysées : 3 au-delà de la frontière et 5 à l'intérieur. Pour le centre de Genève une zone avec rayon de 3.5 km était sélectionnée à partir des coordonnées du Jet d'eau 46.207964, 6.156013, pour Annemasse une région d'un rayon de 2 km à partir des cordonnées 46.185870, 6.229634, pour St-Julien-en-Genevois de 2 km à partir du point 46.144973, 6.086811, pour Versoix de 2.3 km à partir du point 46.277169, 6.151289, pour Ferney-Voltaire de 1.5 km à partir du point 46.261268, 6.094984, pour Veyrier de 2 km à partir du point 46.175513, 6.155752, pour Bernex de 3 km à partir du point 46.179037, 6.073664, et pour Satigny de 2.15 km à partir du point 46.203576, 6.018217.

Figure 25 : Zones d'analyse




(Banfi 2015)

Figure 26 : Variation des langues des tweets géolocalisés entre les zones centrales du canton et sa périphérie



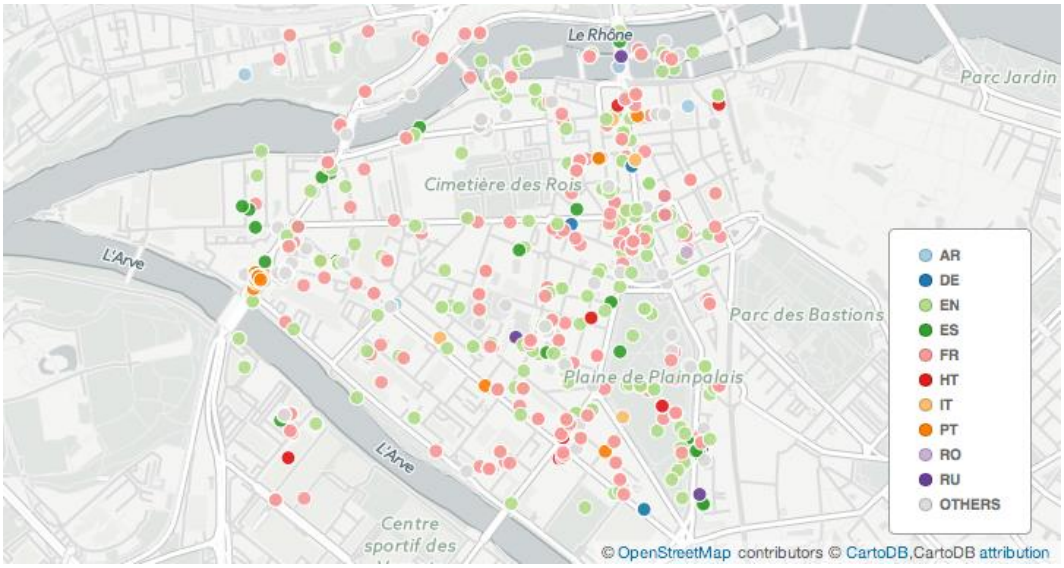

(Banfi 2015)

Au-delà d'Annemasse et de Saint-Julien, le français prévaut sur les autres langues (72%). Autour de Veyrier le français reste majoritaire (55%) bien que l'anglais soit la langue de 31% de tweets. Autour de Bernex, la situation est similaire à la zone de Veyrier avec une légère augmentation de la variété linguistique (autres langues 22%). Autour de Satigny, l'anglais prévaut sur le français et les autres langues. La variété linguistique au centre-ville et à Versoix augmente au détriment du français pour toucher les pourcentages les plus élevés à Ferney-Voltaire et autour de l'aéroport international de Genève. L'analyse des tweets confirme le profil international et multiculturel de la ville de Genève. Les données révèlent une influence sur les villes françaises au-delà de la frontière pour ce qui concerne la présence de l'anglais et des autres langues.

L'exploration de la visualisation avec CartoDB confirme la validité de l'hypothèse H1.2b, c'est-à-dire que selon les quartiers, la distribution des tweets et leur diversité linguistique varient.

Pour l'analyse des tweets par quartiers, 6 zones étaient analysées. Pour la zone entre la Jonction et Plainpalais, un rayon de 0.545 km a été défini à partir des coordonnées 46.200302, 6.135676. Pour la zone entre les bois de la Bâtie et le Cimetière St-Georges, le rayon est de 0.525 km à partir du point 46.198698, 6.118381. Pour celle entre la vieille ville et le Jardin Anglais, le rayon est de 0.470 km à partir du point 46.201787, 6.148894. Pour les Eaux-Vives, il est de 0.5 km à partir du point de 46.204044, 6.163614. Pour les Pâquis, le rayon est de 0.4 km à partir du point 46.211944, 6.148937. Pour la zone entre Servette et Saint-Jean, il est de 0.720 km à partir du point 46.209093, 6.121043. Enfin, pour la zone entre le Grand-Saconnex et les Nations Unies, un rayon de 1 km à partir du point 46.233173, 6.136664 a été défini.

Figure 27 : Variété linguistique des tweets géolocalisés entre Plainpalais et la Jonction




(Banfi 2015, réalisé avec CartoDB)

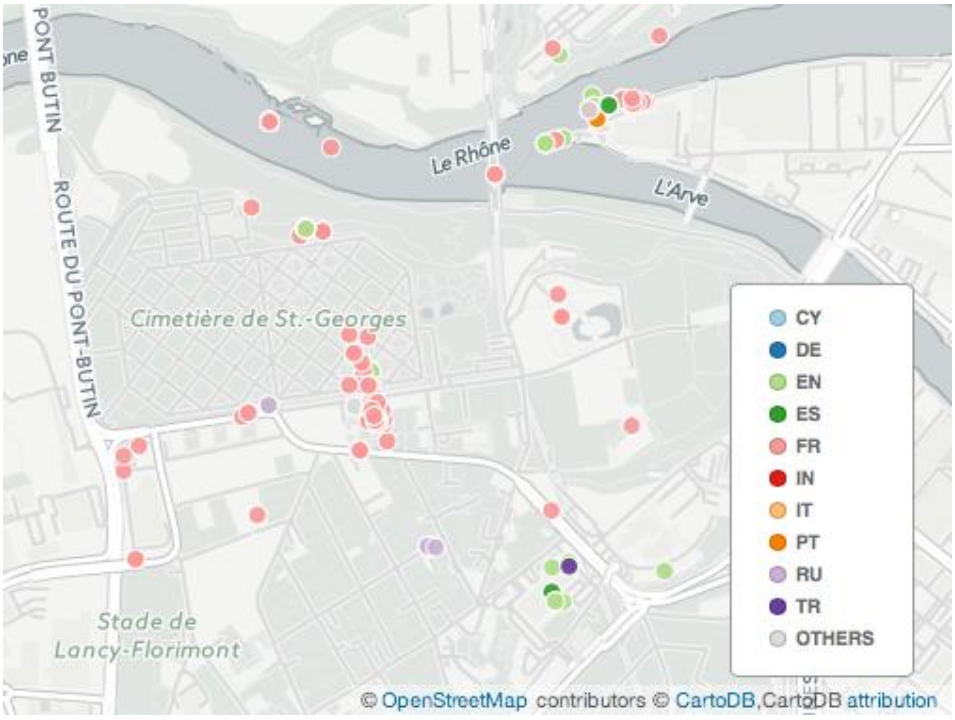
Tableau 5 : Variété linguistique des tweets géolocalisés entre Plainpalais et la Jonction

| | | |
|---|---|---|
| en | 1040 | 42.22 % |
| fr | 757 | 30.73 % |
| pt | 301 | 12.22 % |
| autres | 365 | 14.82 % |
| Total | 2463 | 100.00 % |

(Banfi 2015)

Dans la zone de Plainpalais, la Jonction et le Rhône nous avons une prévalence de l'anglais suivi par le français et une présence importante du portugais.

Figure 28 : Variété linguistique des tweets géolocalisés entre le Bois de la Bâtie et le cimetière St-Georges




(Banfi 2015)

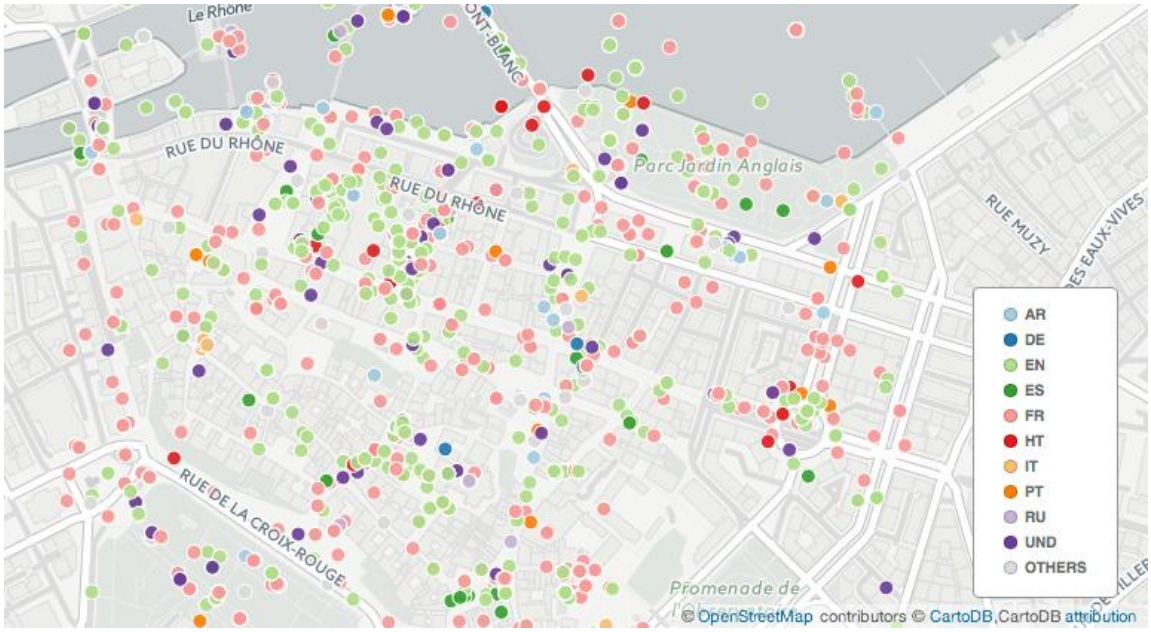
Tableau 6 : Variété linguistique des tweets géolocalisés entre le Bois de la Bâtie et le cimetière St-Georges

| | | |
|---|---|---|
| fr | 168 | 66.40 % |
| en | 29 | 11.46 % |
| autres | 56 | 22.13 % |
| Total | 253 | 100.00 % |

(Banfi 2015)

La zone du cimetière de St-Georges est presque exclusivement francophone.

Figure 29 : Variété linguistique des tweets géolocalisés entre la vieille ville et le Jardin Anglais




(Banfi 2015)

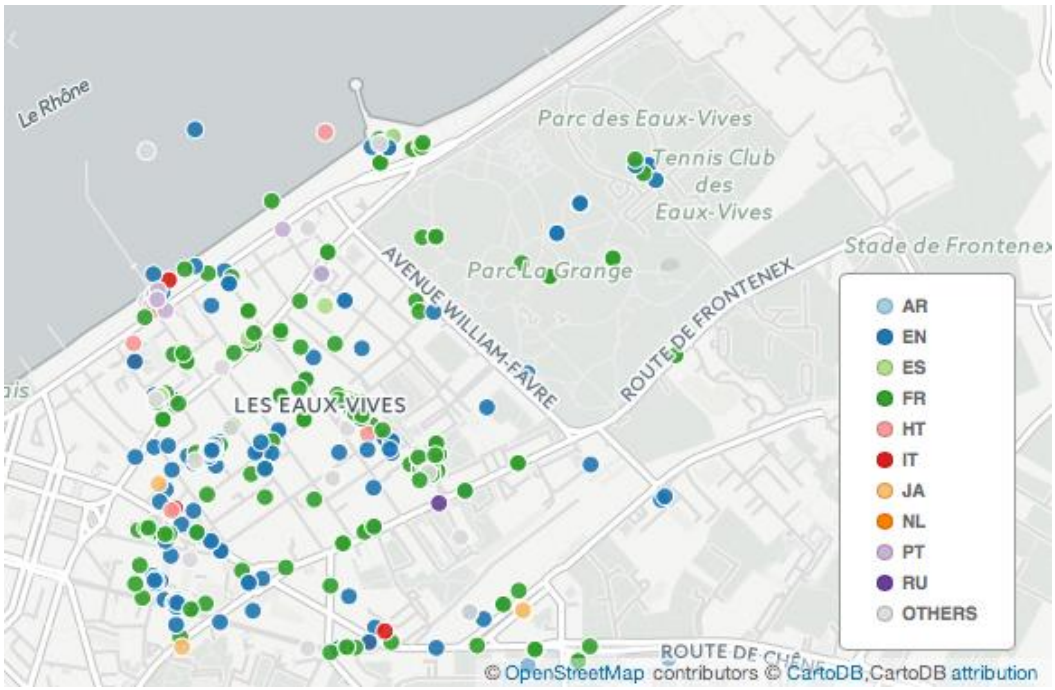
Tableau 7 : Variété linguistique des tweets géolocalisés entre la vieille ville et le Jardin Anglais

| | | |
|---|---|---|
| en | 2118 | 44.54 % |
| fr | 1506 | 31.67 % |
| ar | 169 | 3.55 % |
| pt | 81 | 1.70 % |
| it | 69 | 1.45 % |
| autres | 812 | 17.08 % |
| Total | 4755 | 100.00 % |

(Banfi 2015)

Entre la vieille ville et le Jardin Anglais, nous retrouvons une forte présence de l'anglais suivi du français. La variété linguistique reste importante avec le 3.5% des tweets en arabe, le 1,7% en italien et le 1,45% en portugais.

Figure 30 : Variété linguistique des tweets géolocalisés aux Eaux-Vives




(Banfi 2015)

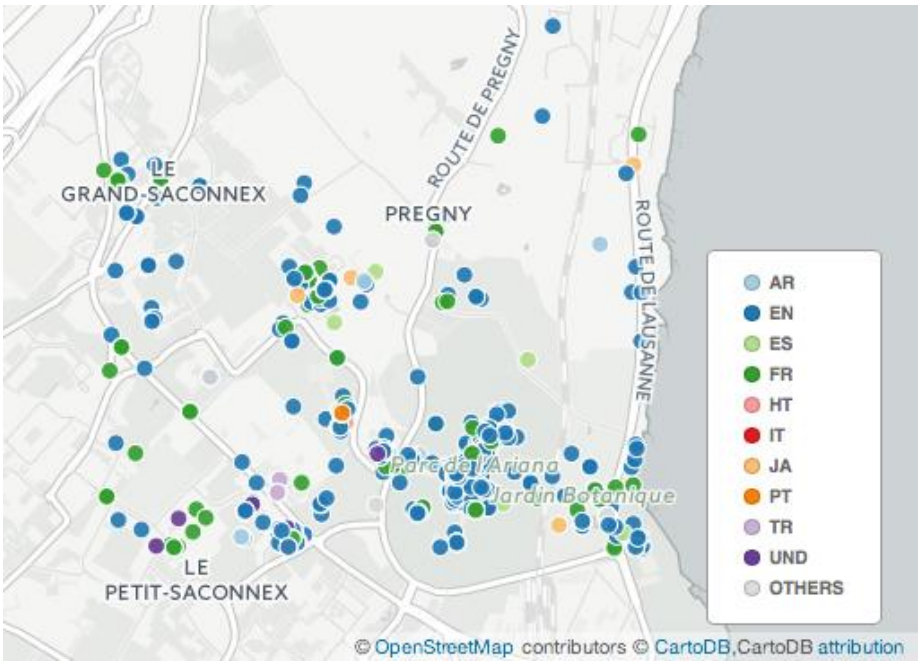
Tableau 8 : Variété linguistique des tweets géolocalisés aux Eaux-Vives

| | | |
|---|---|---|
| fr | 625 | 58.19 % |
| en | 311 | 28.96 % |
| pt | 18 | 1.68 % |
| it | 10 | 0.93 % |
| autres | 110 | 10.24 % |
| Total | 1074 | 100.00 % |

(Banfi 2015)

Aux Eaux-Vives le français est plus présent que l'anglais. La variété linguistique reste importante avec 1.7% des tweets en portugais et 1% en italien.

Figure 31 : Variété linguistique des tweets géolocalisés au Grand-Saconnex




(Banfi 2015)

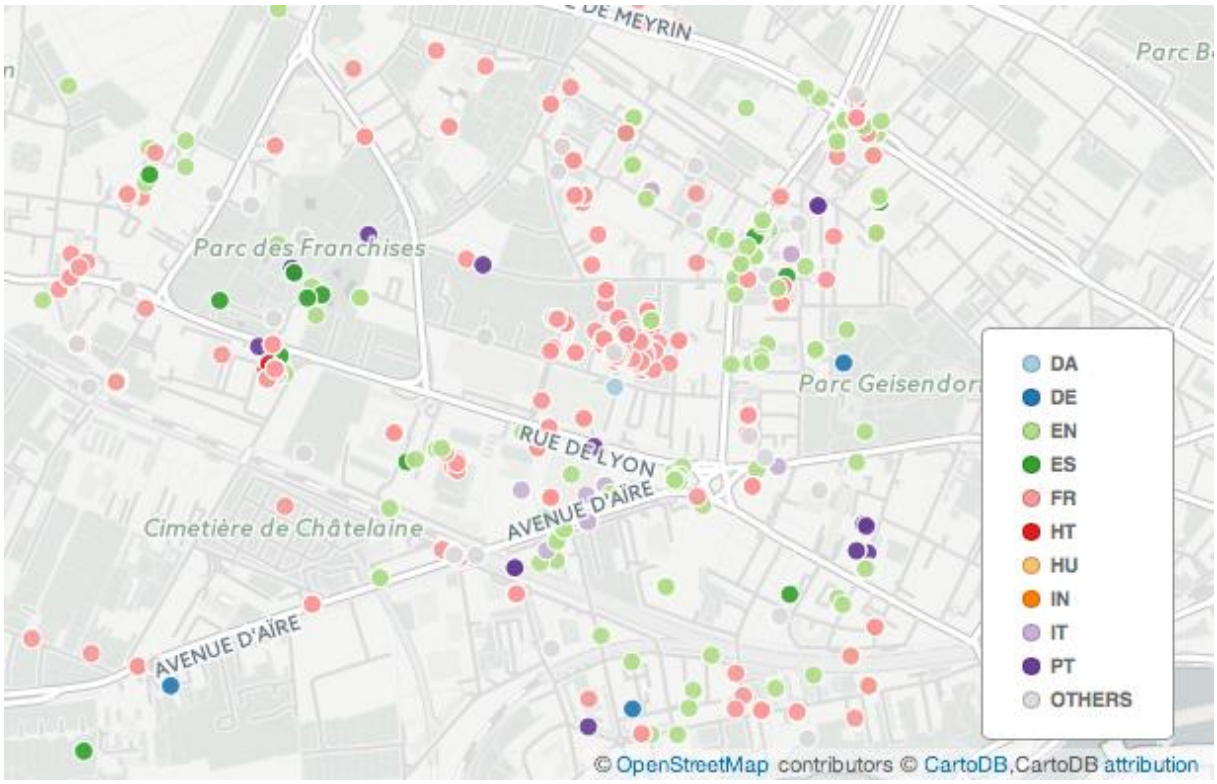
Tableau 9 : Variété linguistique des tweets géolocalisés au Grand-Saconnex

| | | |
|---|---|---|
| en | 869 | 67.47 % |
| fr | 167 | 12.97 % |
| es | 70 | 5.43 % |
| ar | 62 | 4.81 % |
| autres | 120 | 9.32 % |
| Total | 1288 | 100.00 % |

(Banfi 2015)

Au Grand-Saconnex, l'anglais est plus présent que le français. La variété linguistique reste importante avec 5.43% des tweets en espagnol et 4.81% en arabe.

Figure 32 Variété linguistique des tweets géolocalisés autour des Charmilles




(Banfi 2015)

Tableau 10 : Variété linguistique des tweets géolocalisés autour des Charmilles

| | | |
|---|---|---|
| fr | 420 | 55.93 % |
| en | 154 | 20.51 % |
| it | 34 | 4.53 % |
| es | 33 | 4.39 % |
| pt | 25 | 3.33 % |
| autres | 85 | 11.32 % |
| Total | 751 | 100.00 % |

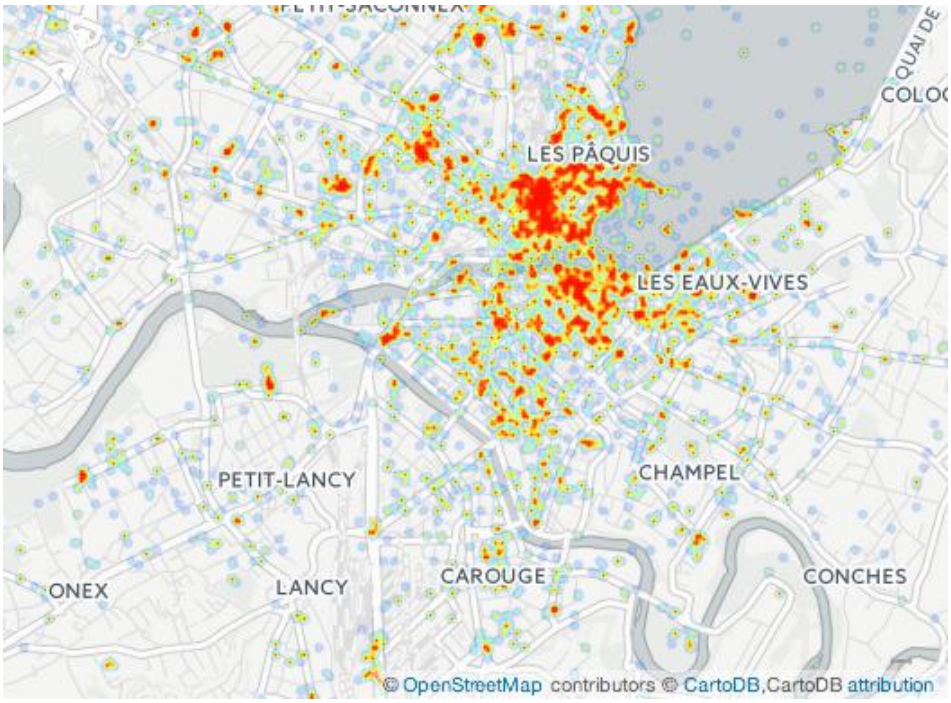
(Banfi 2015)

Autour du quartier des Charmilles, le français prévaut sur l'anglais. La variété linguistique reste importante avec le 4.53% des tweets en italien, 4.39% en espagnol et 3.33% en portugais.

## 6.1 Comparaison entre densité résidentielle et densité de tweets

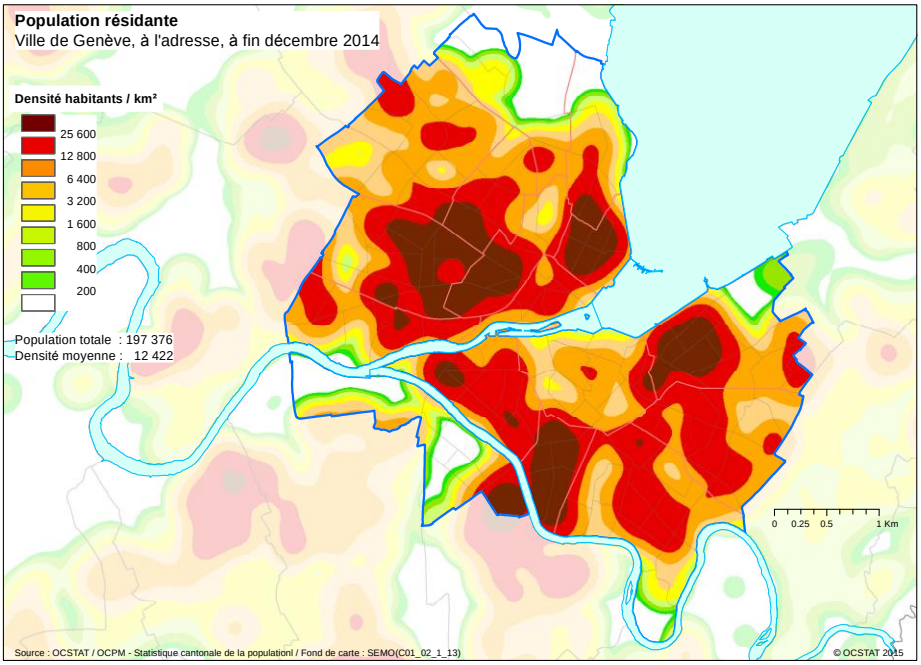
Nous observons de nombreuses différences entre la densité d'habitation et la densité des tweets. Les données virtuelles se concentrent surtout entre la vieille ville et la gare avec des pôles de concentration entre la Jonction et Plainpalais, mais présentent une faible densité à Carouge, au Petit-Lancy, à Onex et à Champel alors que ces zones révèlent une forte concentration d'habitants selon les statistiques officielles.

Figure 33 : Densité des tweets à Genève




(Banfi 2015, réalisé avec CartoDB)

Figure 34 : Densité résidentielle à Genève



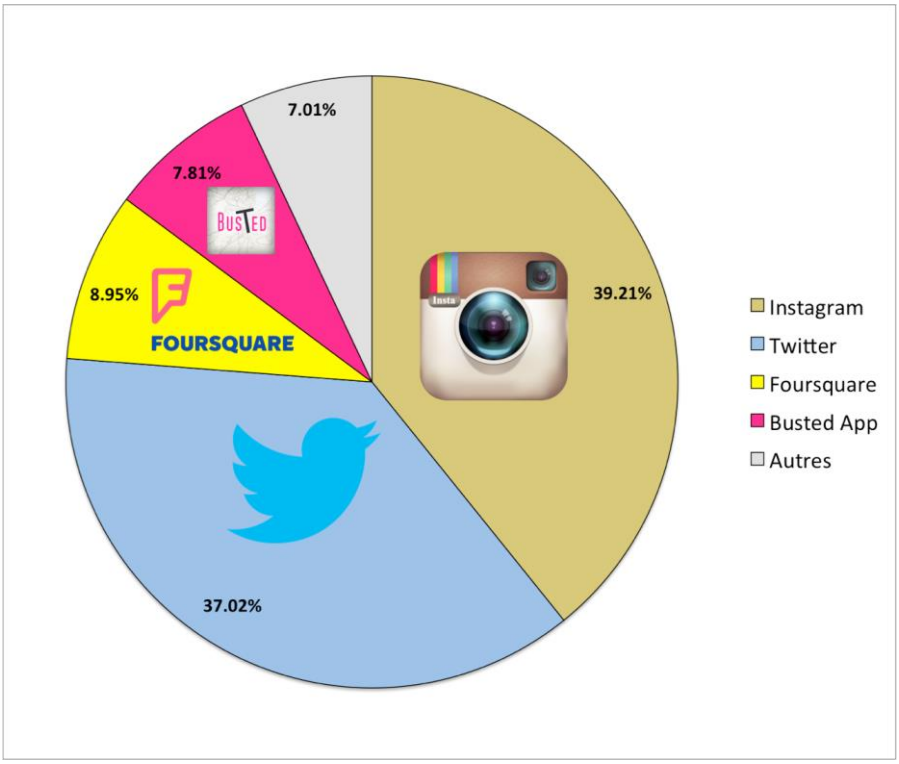

(Office cantonal de la statistique Genève, 2015)

# 7. Source des tweets

Une part importante de tweets est émise à partir d'autres applications que Twitter. Il est en effet possible de lier par exemple un compte Instagram ou Foursquare avec un compte Twitter, relayant ainsi sous forme de tweets tous les contenus publiés sur ces applications. De plus, certains outils comme Hootsuite permettent de gérer les flux de plusieurs réseaux sociaux en même temps via un tableau de bord. En plus des tweets rédigés dans Twitter par les utilisateurs, un grand nombre de contenus transitent donc par cette application.

A l'échelle de notre corpus de 36'904 tweets genevois (uniquement géolocalisés, et provenant d'utilisateurs humains), 39.2 % sont des publications nativement d'Instagram, plaçant ainsi ce réseau en tête des sources de tweets de notre corpus. 37.02% proviennent de l'application Twitter (toutes plateformes confondues), 8.95% sont des publications nativement de Foursquare, et 7.81% de BustedApp. Le reste (7.01%) provient de diverses applications, dont des outils tels que Hootsuite. La liste complète des applications sources est disponible en annexe 4.

Figure 35 : Applications sources des tweets




(Béguelin 2016)

Instagram occupe donc beaucoup de place dans le trafic de Twitter, et dans le panorama des réseaux sociaux. Il a en effet relégué Twitter à la troisième place, en se hissant juste derrière Facebook en nombre d'utilisateurs actifs. Actuellement, Instagram compte plus de 400 millions d'utilisateurs actifs par mois[9], et Twitter 320 millions[10]. Les personnalités les plus influentes

[9] INSTAGRAM, 2015. Celebrating a community of 400 Million. *Blog Instagram* [en ligne]. 22 septembre 2015. [Consulté le 15 janvier 2016]. Disponible à l'adresse : http://blog.instagram.com/post/129662501137/150922-400million

[10] TWITTER, 2016. Utilisation de Twitter : les chiffres de l'entreprise. *Twitter* [en ligne]. 2016. [Consulté le 15 janvier 2016]. Disponible à l'adresse : https://about.twitter.com/fr/company

sur Twitter ont d'ailleurs reçu début 2015 un message du réseau social, leur demandant de ne plus passer par Instagram pour joindre une photo à leurs tweets, mais d'utiliser l'outil d'intégration de photo directement présent dans Twitter, afin de ralentir la progression d'Instagram[11].

Twitter est un réseau social très utilisé par les professionnels de l'information et de la communication, les politiciens et les « people », mais a du mal à séduire le grand public. Pour y remédier, le réseau social a annoncé début 2016 être en train de tester une fonctionnalité permettant de se soustraire à la règle limitant les tweets à 140 caractères, qui fait pourtant partie de l'identité de Twitter[12]. L'avenir dira si le petit oiseau bleu prendra son envol, ou si au contraire le nombre d'utilisateurs continuera de stagner.

[11]FROMENT, Etienne, 2015. Twitter demande à ses membres d'arrêter de poster des liens vers Instagram. *Le Soir* [en ligne]. 23 janvier 2015. [Consulté le 7 janvier 2016]. Disponible à l'adresse : http://geeko.lesoir.be/2015/01/23/twitter-demande-a-ses-membres-darreter-de-poster-des-liens-vers-instagram/

[12] AWP, 2016. Twitter sur le point d'abandonner les 140 caractères. *Bilan* [en ligne]. 6 janvier 2016. [Consulté le 15 janvier 2016]. Disponible à l'adresse : http://www.bilan.ch/techno/twitter-point-dabandonner-140-caracteres

# 8. Utilisateurs uniques et profils

En nous basant sur le corpus de 36'904 tweets récoltés entre le 28 avril et le 26 octobre 2015 géolocalisés à Genève (rayon de 20 km), nous avons dénombré 6'365 utilisateurs uniques.

La moitié de ces utilisateurs (50.95%) n'a envoyé qu'un seul tweet durant les six mois de notre capture, et la grande majorité (92.69%) n'a pas envoyé plus de dix tweets.

Toutefois, certains utilisateurs se situent aux extrêmes, et ont envoyé plusieurs centaines de tweets, voire des milliers.

Le tableau ci-dessous représente la répartition du nombre de tweets envoyés par utilisateur :

Tableau 11 : Nombre de tweets envoyés par utilisateur unique

| | Nombre d'utilisateurs uniques | Pourcentage du total d'utilisateurs uniques |
|---|---|---|
| Envoyé 1 tweet | 3243 | 50.95% |
| Envoyé 2 tweets | 1125 | 17.67% |
| Envoyé 3 tweets | 582 | 9.14% |
| Envoyé 4 tweets | 310 | 4.87% |
| Envoyé 5 tweets | 199 | 3.13% |
| Envoyé entre 6 et 10 tweets | 441 | 6.93% |
| Envoyé entre 11 et 15 tweets | 165 | 2.59% |
| Envoyé entre 16 et 20 tweets | 76 | 1.19% |
| Envoyé entre 21 et 30 tweets | 71 | 1.12% |
| Envoyé entre 31 et 40 tweets | 43 | 0.68% |
| Envoyé entre 41 et 50 tweets | 25 | 0.39% |
| Envoyé entre 51 et 100 tweets | 53 | 0.83% |
| Envoyé entre 101 et 150 tweets | 12 | 0.19% |
| Envoyé entre 151 et 200 tweets | 6 | 0.09% |
| Envoyé entre 201 et 300 tweets | 3 | 0.05% |

| | | |
|---|---|---|
| Envoyé entre 301 et 400 tweets | 3 | 0.05% |
| Envoyé entre 401 et 500 tweets | 2 | 0.03% |
| Envoyé entre 501 et 600 tweets | 1 | 0.02% |
| Envoyé entre 601 et 700 tweets | 2 | 0.03% |
| Envoyé entre 701 et 800 tweets | 0 | |
| Envoyé entre 801 et 900 tweets | 1 | 0.02% |
| Envoyé entre 901 et 1000 tweets | 0 | |
| Envoyé plus de 1000 tweets | 2 | 0.03% |

(Béguelin 2016)

## 8.1 Qui tweete le plus à Genève ?

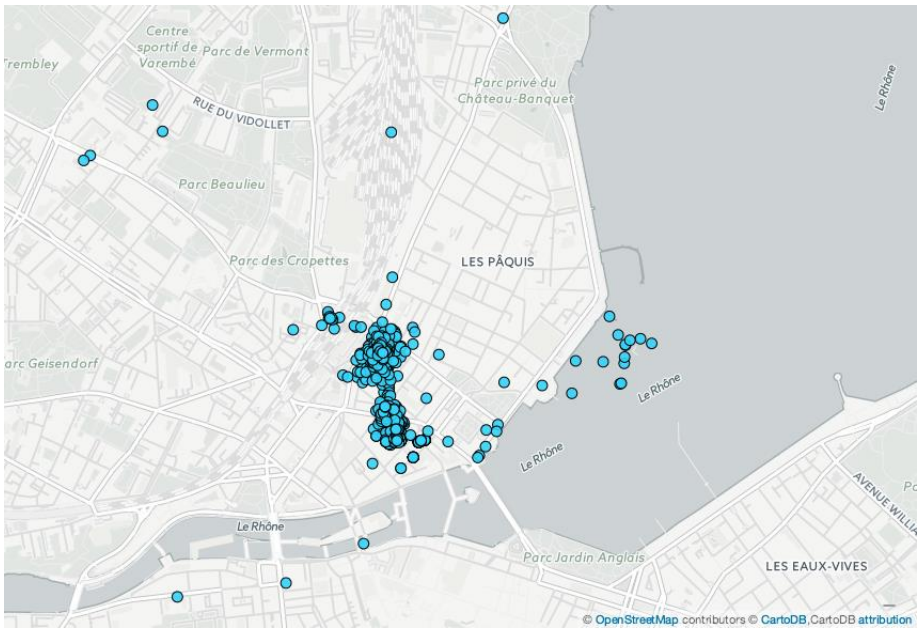
Nous allons nous intéresser aux utilisateurs ayant envoyé plus de 500 tweets en six mois.

Le compte ayant envoyé le plus de tweets à Genève sur cette période (2'882 tweets envoyés) est celui de Busted Genève. Il ne s'agit pas d'un utilisateur humain, mais du compte Twitter de l'application Busted, regroupant toutes les alertes liées aux transports publics (retards, contrôleurs, pickpockets, etc.), envoyées par des utilisateurs géolocalisés.

Le deuxième utilisateur ayant envoyé plus de mille tweets (1'939) est celui d'une véritable personne cette fois. Il s'agit d'une jeune femme japonaise résidant à Genève, qui utilise Twitter pour communiquer avec sa famille et ses amis au Japon. Elle n'utilise pas d'application tierce pour envoyer ses tweets, mais uniquement l'application Twitter pour iPhone. A elle seule, elle représente 93% des tweets en japonais du corpus. Si on l'exclut de nos calculs, le japonais passe de la troisième à la treizième place dans le classement des langues à Genève. Nous avons choisi de la conserver dans notre corpus, car il ne s'agit pas d'un robot publiant des tweets automatiquement, mais bien d'un être humain, qui doit donc faire partie de notre étude.

Comme le montre cette visualisation, elle tweete principalement depuis le quartier de Cornavin.

Figure 36 : Localisation des tweets du deuxième twittos genevois

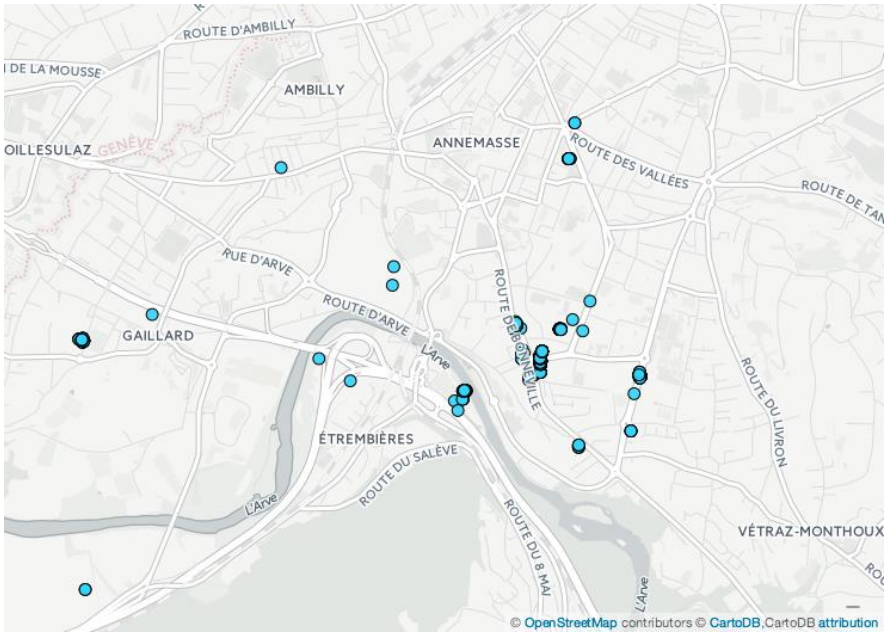



(Béguelin 2016, réalisé avec CartoDB)

L'utilisateur occupant la troisième place des twittos les plus actifs à Genève a envoyé 802 tweets en six mois. Il s'agit d'un jeune homme de 19 ans, résidant en Haute-Savoie et se décrivant comme un youtubeur. Il n'utilise pas non plus d'application tierce pour publier ses tweets, qui proviennent tous de l'application Twitter pour Android. Tous ses tweets sont en français.

La visualisation de la géolocalisation de ses tweets révèle qu'il tweete principalement entre Gaillard et Annemasse.

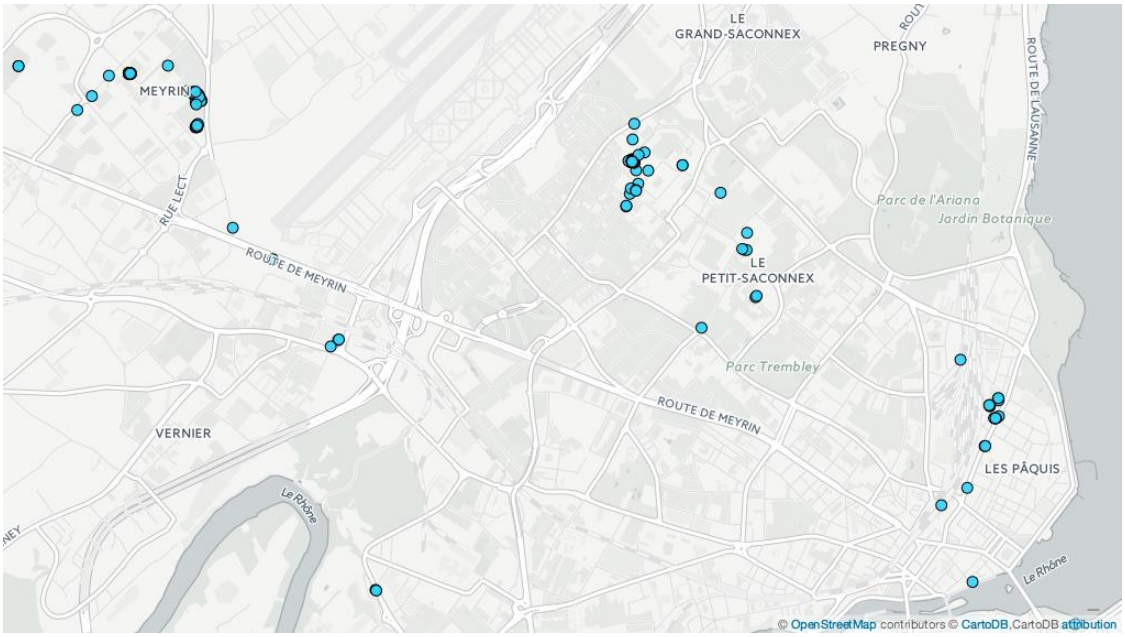
Figure 37 : Localisation des tweets du troisième twittos genevois




(Béguelin 2016, réalisé avec CartoDB)

Avec 665 tweets envoyés en six mois, la quatrième place est tenue par une jeune femme habitant au Grand-Saconnex. Elle tweete en français, et n'utilise pas d'application tierce. Tous ses tweets proviennent de l'application Twitter pour iPhone, et sont envoyés principalement depuis Meyrin et le Grand-Saconnex.

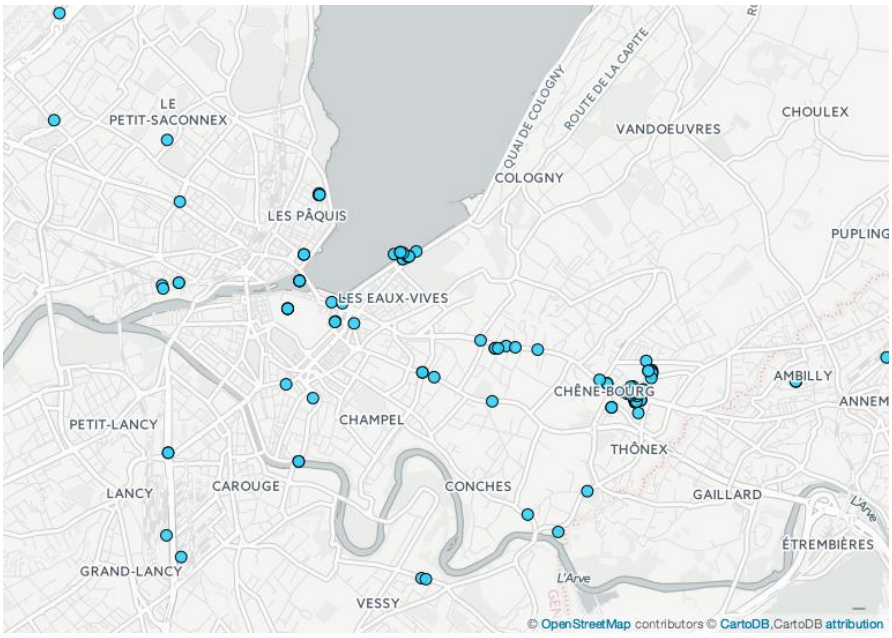
Figure 38 : Localisation des tweets du quatrième twittos genevois




(Béguelin 2016, réalisé avec CartoDB)

La cinquième place est tenue par un jeune homme ayant tweeté 638 fois en six mois. Tous ses tweets proviennent de l'application Twitter pour iPhone, et si la plupart sont en français, un dixième environ est en slovène. La visualisation de ses tweets géolocalisés indique qu'il tweete principalement depuis Saint-Genis-Pouilly, Chêne-Bourg et les Eaux-Vives, mais ses tweets sont répartis sur toute la ville de Genève.

Figure 39 : Localisation des tweets du cinquième twittos genevois

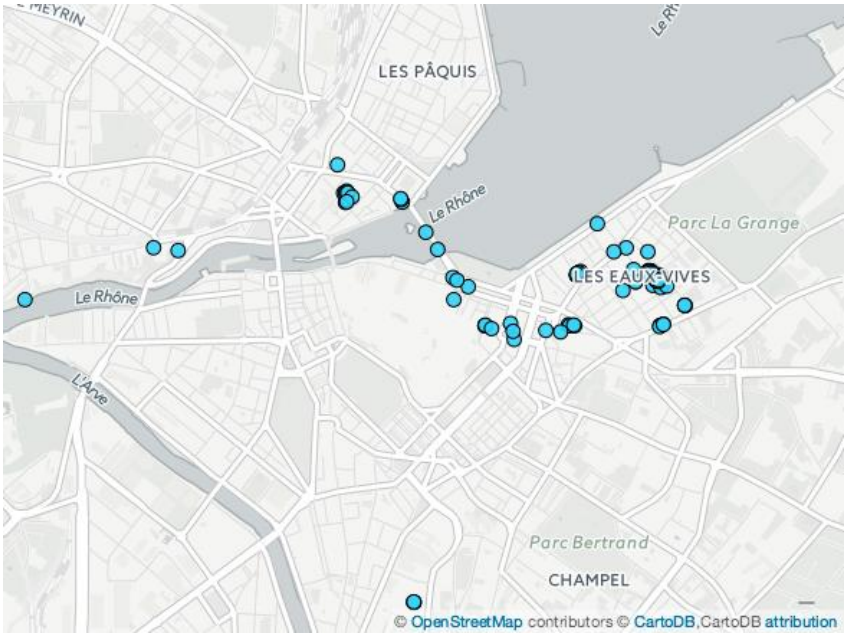



(Béguelin 2016, réalisé avec CartoDB)

La dernière utilisatrice de Twitter ayant envoyé plus de 500 tweets en six mois est une jeune fille habitant à Genève, dont nous comptabilisons 523 tweets dans notre corpus. La plupart de ses tweets sont en français, et environ 10% sont en anglais. Elle utilise exclusivement l'application Twitter pour Android, et ne recourt donc pas à une application tierce.

La visualisation de ses tweets géolocalisés révèle qu'elle tweete principalement entre les Pâquis et les Eaux-Vives.

Figure 40 : Localisation des tweets du sixième twittos genevois




(Béguelin 2016, réalisé avec CartoDB)

Excepté le compte de Busted App, nous pouvons donc constater que les twittos les plus prolifiques de Genève ont des points communs : ils sont tous jeunes, tweetent plusieurs fois par jour pour raconter des anecdotes de leur vie quotidienne, et n'utilisent pas d'application tierce pour envoyer ou gérer leurs tweets. Il ne s'agit donc pas de professionnels des réseaux sociaux, tels que des community managers.

# 9. Problèmes techniques

Nous avons rencontré quelques problèmes techniques au cours de notre projet, que nous jugeons utile de mentionner dans ce rapport car ils peuvent dans certains cas être un biais à nos résultats.

## 9.1 Changement dans l'API de Twitter

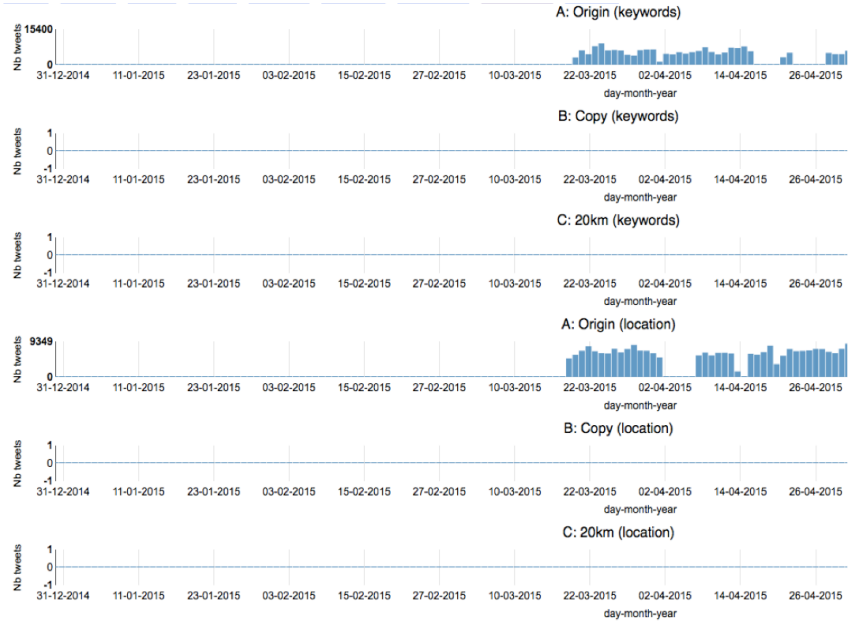
La capture des tweets depuis le streaming public vers les serveurs est en place depuis le 18 mars 2015. Toutefois, le corpus que nous avons utilisé pour nos analyses ne comprend que des tweets à partir du 28 avril 2015. Cela est dû à un changement dans le fonctionnement de l'API de Twitter, survenu le 27 avril 2015. En effet, les tweets géolocalisés provenant de l'application Foursquare étaient auparavant tous dotés de coordonnées géographiques précises, et nous les obtenions donc dans notre capture. A partir du 27 avril, Twitter a mis en application une nouvelle option pour les utilisateurs de Foursquare souhaitant se géolocaliser : au lieu de coordonnées géographiques, ils peuvent être identifiés comme étant dans un restaurant précis, ou un café. Les métadonnées de leur tweet ne comprennent alors plus de coordonnées latitude longitude, ce qui signifie que nous ne les captons plus puisque l'un de nos critères de filtre sur le streaming est la présence de coordonnées géographiques dans les métadonnées du tweet. Un certain nombre d'utilisateurs de Foursquare a continué de transmettre ses coordonnées géographiques précises, mais une part non négligeable a choisi la nouvelle option. Le nombre moyen de tweets géolocalisés a donc chuté le 27 avril 2015. Afin d'avoir un corpus le plus représentatif possible, nous avons choisi d'exclure les tweets antérieurs au 27 avril, et de constituer un corpus de tweets à partir du 28 avril 2015.

## 9.2 Trous dans la capture

L'un des problèmes techniques que nous avons rencontré est la coupure de la capture à cause de l'arrêt temporaire des serveurs. Nous avions anticipé ce problème en mettant en place deux serveurs, afin de dupliquer les données. Toutefois, les coupures étant plus fréquentes que ce que nous pensions, un troisième serveur a été mis en place le 28 août 2015.

Comme le montrent les figures ci-dessous, les coupures les plus importantes ont eu lieu en avril, mai et août.

Figure 41 : Capture des tweets par serveur (mars-avril 2015)




(Béguelin 2016)

Figure 42 : Capture des tweets par serveur (mai-octobre 2015)


(Béguelin 2016)

# 10. Conclusion

Le projet GEoTweet a contribué à la recherche sur les données géolocalisées et leur exploitation scientifique de façon innovante.

Avant tout, il a implémenté des modalités techniques d'accès, de stockage et d'exploitation des tweets géolocalisés à partir du streaming public mis à disposition par Twitter via son API. Ensuite il s'est penché sur l'exploration des tweets géolocalisés dans un contexte spatio-temporel défini, Genève et ses alentours, en vérifiant leur pertinence scientifique et leur relation à des données sociodémographiques classiques. Enfin, il a testé des formes de visualisation des données géolocalisées en considérant leur potentiel analytique.

D'un point de vue méthodologique, le projet a mis en évidence les résultats suivants :

- Bien que l'API de Twitter ne mette à disposition que 1% du trafic des tweets, les échantillons de tweets géolocalisés ainsi obtenus sont souvent exhaustifs grâce au fait que seul un faible pourcentage des tweets est géolocalisé.
- L'outil Solr peut être recommandé pour son efficacité dans le traitement de données géolocalisées. En effet, le projet a testé les possibilités de manipulation de données notamment en se concentrant sur l'extraction des tweets géolocalisés par secteur circulaire (à rayon variable).
- Le projet a également testé les possibilités de visualisation/exploration de l'interface Banana en identifiant des pistes analytiques à l'aide de l'affichage juxtaposé de statistiques de synthèse, de graphiques et de tweets individuels.
- La recherche d'outils de visualisation s'est orientée après plusieurs semaines de tests, vers le logiciel CartoDB pour sa flexibilité, sa gratuité et ses possibilités d'archivage des bases de données et des cartes qui en découlent.

D'un point de vue empirique, les résultats peuvent être synthétisés selon les trois groupes d'hypothèses.

L'axe de recherche concernant les frontières linguistiques virtuelles a permis de mettre en valeur la spécificité des informations géolocalisées provenant de tweets par rapport aux informations statistiques officielles.

En effet, les données sociodémographiques classiques ne permettent pas d'intégrer dans l'analyse des pratiques linguistiques par exemple des touristes, des résidents non déclarés ou des pendulaires. Les tweets géolocalisés donnent des informations sur la population résidente et, en même temps, sur les flux volatiles d'autres catégories d'utilisateurs de l'espace urbain genevois.

- Parmi les tweets géolocalisés dans un rayon de 20 km autour de Genève, l'analyse des langues détectées par Twitter révèle une extraordinaire variété linguistique.
- Parmi les 40 langues détectées, la polarisation entre le français (38.4%) et l'anglais (33.9%) nous renseigne sur le caractère international de la ville. Les langues latines (espagnol, portugais et italien) qui s'élèvent également à 8% témoignent du fait que la ville a été et est encore un pôle d'attraction de main-d'œuvre européenne. Par contre la présence mineure, mais néanmoins non-négligeable, des autres langues met en évidence sa récente ouverture à des flux migratoires extra-européens.

- Les tweets géolocalisés se distribuent de façon homogène parmi les jours de la semaine.
- Nous observons également une variation significative de la production de tweets entre les périodes diurnes et nocturnes. La production de tweets suit les phases de la vie ordinaire d'un citoyen genevois car elle est plus importante pendant les heures de travail que pendant les moments de repos. En effet, la majorité de la population tweete pendant les heures diurnes au travail ou à l'école.
- Les différences de comportement et d'utilisation des espaces urbains des communautés linguistiques virtuelles sont observables via l'analyse visuelle dans CartoDB des tweets géolocalisés. Les utilisateurs qui tweetent en anglais se concentrent dans les zones internationales de la ville telles que le parc de l'Ariana, la place des Nations Unies, l'Organisation mondiale du commerce, l'Hôtel intercontinental Genève, Palexpo et le CERN. Ils sont présents dans les boutiques et les restaurants de la vieille ville. Les utilisateurs qui tweetent en français investissent des espaces plus ordinaires tels que les écoles, les supermarchés, les Pâquis, le bord du lac, Carouge et les arrêts de bus. Les utilisateurs tweetent en italien souvent dans des restaurants et pizzerias. Ceux tweetant en russe montrent une préférence pour les bords du Rhône au centre-ville et les lieux les plus touristiques. Les utilisateurs qui tweetent en arabe se polarisent entre des lieux très populaires et d'autres plus exclusifs en témoignant la stratification sociodémographique de cette population de twittos (touristes du Golfe, immigrés économiques et/ou réfugiés). Les utilisateurs qui tweetent en portugais ont une préférence pour les cafés et les bistrots et leurs tweets proviennent souvent de la zone entre la Servette et les Charmilles. Les utilisateurs qui tweetent en espagnol produisent leurs tweets à partir du Jardin Anglais, des Acacias, de la Jonction et des Pâquis.

Le deuxième axe de recherche qui concerne les convergences et les divergences entre les frontières géopolitiques et les frontières virtuelles nous a permis de souligner leur perméabilité et leur influence réciproque.

- Les territoires français au-delà de la frontière orientale et méridionale du canton de Genève, c'est-à-dire Annemasse et Saint-Julien, affichent clairement leur identité française avec 72% de tweets géolocalisés en français.
- Dans les régions centre-méridionales du territoire interne au canton de Genève, la présence de l'anglais et le multilinguistique augmente. La variété linguistique au centre-ville et à Versoix s'accentue au détriment du français pour atteindre les pourcentages les plus élevés à Ferney-Voltaire et autour de l'aéroport international de Genève.
- Les données révèlent une influence du caractère multinational des populations gravitant autour du canton de Genève sur les villes françaises.
- Au sein de la ville de Genève, selon les quartiers, la distribution des tweets et leur composition linguistique varient. Dans les zones de Plainpalais, la Jonction et le Rhône nous avons une prévalence de l'anglais suivi par le français et une présence importante du portugais. Les utilisateurs tweetent à partir d'espaces publics, notamment des cafés et de l'Université. Au Grand-Saconnex, l'anglais est également plus présent que le français avec une variété linguistique importante (espagnol et arabe), les tweets se concentrent autour des Nations Unies et des parcs publics adjacents. Entre la vieille ville et le Jardin Anglais, nous retrouvons également une forte présence de l'anglais suivi du français avec une variété linguistique qui reste importante.

- Par contre, aux Eaux-Vives et entre les quartiers de Châtelaine et des Charmilles le français est plus présent que l'anglais et les utilisateurs tweetent à partir des lieux très ordinaires (par exemple les préaux des écoles). La zone du cimetière St-Georges est presque exclusivement francophone.
- La comparaison entre la distribution des tweets géolocalisés à Lausanne et à Genève révèle une plus grande diversité linguistique à Genève. Toutefois, parmi les dix premières langues des tweets, les deux villes affichent le français, l'anglais, le portugais, l'espagnol, l'arabe et l'italien. Seuls l'allemand et l'estonien sont plus présents à Lausanne, tandis qu'à Genève le russe et le japonais prennent leur place parmi les dix premières langues des tweets.
- Les tweets genevois sont plus concentrés au centre-ville, alors que les tweets lausannois sont plus répartis.
- Genève compte plus de "gros" utilisateurs que Lausanne. Le corpus lausannois ne comprend donc pas de twittos intensifs, contrairement à celui de Genève.

Le troisième axe de recherche qui concerne la qualité de données récoltées et leur fiabilité a produit des résultats sur l'intégration entre différents media sociaux géolocalisés, sur les procédures d'élimination des robots et sur l'importance des sauvegardes pour le stockage.

- Les analyses montrent comment une part importante de tweets sont émis à partir d'autres applications que Twitter, notamment Instagram.
- La procédure rigoureuse d'élimination des robots doit nécessairement passer par l'analyse individualisée des comptes des utilisateurs qui sont assimilables par leur comportement à des robots.
- La pratique de stockage doit impérativement prévoir la sauvegarde parallèle des données sur différents serveurs.

Le projet GEoTweet ouvre des perspectives additionnelles d'exploration et d'analyse des données. Il serait par exemple envisageable d'approfondir l'analyse textuelle des tweets grâce au text-mining et à la constitution de clusters sémantiques. Cela pourrait entre autres vérifier la relation entre la hausse du nombre de tweets pendant les mois estivaux et l'afflux des touristes en ville.

L'analyse des mots-clés contenus dans les tweets pourrait également servir à étudier le rayonnement de Genève en Europe et dans le monde.

Ce projet de recherche est le point de départ d'une série d'événements grand public visant à la vulgarisation du big data. Il sera notamment le représentant de la Haute école de gestion dans le cadre de « Mapping : l'événement HES » qui se tiendra au mois d'avril 2016.

# Bibliographie

Note : Un état de l'art sur la data visualisation, les méthodes de captures et l'archivage de Twitter, réalisé par Elisa Banfi, Fanny Béguelin et Romaine Kaufmann est disponible en annexe.


**Références applications sources**

AWP, 2016. Twitter sur le point d'abandonner les 140 caractères. *Bilan* [en ligne]. 6 janvier 2016. [Consulté le 15 janvier 2016]. Disponible à l'adresse : http://www.bilan.ch/techno/twitter-point-dabandonner-140-caracteres

FROMENT, Etienne, 2015. Twitter demande à ses membres d'arrêter de poster des liens vers Instagram. *Le Soir* [en ligne]. 23 janvier 2015. [Consulté le 7 janvier 2016]. Disponible à l'adresse : http://geeko.lesoir.be/2015/01/23/twitter-demande-a-ses-membres-darreter-de-poster-des-liens-vers-instagram/

INSTAGRAM, 2015. Celebrating a community of 400 Million. *Blog Instagram* [en ligne]. 22 septembre 2015. [Consulté le 15 janvier 2016]. Disponible à l'adresse : http://blog.instagram.com/post/129662501137/150922-400million

TWITTER, 2016. Utilisation de Twitter : les chiffres de l'entreprise. *Twitter* [en ligne]. 2016. [Consulté le 15 janvier 2016]. Disponible à l'adresse : https://about.twitter.com/fr/company

**Références statistiques**

REPUBLIQUE ET CANTON DE GENEVE, 2015. Statistiques cantonales, les 21 domaines : 01. Population. *Ge.ch* [en ligne]. 2015. [Consulté le 27 décembre 2015]. Disponible à l'adresse : http://www.ge.ch/statistique/domaines/01/01_05/tableaux.asp#1

**Références data visualisation**

ARAMO-IMMONEN, Heli, JUSSILA, Jari et HUHTAMÄKI, Jukka, 2015. Exploring Co-learning Behavior of Conference Participants with Visual Network Analysis of Twitter Data. In : Computers in Human Behavior [en ligne]. mars 2015. [Consulté le 14 juin 2015]. DOI 10.1016/j.chb.2015.02.033. Disponible à l'adresse : http://linkinghub.elsevier.com/retrieve/pii/S0747563215001375 [accès par abonnement].

BARNES, Robert, 2014. Trendsmap Plus: More Trends, Faster Interface and Facebook Content. In : Trendsmap Blog [en ligne]. 11 avril 2014. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://blog.trendsmap.com/2014/11/trends-enhancements-trendsmap-plus.

BAUDOT, Jean-Charles, 2014. See the WDMTG Graph of a Tweet by Jcbaudot. In : app.wdmtg.com [en ligne]. 10 novembre 2014. [Consulté le 14 juin 2015]. Disponible à l'adresse : https://app.wdmtg.com/tweet/531863873780191232/546a32d104e28.

CAO, Nan, LIN, Yu-Ru, SUN, Xiaohua, LAZER, David, LIU, Shixia et QU, Huamin, 2012. Whisper: Tracing the Spatiotemporal Process of Information Diffusion in Real Time. In : Visualization and Computer Graphics, IEEE Transactions [en ligne]. 2012. Vol. 18, n° 12, pp. 2649–2658. [Consulté le 14 juin 2015]. DOI 10.1109/TVCG.2012.291. Disponible à l'adresse : http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=6327271 [accès par abonnement].

CNN ECOSPHERE Project, 2011. CNN ECOSPHERE Project [en ligne]. 17 novembre 2011. [Consulté le 14 juin 2015]. Disponible à l'adresse : https://www.youtube.com/watch?v=60WILmaDA-s.

DERMOUCHE, Mohamed, KHOUAS, Leila, LOUDCHER, Sabine, VELCIN, Julien et FOURBOUL, Eric, 2015. Analyse et visualisation d'opinions dans un cadre de veille sur le Web. In : 15èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2015, 27-30 Janvier 2015, Luxembourg [en ligne]. Luxembourg : EGC 2015, pp. 461‑466. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://editions-rnti.fr/?inprocid=1002110 [accès par abonnement].

EKIMETRICS, 2015. DataMatch - Le FN tient-il un double discours? In : ParisMatch.com [en ligne]. 27 mars 2015. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.parismatch.com/Actu/Politique/Le-FN-tient-il-un-double-discours.

FISCHER, Eric, 2014a. Locals & Tourists. In : https://www.mapbox.com/ [en ligne]. 1 décembre 2014. [Consulté le 14 juin 2015]. Disponible à l'adresse : https://www.mapbox.com/labs/twitter-gnip/locals/#5/38.013/-95.032.

FISCHER, Eric, 2014b. Making the Most Detailed Tweet Map Ever. Mapbox.com [en ligne]. 3 décembre 2014. [Consulté le 31 mars 2015]. Disponible à l'adresse : https://www.mapbox.com/blog/twitter-map-every-tweet.

FONSECA, Margarida, 2012. http://flowingcity.com. In : http://flowingcity.com [en ligne]. 2012. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://flowingcity.com/the-project.

FUJITA, Hideyuki, 2013. Geo-tagged Twitter Collection and Visualization System. In : Cartography and Geographic Information Science. juin 2013. Vol. 40, n° 3, pp. 183‑191. [Consulté le 14 juin 2015]. DOI 10.1080/15230406.2013.800272. Disponible à l'adresse : http://www.tandfonline.com/doi/abs/10.1080/15230406.2013.800272?journalCode=tcag20 [accès par abonnement].

GHINN, Daniel, [2015]. Worldwide Doctors & HCPs on Twitter: Now Explore the Data. In : Creation Pinpoint [en ligne]. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.creationpinpoint.com/worldwide-doctors-on-twitter-explore-the-data.

GUILLE, Adrien et FAVRE, Cécile, 2015. Event Detection, Tracking, and Visualization in Twitter: a Mention-anomaly-based Approach. In : Social Network Analysis and Mining [en ligne]. décembre 2015. Vol. 5, n° 1, pp. 1-17 [Consulté le 14 juin 2015]. DOI 10.1007/s13278-015-0258-0. Disponible à l'adresse : http://link.springer.com/10.1007/s13278-015-0258-0 [accès par abonnement].

HÜGEL, Stephan et ROUMPANI Flora, 2014. Urschrei/CityEngine-Twitter.Urschrei.github.io[en ligne]. 4 mai 2014. [Consulté le 14 juin 2015]. DOI 10.5281/zenodo.9795. Disponible à l'adresse : http://urschrei.github.io/CityEngine-Twitter.

HURON, Samuel, VUILLEMOT, Romain et FEKETE, Jean-Daniel, 2014. La sédimentation visuelle. Outil et technique pour visualiser les flux de données à destination du grand public. In : Ingénierie des Systèmes d'Information. 2014. Vol. 19, n° 3, pp. 155‑158.

JACKSON, Simon, 2014. Language and Swearing Map of Australia: Interactive. In : The Guardian [en ligne]. 30 juillet 2014. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.theguardian.com/news/datablog/interactive/2014/jul/30/language-swearing-map-australia [accès par abonnement].

LÜFKENS, Matthias, 2015. Twiplomacy Study 2015 | Twiplomacy. In : http://twiplomacy.com/ [en ligne]. 28 avril 2015. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://twiplomacy.com/blog/twiplomacy-study-2015.

LULIC, I. et KOVIC, I., 2013. Analysis of Emergency Physicians' Twitter Accounts. In : Emergency Medicine Journal. 1 mai 2013. Vol. 30, n° 5, pp. 371‑376. [Consulté le 14 juin 2015]. DOI 10.1136/emermed-2012-201132. Disponible à l'adresse : http://emj.bmj.com/content/30/5/371.long [accès par abonnement].

MAIN, Guillaume, 2013. L'outil Where Does My Tweet Go calcule le score de propagation de tous vos tweets. In : Statosphere [en ligne]. 6 juin 2013. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://statosphere.fr/outils/2013/06/06/wdmtg-algorithme-tweet-score-spreadrank-mfg-labs-twitter.

MARSHALL, Aarian, 2015. Why You Should Be Skeptical of Most Twitter Maps. In : CityLab [en ligne]. 26 mars 2015. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.citylab.com/housing/2015/03/why-most-twitter-maps-cant-be-trusted/388586/.

NGUYEN, Vu Dung, VARGHESE, Binni et BARKER, Adam, 2013. The Royal Birth of 2013: Analysing and Visualising Public Sentiment in the UK Using Twitter. In : Big Data, 2013 IEEE

International Conference on [en ligne]. IEEE. 2013. pp. 46–54. [Consulté le 14 juin 2015]. arXiv:1308.1847. Disponible à l'adresse : http://arxiv.org/abs/1308.1847.

PALMER, Stuart, 2013. Characterisation of the Use of Twitter by Australian Universities. In : Journal of Higher Education Policy and Management. 2013. Vol. 35, n° 4, pp. 333‑344. [Consulté le 14 juin 2015]. DOI 10.1080/1360080X.2013.812029. Disponible à l'adresse : http://www.tandfonline.com/doi/abs/10.1080/1360080X.2013.812029 http://emj.bmj.com/content/30/5/371.long.

ROGERS, Simon, 2014. CartoDB · How News of #Ferguson Spread across Twitter. In : http://srogers.cartodb.com [en ligne]. 2014. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://srogers.cartodb.com/viz/4a5eb582-23ed-11e4-bd6b-0e230854a1cb/embed_map.

SHELTON, Taylor, POORTHUIS, Ate et ZOOK, Matthew, 2015. Social Media and the City: Rethinking Urban Socio-spatial Inequality Using User-generated Geographic Information. In : Landscape and Urban Planning [en ligne]. 2015. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.sciencedirect.com/science/article/pii/S0169204615000523 http://emj.bmj.com/content/30/5/371.long.

TRENDSMAP, [sans date]. Real-time local Twitter trends - Trendsmap. In : Trendsmap [en ligne]. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://trendsmap.com.

TWEETPING, [sans date]. Tweetping. In : Tweetping [en ligne]. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://tweetping.net.

WAGNER STEFAN, 2015. Sociotope.me. In : https://vimeo.com/ [en ligne]. 17 février 2015. [Consulté le 14 juin 2015]. Disponible à l'adresse : https://vimeo.com/119835746.

WDMTG, 2013. WDMTG - Where Does My Tweet Go. In : WDMTG [en ligne]. 2013. [Consulté le 14 juin 2015]. Disponible à l'adresse : https://wdmtg.com/.

WUEEST, Bruno et MÜLLER, Christian, 2014. Tweetocracy Switzerland: Exploring the evolution, Representativeness and Structuration of Swiss Party Politics on Twitter. In : [en ligne]. 2014. [Consulté le 13 avril 2015]. Disponible à l'adresse : http://www.bruno-wueest.ch/files/imir_bw_cm_20140810.pdf.

ZHAO, Jian, CAO, Nan, WEN, Zhen, SONG, Yale, LIN, Yu-Ru et COLLINS, Christopher, 2014. #FluxFlow: Visual Analysis of Anomalous Information Spreading on Social Media. In : IEEE Transactions on Visualization and Computer Graphics. 31 décembre 2014. Vol. 20, n° 12, pp. 1773‑1782. [Consulté le 14 juin 2015]. DOI 10.1109/TVCG.2014.2346922. Disponible à l'adresse : www.nancao.org/pubs/zhao_vast14_paper.pdf.

**Références méthodes de capture**

*Ouvrages et articles*

COTELO, J.M., CRUZ, F.L., TROYANO, J.A., ORTEGA, F.J., 2015. A modular approach for lexical normalization applied to Spanish tweets. Expert Systems with Applications [en ligne]. Février 2015. Vol. 42, pp. 4743-4754. [Consulté le 13 juin 2015]. Disponible à l'adresse : http://www.sciencedirect.com/science/article/pii/S0957417415000962 [accès par abonnement].

EISENSTEIN, Jacob, 2013. What to do about bad language on the internet. Georgia Institute of Technology NAACL, 2013 [en ligne]. 11p. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.cc.gatech.edu/~jeisenst/papers/naacl2013-badlanguage.pdf.

FRIAS-MARTINEZ, Vanessa, FRIAS-MARTINEZ, Enrique, 2014. Spectral clustering for sensing urban land use using Twitter activity. Engineering Applications of Artificial Intelligence [en ligne]. Juillet 2014. Volume 35, pp. 237-245. [Consulté le 13 juin 2015]. Disponible à l'adresse : http://www.sciencedirect.com/science/article/pii/S0952197614001419 [accès par abonnement].

MORSTATTER, Fred, PFEFFER, Jürgen, LIU, Huan, et al. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. arXiv preprint

arXiv:1306.5204, 2013. [Consulté le 30 mars 2015]. Disponible à l'adresse : http://arxiv.org/pdf/1306.5204v1.pdf.

*Conférences*

KONDOR, Daniel et al., 2013. Using Robust PCA to estimate regional characteristics of language use from geo-tagged Twitter messages. In : COGINFOCOM 2013. 4th IEEE International Conference on Cognitive Infocommunications, Budapest (Hongrie), 2-5 décembre 2013 [en ligne]. [Consulté le 13 juin 2015]. Disponible à l'adresse : http://apps.webofknowledge.com/full_record.do?product=INSPEC&search_mode=GeneralSearch&qid=4&SID=R2oIFEwvMmH365JzQWR&page=1&doc=6 [accès par abonnement].

THOM, Dennis, BOSCH, Harald, KOCH, Steffen, WÖRNER, Michael, ERTL, Thomas, 2012. Spatiotemporal Anomaly Detection through Visual Analysis of Geolocated Twitter Messages. In : IEEE Pacific Visualization Symposium, Songdo (Corée), 28 février – 2 mars 2012 [en ligne]. Stuttgart : University of Stuttgart, 2012, pp. 41-48. [Consulté le 13 juin 2015]. Disponible à l'adresse : http://apps.webofknowledge.com/full_record.do?product=INSPEC&search_mode=GeneralSearch&qid=7&SID=R2oIFEwvMmH365JzQWR&page=1&doc=2 [accès par abonnement].

ZHAO, Yanchang, 2015. Text Mining with R – Twitter Data Analysis [en ligne]. Melbourne : Deakin University, 28 mai 2015. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.rdatamining.com/docs/text-mining-with-r-of-twitter-data-analysis.

# Annexe 1 : Tableau des tweets par heures de la journée et par jours de la semaine

| Heures | lundi | mardi | mercredi | jeudi | vendredi | samedi | dimanche | Total |
|---|---|---|---|---|---|---|---|---|
| 0 | 46 | 45 | 33 | 64 | 92 | 92 | 106 | 478 |
| 1 | 23 | 44 | 35 | 73 | 41 | 51 | 46 | 313 |
| 2 | 24 | 41 | 22 | 27 | 39 | 36 | 24 | 213 |
| 3 | 31 | 30 | 32 | 20 | 25 | 51 | 37 | 226 |
| 4 | 71 | 57 | 50 | 54 | 52 | 57 | 41 | 382 |
| 5 | 139 | 151 | 171 | 149 | 94 | 120 | 95 | 919 |
| 6 | 168 | 193 | 231 | 186 | 190 | 154 | 108 | 1230 |
| 7 | 185 | 170 | 206 | 182 | 199 | 210 | 183 | 1335 |
| 8 | 220 | 229 | 253 | 232 | 251 | 260 | 218 | 1663 |
| 9 | 263 | 273 | 312 | 262 | 303 | 308 | 296 | 2017 |
| 10 | 301 | 290 | 302 | 264 | 312 | 329 | 350 | 2148 |
| 11 | 326 | 285 | 304 | 306 | 303 | 337 | 289 | 2150 |
| 12 | 265 | 289 | 289 | 312 | 309 | 327 | 331 | 2122 |
| 13 | 273 | 248 | 266 | 292 | 303 | 329 | 313 | 2024 |
| 14 | 282 | 276 | 302 | 294 | 315 | 314 | 301 | 2084 |
| 15 | 322 | 307 | 307 | 375 | 316 | 317 | 260 | 2204 |
| 16 | 297 | 316 | 362 | 319 | 349 | 341 | 286 | 2270 |
| 17 | 265 | 305 | 313 | 354 | 303 | 371 | 319 | 2230 |
| 18 | 279 | 311 | 344 | 323 | 318 | 299 | 343 | 2217 |
| 19 | 269 | 328 | 333 | 321 | 276 | 315 | 306 | 2148 |
| 20 | 290 | 358 | 376 | 280 | 264 | 349 | 310 | 2227 |
| 21 | 246 | 269 | 314 | 252 | 212 | 284 | 286 | 1863 |
| 22 | 185 | 222 | 206 | 275 | 163 | 219 | 219 | 1489 |
| 23 | 106 | 119 | 113 | 144 | 182 | 163 | 125 | 952 |
| Total | 4876 | 5156 | 5476 | 5360 | 5211 | 5633 | 5192 | 36904 |

# Annexe 2 : Population résidante en 2000 selon domicile civil, commune (Genève et Lausanne) et langue principale

| *Domicile civil* | **Ville de Lausanne** | | **Ville de Genève** | |
|---|---|---|---|---|
| Langues | **Total** | % | **Total** | % |
| **Allemand** | 4118 | 3.45 | 6208 | 3.55 |
| **Français** | 94896 | 79.57 | 126854 | 72.51 |
| **Italien** | 4197 | 3.52 | 6977 | 3.99 |
| **Romanche** | 55 | 0.05 | 111 | 0.06 |
| **Anglais** | 1796 | 1.51 | 7828 | 4.47 |
| **Néerlandais** | 221 | 0.19 | 361 | 0.21 |
| **Espagnol** | 3409 | 2.86 | 7461 | 4.26 |
| **Portugais** | 3152 | 2.64 | 6620 | 3.78 |
| **Autres langues d'Europe de l'Ouest** | 8 | 0.01 | 4 | 0.00 |
| **Danois** | 43 | 0.04 | 132 | 0.08 |
| **Norvégien** | 52 | 0.04 | 85 | 0.05 |
| **Suédois** | 174 | 0.15 | 300 | 0.17 |
| **Finnois** | 43 | 0.04 | 128 | 0.07 |
| **Autres langues d'Europe du Nord** | 4 | 0.00 | 15 | 0.01 |
| **Serbe et croate** | 1331 | 1.12 | 1001 | 0.57 |
| **Russe** | 298 | 0.25 | 1660 | 0.95 |
| **Polonais** | 156 | 0.13 | 257 | 0.15 |
| **Tchèque** | 84 | 0.07 | 112 | 0.06 |
| **Slovaque** | 37 | 0.03 | 52 | 0.03 |
| **Macédonien** | 34 | 0.03 | 27 | 0.02 |
| **Slovène** | 6 | 0.01 | 22 | 0.01 |
| **Bulgare** | 44 | 0.04 | 167 | 0.10 |
| **Langues slaves restantes** | 0 | 0.00 | 1 | 0.00 |
| **Albanais** | 1248 | 1.05 | 1249 | 0.71 |
| **Turc** | 394 | 0.33 | 474 | 0.27 |
| **Autres langues d'Europe de l'est** | 9 | 0.01 | 52 | 0.03 |
| **Hongrois** | 126 | 0.11 | 244 | 0.14 |
| **Roumain** | 215 | 0.18 | 256 | 0.15 |
| **Grec** | 129 | 0.11 | 238 | 0.14 |
| **Autres langues européennes** | 0 | 0.00 | 3 | 0.00 |
| **Langues africaines** | 457 | 0.38 | 859 | 0.49 |
| **Arabe** | 635 | 0.53 | 1508 | 0.86 |
| **Langues d'Asie de l'Ouest** | 459 | 0.38 | 796 | 0.45 |
| **Langues indo-aryennes et dravidiennes** | 571 | 0.48 | 670 | 0.38 |
| **Langues d'Asie de l'est** | 785 | 0.66 | 2100 | 1.20 |
| **Autres indications de langues** | 76 | 0.06 | 120 | 0.07 |
| **Total** | 119262 | 100.00 | 174952 | 100.00 |

# Annexe 3a : Applications sources des tweets corpus Genève 10 km

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Instagram | 13262 | 39.34% | | |
| 2 | Twitter | 12125 | 35.96% | | |
| 3 | Foursquare | 3119 | 9.25% | | |
| 4 | Busted App | 2882 | 8.55% | | |
| 5 | iOS | 632 | 1.87% | | |
| 6 | TweetMyJOBS | 580 | 1.72% | | |
| 7 | Path | 307 | 0.91% | | |
| 8 | Tweetbot for iOS | 170 | 0.50% | | |
| 9 | SafeTweet by TweetMyJOBS | 71 | 0.21% | | |
| 10 | Endomondo | 60 | 0.18% | | |
| 11 | Hootsuite | 43 | 0.13% | | |
| 12 | locavoresco | 34 | 0.10% | | |
| 13 | OS X | 33 | 0.10% | | |
| 14 | Docannonce | 32 | 0.09% | | |
| 15 | chat-perdu.org | 31 | 0.09% | | |
| 16 | Been there\ | 29 | 0.09% | | |
| 17 | DanceDeets | 26 | 0.08% | | |
| 18 | WordPress.com | 26 | 0.08% | | |
| 19 | Untappd | 25 | 0.07% | | |
| 20 | Fenix for Android | 20 | 0.06% | | |
| 21 | MeTweets for Windows Phone | 16 | 0.05% | | |
| 22 | Hipstamatic | 15 | 0.04% | | |
| 23 | TweetCaster for Android | 14 | 0.04% | | |
| 24 | Plume for Android | 12 | 0.04% | | |
| 25 | Tweet It! for WP8 | 12 | 0.04% | | |
| 26 | Bahnbilder.de | 11 | 0.03% | | |
| 27 | twitterfeed | 11 | 0.03% | | |
| 28 | chien-perdu.org | 10 | 0.03% | | |
| 29 | Slices Pro for iPhone | 9 | 0.03% | | |
| 30 | Squarespace | 9 | 0.03% | | |
| 31 | Remixjobs | 8 | 0.02% | | |
| 32 | Tweetbot for Mac | 8 | 0.02% | | |
| 33 | 大空お天気 | 8 | 0.02% | | |
| 34 | Ecotour.com Deals | 6 | 0.02% | | |
| 35 | Everypost | 6 | 0.02% | | |
| 36 | ptext-API2 | 6 | 0.02% | | |
| 37 | TweetList! | 5 | 0.01% | | |
| 38 | Twitterrific | 4 | 0.01% | | |
| 39 | Hackathons Near Me | 3 | 0.01% | | |
| 40 | JenniferSpency App | 3 | 0.01% | | |
| 41 | Karen Smith FL App | 3 | 0.01% | | |
| 42 | MontreuxJazzFestival2015 | 3 | 0.01% | | |
| 43 | SOTAwatch | 3 | 0.01% | | |
| 44 | Day One | 2 | 0.01% | | |
| 45 | Hautetfort | 2 | 0.01% | | |
| 46 | Kalimmo | 2 | 0.01% | | |
| 47 | MyRegus | 2 | 0.01% | | |
| 48 | SimonARoberts dot Com | 2 | 0.01% | | |
| 49 | Twittimer | 2 | 0.01% | | |
| 50 | vk.com | 2 | 0.01% | | |
| 51 | ツイタマ for Android | 2 | 0.01% | | |
| 52 | Camera on iOS | 1 | 0.00% | | |
| 53 | Camera+ | 1 | 0.00% | | |
| 54 | Echofon | 1 | 0.00% | | |
| 55 | Mobypicture | 1 | 0.00% | | |
| 56 | Photos on iOS | 1 | 0.00% | | |
| 57 | Simugalicia Web | 1 | 0.00% | | |
| 58 | Squarespace 6 | 1 | 0.00% | | |
| 59 | | | | | |

# Annexe 3b : Applications sources des tweets corpus Lausanne 10 km

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Instagram | 4541 | 46.21% | | | |
| 2 | Twitter | 2503 | 25.47% | | | |
| 3 | Busted App | 1418 | 14.43% | | | |
| 4 | Foursquare | 678 | 6.90% | | | |
| 5 | iOS | 178 | 1.81% | | | |
| 6 | Tweetbot for iOS | 106 | 1.08% | | | |
| 7 | TweetMyJOBS | 91 | 0.93% | | | |
| 8 | Endomondo | 38 | 0.39% | | | |
| 9 | Path | 31 | 0.32% | | | |
| 10 | Slices Pro for iPhone | 29 | 0.30% | | | |
| 11 | Fenix for Android | 27 | 0.27% | | | |
| 12 | Twitterrific | 26 | 0.26% | | | |
| 13 | WordPress.com | 20 | 0.20% | | | |
| 14 | ptext-API2 | 18 | 0.18% | | | |
| 15 | Tweetbot for Mac | 12 | 0.12% | | | |
| 16 | Untappd | 12 | 0.12% | | | |
| 17 | Hootsuite | 11 | 0.11% | | | |
| 18 | Plume for Android | 11 | 0.11% | | | |
| 19 | DanceDeets | 7 | 0.07% | | | |
| 20 | OS X | 6 | 0.06% | | | |
| 21 | SOTAwatch | 5 | 0.05% | | | |
| 22 | Yakaz | 5 | 0.05% | | | |
| 23 | Bahnbilder.de | 4 | 0.04% | | | |
| 24 | Been there\ | 4 | 0.04% | | | |
| 25 | twitterfeed | 4 | 0.04% | | | |
| 26 | #Träwelling | 3 | 0.03% | | | |
| 27 | chat-perdu.org | 3 | 0.03% | | | |
| 28 | Hipstamatic | 3 | 0.03% | | | |
| 29 | Post with Klout | 3 | 0.03% | | | |
| 30 | Squarespace | 3 | 0.03% | | | |
| 31 | WMaker | 3 | 0.03% | | | |
| 32 | GitHubJobs | 2 | 0.02% | | | |
| 33 | MyRegus | 2 | 0.02% | | | |
| 34 | SimonARoberts dot Com | 2 | 0.02% | | | |
| 35 | 大空お天気 | 2 | 0.02% | | | |
| 36 | A Travel Bot | 1 | 0.01% | | | |
| 37 | Blockade2204 | 1 | 0.01% | | | |
| 38 | chien-perdu.org | 1 | 0.01% | | | |
| 39 | Dexigner iPhone App | 1 | 0.01% | | | |
| 40 | Frontback | 1 | 0.01% | | | |
| 41 | Google | 1 | 0.01% | | | |
| 42 | Hautetfort | 1 | 0.01% | | | |
| 43 | JobsHub.IO Site | 1 | 0.01% | | | |
| 44 | MarsBots | 1 | 0.01% | | | |
| 45 | morning_relay | 1 | 0.01% | | | |
| 46 | Squarespace 6 | 1 | 0.01% | | | |
| 47 | Trainspo | 1 | 0.01% | | | |
| 48 | Tweetlogix | 1 | 0.01% | | | |
| 49 | vk.com | 1 | 0.01% | | | |
| 50 | YouTube on iOS_ | 1 | 0.01% | | | |
| 51 | | | | | | |

# Annexe 4 : Applications sources des tweets corpus Genève 20 km

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Archives envoie automatique | 3 | 0.01% | | | | | |
| 2 | Bahnbilder.de | 11 | 0.03% | | | | | |
| 3 | Been there\ | 29 | 0.08% | | | | | |
| 4 | Busted App | 2882 | 7.81% | | | | | |
| 5 | Camera on iOS | 1 | 0.00% | | | | | |
| 6 | Camera+ | 1 | 0.00% | | | | | |
| 7 | chat-perdu.org | 108 | 0.29% | | | | | |
| 8 | chien-perdu.org | 25 | 0.07% | | | | | |
| 9 | DanceDeets | 26 | 0.07% | | | | | |
| 10 | Day One | 2 | 0.01% | | | | | |
| 11 | Docannonce | 45 | 0.12% | | | | | |
| 12 | Echofon | 1 | 0.00% | | | | | |
| 13 | Ecotour.com Deals | 7 | 0.02% | | | | | |
| 14 | Endomondo | 68 | 0.18% | | | | | |
| 15 | Everypost | 6 | 0.02% | | | | | |
| 16 | Fenix for Android | 20 | 0.05% | | | | | |
| 17 | Foursquare | 3304 | 8.95% | | | | | |
| 18 | Hackathons Near Me | 3 | 0.01% | | | | | |
| 19 | Hautetfort | 3 | 0.01% | | | | | |
| 20 | Hipstamatic | 15 | 0.04% | | | | | |
| 21 | Hootsuite | 45 | 0.12% | | | | | |
| 22 | Instagram | 14471 | 39.21% | | | | | |
| 23 | iOS | 702 | 1.90% | | | | | |
| 24 | JenniferSpency App | 3 | 0.01% | | | | | |
| 25 | Kalimmo | 2 | 0.01% | | | | | |
| 26 | Karen Smith FL App | 3 | 0.01% | | | | | |
| 27 | locavoresco | 51 | 0.14% | | | | | |
| 28 | MeTweets for Windows Phone | 16 | 0.04% | | | | | |
| 29 | Mobypicture | 1 | 0.00% | | | | | |
| 30 | MontreuxJazzFestival2015 | 3 | 0.01% | | | | | |
| 31 | MyRegus | 2 | 0.01% | | | | | |
| 32 | OS X | 37 | 0.10% | | | | | |
| 33 | Path | 310 | 0.84% | | | | | |
| 34 | Photos on iOS | 1 | 0.00% | | | | | |
| 35 | Plume for Android | 15 | 0.04% | | | | | |
| 36 | ptext-API2 | 6 | 0.02% | | | | | |
| 37 | Remixjobs | 8 | 0.02% | | | | | |
| 38 | SafeTweet by TweetMyJOBS | 71 | 0.19% | | | | | |
| 39 | SimonARoberts dot Com | 3 | 0.01% | | | | | |
| 40 | Simugalicia Web | 1 | 0.00% | | | | | |
| 41 | Slices Pro for iPhone | 9 | 0.02% | | | | | |
| 42 | SOTAwatch | 21 | 0.06% | | | | | |
| 43 | Squarespace | 9 | 0.02% | | | | | |
| 44 | Squarespace 6 | 1 | 0.00% | | | | | |
| 45 | Tweet It! for WP8 | 12 | 0.03% | | | | | |
| 46 | Tweetbot for iOS | 177 | 0.48% | | | | | |
| 47 | Tweetbot for Mac | 8 | 0.02% | | | | | |
| 48 | TweetCaster for Android | 18 | 0.05% | | | | | |
| 49 | TweetList! | 5 | 0.01% | | | | | |
| 50 | TweetMyJOBS | 580 | 1.57% | | | | | |
| 51 | Twitter for  Android | 2 | 0.01% | | | | | |
| 52 | Twitter for Android | 4332 | 11.74% | | | | | |
| 53 | Twitter for Android Tablets | 64 | 0.17% | | | | | |
| 54 | Twitter for BlackBerry | 170 | 0.46% | | | | | |
| 55 | Twitter for BlackBerry® | 183 | 0.50% | | 37.02% | | | |
| 56 | Twitter for iPad | 142 | 0.38% | | | | | |
| 57 | Twitter for iPhone | 7542 | 20.44% | | | | | |
| 58 | Twitter for Windows | 2 | 0.01% | | | | | |
| 59 | Twitter for Windows Phone | 1224 | 3.32% | | | | | |
| 60 | twitterfeed | 11 | 0.03% | | | | | |
| 61 | Twitterrific | 4 | 0.01% | | | | | |
| 62 | Twittimer | 2 | 0.01% | | | | | |
| 63 | Untappd | 32 | 0.09% | | | | | |
| 64 | vk.com | 2 | 0.01% | | | | | |
| 65 | wezzoo | 3 | 0.01% | | | | | |
| 66 | WordPress.com | 27 | 0.07% | | | | | |
| 67 | Yakaz | 1 | 0.00% | | | | | |
| 68 | ツイタマ for Android | 2 | 0.01% | | | | | |
| 69 | 大空お天気 | 8 | 0.02% | | | | | |
| 70 | | **36904** | 100.00% | | | | | |
| 71 | | | | | | | | |

# Annexe 5 : Bibliographie commentée

## Contexte

La présente bibliographie commentée s'inscrit dans le projet de recherche GEoTweet, détaillé dans le cahier des charges d'avril 2015.

Elle consiste en la réalisation du premier de nos objectifs, rappelés ci-après.

## Objectifs

1. Fournir une revue de la littérature concernant la visualisation des données, les méthodes de capture et l'archivage des tweets

2. Proposer des visualisations
   2.1. Spatialiser la diversité linguistique de Genève en utilisant les tweets géolocalisés
   2.2. Représenter le rayonnement de Genève

3. Comparer les résultats des différentes méthodes de capture utilisées lors des visualisations

## Etat de l'art

Cet état de l'art est séparé en trois axes : visualisation, méthodes de capture, archivage. La bibliographie correspondant à chacun de ces axes est développée à la fin de chaque partie, afin de gagner en clarté et de répartir les sources par sujet.

### Visualisation

La littérature qui s'intéresse aux formes de visualisation de tweets provient de deux sources différentes : d'une part nous retrouvons des publications académiques dans des revues à comité de lecteurs (1) et d'autre part des blogs et des images diffusés par les réseaux sociaux, notamment Twitter et Pinterest (2).

### 1. Publications académiques

La littérature scientifique sur les visualisations de données obtenues par Twitter se divise en deux groupes : le premier traite des données géolocalisées tandis que le second se concentre sur les données non-géolocalisées.

#### *1.1 Tweets géolocalisés*

Notre projet de recherche s'intéresse essentiellement à la première catégorie de données et notamment aux visualisations qui concernent le nombre agrégé de tweets géolocalisés par unité temporelle et/ou spatiale. A travers des graphiques (lignes, nuages de points, secteurs, histogrammes, barres, barres empilées, aires), la variation temporelle des tweets est visualisée dans des aires géographiques définies.

Par exemple, Fujita (2013) utilise des visualisations traditionnelles de manière novatrice en les associant à des parcelles géographiques. Il propose également des cartes de densité qui visualisent la variation des tweets par aires et par périodes temporelles.

Hügel et Roumpani (2014) utilisent également des formes de visualisation 3D dans le projet City Engine Twitter. Ce projet visualise la quantité de tweets en les transformant en formes 3D géolocalisées.

Vu Dung Nguyen, Blesson Vargheseb et Adam Barkerb (2013) saisissent l'occasion du Royal Birth en 2013 pour analyser le sentiment du public du Royaume-Uni. Avec leur visualisation des sentiments sur Twitter, ils arrivent également à mener une analyse sur la relation entre les formes de visualisation et la récolte des données. C'est pourquoi dans leur travail, ils analysent la création de leur corpus en explicitant la méthodologie adoptée. Ils analysent l'étape du parsing (parser les tweets : repérer les mots clés utilisés dans les tweets afin d'analyser les sentiments dans l'opinion publique). Les auteurs proposent deux formes de visualisation : des tile-map et des cartographies produites à l'aide de Google Earth.

Dans un article très critiqué, les géographes Taylor Shelton, Ate Poorthuis et Matthew Zook (2015) étudient la ségrégation spatiale de la population dans la ville de Louisville au Kentucky à travers l'analyse des tweets géolocalisés. La première carte produite par ces chercheurs donnait l'idée d'une ville caractérisée par une ségrégation spatiale très marquée. A la suite des critiques sur leur méthodologie de visualisation des tweets géolocalisés, ces chercheurs ont modifié la modalité de visualisation en proposant une deuxième carte qui donne cette fois-ci des résultats plus nuancés que la première (Marshall 2015).

En soulignant les fausses représentations produites par une utilisation non rigoureuse des tweets géolocalisés, Aarian Marshall (2015) signale également le manque de significativité de l'étude sur la diffusion de l'hashtag #ferguson13 en critiquant une utilisation paresseuse des formes de visualisation de tweets géolocalisés (Rogers 2014).

Nan Cao et al. (2012) concentrent leur travail sur une forme de visualisation « Whisper » capable de décrire en temps réel la diffusion de l'information liée au tremblement de terre qui a frappé la région de Hokkaido en 2012. Leur visualisation permet d'analyser la diffusion des tweets dans son aspect qualitatif (émotions), spatial (géo-groupes), temporel (temps) et quantitatif (intensité). Comme dans le cas précédent, cette visualisation est conçue pour être une aide pour les analystes de très larges flux de données.

### *1.2 Tweets non-géolocalisés*

Cette deuxième catégorie de publications scientifiques présente des formes de visualisation de données moins dynamiques et interactives. Par exemple, Wueest et Müller (2014) présentent la quantité et la qualité des tweets ou des comptes observés selon des thématiques étudiées à travers des formes conventionnelles de visualisation.

Dans cette catégorie, une étude de référence est sûrement le rapport annuel de « twiplomacy » financé par Burson-Marsteller et qui étudie l'utilisation de Twitter par les gouvernements de la planète (Lüfkens 2015).

Dans leur article, Lulic et Kovic (2012) sélectionnent des profils Twitter qui appartiennent à certains groupes d'utilisateurs, notamment des physiciens. Ensuite ils utilisent les logiciels Twiangulate tool and NodeXL pour identifier les relations entre les comptes Twitter et pour calculer les métriques de réseau et visualiser celles-ci avec des histogrammes ou des graphiques de réseau.

Guille et Favre (2014) visualisent les évènements mentionnés dans des tweets. D'abord ils identifient les évènements mentionnés en employant la méthodologie MABED (mention-anomaly- based Event Detection). Ensuite ils visualisent ces évènements mentionnés à travers d'une ligne de temps orientée sur laquelle les évènements sont alignés en ordre

[13] De nombreuses visualisations dans la blogosphère sont réalisées par Carto DB (https://cartodb.com). Le logiciel est une plate-forme de cloud computing qui permet d'intégrer des outils de SIG et de cartographie Web. Ce logiciel permet aux utilisateurs avec des compétences de programmation très diverses de créer des visualisations à partir des big data et de les diffuser sur le web. CartoDB permet également aux utilisateurs de développer des versions personnalisées du logiciel.

chronologique et à travers des graphiques plutôt classiques, par exemple des aires empilées en 3D.

Heli Aramo-Immonen, Jari Jussila et Jukka Huhtamäki (2015) explorent les pratiques d'apprentissage collaboratif chez les participants de conférences en Finlande à travers l'analyse de leurs tweets. Ils visualisent les résultats à travers une image de réseau réalisée grâce au logiciel Gephi et une visualisation matricielle qui détaille les différents comportements.

Jian Zhao, Nan Cao, Zhen Wen, Yale Song, Yu-Ru Lin, et Christopher Collins (2014) présentent un système interactif de visualisation FluxFlow #FluxFlow, conçu pour révéler et analyser des modèles de diffusion anormale dans les médias sociaux. Pour cela, ils visualisent les 100 retweeting/fils de discussion les plus anormales en 2012 dans un laps de temps de 18 heures suivant l'ouragan Sandy. La forme de visualisation est présentée par les auteurs comme un outil opérationnel capable d'aider en temps réel les analystes dans leur travail d'identification des modalités extraordinaires de diffusion de l'information.

Huron, Vuillemot et Fekete (2014) illustrent leur technique de visualisation, appelée sédimentation visuelle. Les auteurs expliquent avoir conçu cette technique pour répondre au besoin de visualiser « des flux des données dynamiques, volumineux et hétérogènes » pour accomplir une analyse finalisée à la prise de décision. Dans leur outil de visualisation, Visual sedimentation, les flux de données sont réduits en de simples unités représentées à l'aide de jetons (élément graphique). Chaque jeton est visualisé en mouvement sous l'influence des forces magnétiques ou gravitationnelles. A la fin de leur trajet, les jetons s'empilent dans des conteneurs dans lesquels ils vont se retrouver agrégés.

Dans cette catégorie, nous retrouvons également des formes de visualisation de données obtenues par Twitter qui illustrent les attitudes relationnelles des utilisateurs.

Par exemple, Palmer (2013) visualise les caractéristiques de l'utilisation de Twitter par les universités australiennes à travers des diagrammes relationnels qui montrent l'intensité des échanges des tweets et la distribution des retweets entre/parmi les pôles universitaires étudiés.

Dermouche et al. (2015) proposent dans le cadre d'une analyse d'opinion de visualiser l'opinion contenue dans un corpus de tweets discriminants par un "nuage de termes" de taille proportionnelle à leur fréquence d'apparition et par une visualisation temporelle des mêmes termes.

## 2. Autres sources

En dressant un état de l'art concernant la visualisation des données, il est crucial d'analyser l'information sur ce sujet, disponible à travers les réseaux sociaux et la blogosphère en général. Nous avons recherché dans Twitter et dans Pinterest les mots-clés suivants : #dataviz #tweets #twitter #datavisualization #bigdata. Grâce à l'analyse des résultats, nous avons identifié des logiciels, des plates-formes technologiques et des blogs qui pourraient nous aider dans notre projet de recherche.

### *2.1 Plates-formes technologiques*

Dans certains cas, les formes de visualisation des tweets sont le produit d'un travail de sélection des profils, de la constitution de bases des données et d'intégration de plusieurs logiciels qui permettent de visualiser les tweets récoltés. Par exemple, Creation Pinpoint Global HCP Profiles Engine est une plate-forme technologique qui utilise des algorithmes intelligents pour identifier les profils (notamment dans Twitter) des professionnels de la

santé. Creation Pinpoint Global HCP Profiles Engine est produit par Pinpoint Creation, une branche commerciale de Creation Healthcare, une société de conseil financée par les entreprises pharmaceutiques. Dans ce cas, la visualisation est obtenue par l'intermédiaire du logiciel Brandwatch. L'image ainsi obtenue permet de visualiser la géolocalisation des profils en y associant la spécialité, la langue, l'ancienneté du profil, les nombres de tweets et les followers (Ghinn 2015).

Le site Trendsmap permet de cartographier les hashtags les plus utilisés dans le monde. Il permet donc à chaque cybernaute de sélectionner gratuitement certains hashtags et de visualiser leur distribution dans le monde entier (Barnes 2014). Les sites http://tweetping.net/# et http://onemilliontweetmap.com proposent de visualiser en temps réel les tweets dans le monde.

Jean-Charles Baudot (2014) propose une visualisation très réussie d'un processus de retweeting réalisée à travers le site "Where Does My Tweet Go?" (ou WDMTG) qui opérationnalise le concept créé par MFG Labs, une agence de mathématiciens renommés. Grâce à ce site, le processus de retweeting peut être visualisé à travers les reprises de followers d'un compte (Guillame 2013).

### *2.2 Le Data journalism*

De nombreuses formes de visualisation de tweets peuvent être retrouvées parmi les articles de presse qui adhèrent au courant du Data journalism. Dans ce sens le site de Datamatch est un exemple réussi de visualisation des données Twitter pour un public large. La visualisation des tweets dans l'article de Datamatch sur le double discours du Front National a créé en France un débat très important. La visualisation était très classique dans sa représentation (un simple histogramme), toutefois la présentation très esthétique de l'histogramme et l'originalité de la démarche d'analyse de tweets ont permis à cette visualisation Twitter d'influencer le débat politique en France de façon inattendue (Ekimetrics 2015). A l'aide de @ekimetrics, 633 comptes Twitter, les plus actifs pour leur soutien au Front National, ont été individués et triés selon la fonction de leur propriétaire au sein du FN (candidats, cadres, élus). Datamach  a donc analysé les mots et les hashtags les plus cités parmi les 45'675 tweets postés entre le 27 février et le 18 mars 2015 en comparant leur occurrence entre les sous-groupes analysés.

### *2.3 Les Data Artists*

Une série de visualisations très pertinentes pour notre projet sont produites par des artistes de données.

Eric Fischer (2014) visualise les tweets géolocalisés non agrégés à l'aide des cartes géographiques interactives. Par exemple, il utilise MapBox pour visualiser la concentration de touristes dans le monde grâce à des données Twitter via Gnip.

Simon Jackson (2014) utilise les tweets pour cartographier les langues utilisées en Australie à l'aide d'une carte interactive.

Stefan Wagner utilise sociotype.me pour créer des monstres digitaux qui présentent avec leurs tentacules l'activité sur Twitter d'un utilisateur.

A ce propos, la visualisation ECOSPHERE (CNN 2011), réalisée par STINK NUMÉRIQUE LONDRES / NEW YORK, développée et conçue par MINI VEGAS Amsterdam / Los Angeles est l'un des projets de visualisation interactive les plus avant-gardistes de ces dernières années. Elle a permis de voir en temps réel la contribution des utilisateurs de Twitter (tweet tagués avec hashtag # COP17) à la discussion mondiale sur le changement climatique lors de la Conférence COP17 de Durban, en Afrique du Sud en 2011 durant trois semaines.

Dans la visualisation, chaque tweet a stimulé la croissance d'une plante virtuelle localisée dans la région de provenance du tweet.

En conclusion, parmi les blogs, une archive intéressante de visualisation du big data dans des contextes urbains est celle de Margarita Fonseca (2012).

*Bibliographie : voir bibliographie principale du projet de recherche*

# Méthodes de capture

### Données de Twitter

Twitter met à disposition du public sous forme de flux un échantillon des tweets émis, selon des critères définis par l'utilisateur (par exemple la géolocalisation). Il s'agit de l'API de Twitter. Cet outil est utilisé par beaucoup de chercheurs pour analyser les données et les informations circulant sur Twitter, mais présente un défaut : en effet, il n'existe pratiquement pas de documentation concernant la quantité et la nature des données obtenues via l'API. Il est donc difficile de savoir si l'échantillon obtenu est représentatif de tout Twitter ou non.

Il existe cependant un flux contenant la totalité des tweets publics, le Firehose. Ce flux n'est pas accessible au grand public : seuls quelques partenaires de Twitter y ont accès, et cet accès est très coûteux tant en termes de prix qu'en termes d'infrastructures nécessaires (serveurs, espace disque, réseau).

La capture des tweets peut donc se faire avec deux outils différents : le streaming API, gratuit mais limité, et le Firehose, exhaustif mais très cher.

Une étude parue en 2013 (Morstatter et al., 2013) compare les données obtenues avec le streaming de l'API et celles obtenues avec le Firehose. Cette étude démontre que le streaming de l'API couvre en réalité bien souvent plus de tweets que les 1% annoncés, et qu'il est possible d'augmenter le nombre de résultats obtenus en affinant les critères de recherche. De plus, il en ressort que les échantillons de tweets géolocalisés sont souvent exhaustifs même dans l'API, grâce au fait que seul un faible pourcentage des tweets sont géolocalisés.

### Text-mining des tweets

Les tweets que nous récupérons grâce au streaming de l'API comprennent différents types de données: d'une part les métadonnées (telles que le nom d'utilisateur, la langue, la géolocalisation, etc.) et d'autre part le contenu brut du message lui-même.

L'analyse et l'exploitation de ce dernier couplées aux données géolocalisées sont très riches et intéressantes pour étudier l'activité d'une région.

Toutefois, la spécificité de Twitter (limitation des tweets à 140 caractères) enjoint les utilisateurs à recourir au langage SMS, à des abréviations, aux émoticônes et autres modifications de l'orthographe et de la syntaxe, rendant difficile l'usage du traditionnel NLP (Natural Language Processing, ou Traitement automatique du langage naturel) pour extraire les données lexicales. Afin de tout de même pouvoir utiliser ces outils sur les tweets, il existe deux possibilités : soit on adapte le texte à l'outil (normalisation), soit on adapte l'outil au texte (adaptation au domaine) (Eisenstein, 2013).

Cotelo et al. (2015) critiquent le fait que ces approches sont souvent peu flexibles et difficiles à adapter à d'autres médias ou d'autres domaines de texte que ceux pour lesquels elles ont été pensées. Ils proposent donc une nouvelle approche de normalisation modulaire, plus facile à mettre en œuvre et destinée à l'analyse lexicale des tweets.

Kondor et al. (2013) combinent la géolocalisation avec l'analyse du langage pour observer les variations géographiques de la langue. Leur méthode permet d'isoler hors du corpus les messages indésirables (spam et publicité, stations météorologiques, …) afin de n'avoir que les tweets de « vrais » utilisateurs. L'analyse des mots les plus utilisés permettent de mettre en évidence la séparation entre ville et campagne : « downtown », « sushi » ou « mall » sont opposés à « truck » par exemple.

L'aménagement urbain est également une thématique pouvant être analysée avec les tweets géolocalisés, comme le montre l'étude de Frias-Martinez (2014) qui propose une méthode pour identifier la segmentation du territoire grâce à Twitter.

Thom et al. (2012) proposent quant à eux une approche pour utiliser les tweets géolocalisés dans l'identification de catastrophes naturelles ou d'événements tels que les émeutes de Londres. Utiliser Twitter dans ces cas-là permet aux gens de se tenir au courant en temps réel, sans devoir attendre les informations des médias traditionnels. Analyser les tweets géolocalisés permettrait de connaître en temps réel le ressenti de la population durant ces événements, et éventuellement d'optimiser l'intervention des secours. L'étude explique la démarche d'extraction des termes depuis les tweets pour les transformer en « term artifacts », puis la génération de clusters spatiotemporels, que l'on représente ensuite sur une carte.

Le logiciel libre R permet de faire du text-mining dans les données de Twitter. Le processus d'extraction des tweets, de nettoyage du texte, d'identification des mots les plus fréquents et des associations est expliqué par Zhao (2015) dans une présentation.

*Bibliographie : voir bibliographie principale du projet de recherche*

# Archivage

L'archivage de Twitter, et plus largement des réseaux sociaux, pose de multiples questions sur différents aspects. Nous avons choisi de nous concentrer sur quatre angles d'attaque :

- les recherches générales qui réfléchissent sur la nécessité d'un tel archivage, et comment les professionnels de l'information peuvent s'inscrire dans les nouveaux défis qu'amènent les réseaux sociaux
- les outils développés et les actions qui ont été menées
- la question de l'utilisateur et de l'utilisation des archives
- l'exploitabilité des corpus collectés

### Recherches générales

L'importance grandissante des réseaux sociaux oblige les professionnels de l'information à se remettre en question et à réfléchir sur les impacts que ces nouveaux usages ont sur la mission de préservation qui leur incombe. Comme le souligne Antony Funnell (2014), il est nécessaire de comprendre à la fois les changements technologiques, qui affectent directement ce qui est ou non préservé, mais aussi les déplacements qui surviennent dans les attitudes sociales par rapport à ces données. La question de la valeur de ces dernières est discutée : créées bien souvent pour être « consommées », (c'est-à-dire pour répondre à un besoin ou à une situation éphémère), les individus ne voient pas un tweet comme un document propre à être conservé.

Pourtant, la préservation du contenu public de Twitter est devenue, pour Thomas Risse et al., (2013), une nécessité culturelle. Une valeur peut en effet être donnée à ce genre de contenu, non dans leur spécificité, mais dans la possibilité de percevoir les individus, groupes et organisation dans leur globalité. Il s'ensuit que les défis qui surviennent concernent la préservation du contexte des documents, afin de rendre une image globale

des interactions. De plus, l'enrichissement des données est un point central pour garantir la qualité et l'utilisation de ces corpus « démesurés ».

Les pratiques d'archivage traditionnelles sont mises à mal par la nature éphémère et en constante évolution de ces objets. Comme le soulignent Ross Harvey et Martha Mahard (2013), les concepts de longévité, choix, qualité, intégrité et accès doivent être remodelés. Lisa P. Nathan et Elizabeth M. Shaffer (2012) affirme qu'une collaboration entre les archivistes et le champ de l'interaction homme-machine est nécessaire pour affronter le défi de la préservation à long terme de ces documents. C'est une attitude proactive qui doit être engagée et de nouvelles compétences, notamment techniques, sont requises (Sonya Sherman, 2014).

**Outils développés et actions menées**

Au vu de la popularité des réseaux sociaux, des outils émergent pour gérer ce nouveau type d'archives. Pour que les internautes puissent récupérer leurs archives personnelles, Twitter a introduit la possibilité de télécharger ses « archives Twitter », contenant ses propres interactions (Mollie Vandor, 2012). A l'échelle des organisations et des gouvernements, des entreprises leur proposent des solutions pour s'occuper de la gestion et de l'archivage de leurs utilisations des médias sociaux (Archivesocial, archive-it, erado,...)

Des applications et des systèmes sont développés pour assurer la préservation des ressources 2.0 et supporter leurs spécificités. Citons par exemple les articles de Daniel Chudnov et al. (2014), avec une application basée sur Python et Django (Social Feed Manager), de Gregory Roland (2014), qui se concentre sur le problème de l'interopérabilité des systèmes d'archivage, et de Anqi Cui et al. (2012).

Pour assurer la préservation du contenu de Twitter, deux sortes d'interventions ont été menées aux USA. D'une part, un archivage « unifié » : depuis 2010, la Library of Congress conserve tous les tweets publics (Laura E. Campbell, Beth Dulabahn, 2010) – elle a reçu en outre le corpus de tweets émis entre 2006 et 2010. D'autre part, un archivage plus « morcelé » : des guides de bonnes pratiques sont fournis, pour responsabiliser les grandes entreprises ou institutions – par rapport à l'utilisation, la gestion et la conservation de leurs utilisations des médias sociaux –, comme celui de la National archives and records administration (2013) ou celui de la State archives of North Carolina (2012).

**L'utilisateur et l'utilisation des archives**

L'annonce de l'archivage du contenu public de Twitter par la LoC a soulevé des controverses parmi le public, comme le montrent les commentaires qui ont suivi cette déclaration (Matt Raymond, 2010). D'un côté, ce contenu est perçu comme trop personnel, et de l'autre l'archivage de certains messages semble être du gaspillage de temps et d'argent au vu de l'insignifiance de certains posts.

Catherine C. Marshall a réalisé plusieurs études (http://www.csdl.tamu.edu/~marshall/pubs-by-year.html) par rapport aux potentiels utilisateurs. Elle analyse plusieurs questionnaires pour préciser les attentes et les craintes du public en ce qui concerne la propriété privée et l'utilisation d'un tel corpus (accès à tous, seulement aux chercheurs, après un certain temps,…). Kaitlin L. Costellon et Jason Priem (2011) se sont aussi intéressés aux utilisateurs mais en proposant une étude qualitative : ils en ressortent des informations sur l'opinion quant à la préservation (garder une partie des tweets, tout ou ne rien conserver) et quant à la responsabilité de cette sauvegarde. Concernant les questions éthiques qui peuvent entrer en jeu lorsque les médias sociaux sont archivés, le blog de Michael Zimmer (http://www.michaelzimmer.org) donne des pistes de réflexions.

### Exploitabilité des corpus

La qualité du corpus archivé est primordiale pour que l'utilisation en soit optimale. En effet, au vu de l'énorme quantité des données, un travail doit être fait pour qu'elles soient exploitables. Ainsi, tout un pan de la littérature sur l'archivage de Twitter, et plus généralement des médias sociaux, se concentre sur les actions à réaliser pour donner de la valeur aux corpus collectés.

Tout d'abord, The Arcomem consortium (http://www.arcomem.eu/) travaille sur les nouveaux défis qu'amènent les réseaux sociaux. Il cherche à réduire le risque de perdre ce web éphémère et à soutenir la création d'archives de valeur. Le projet draine de nombreuses publications, propose des outils pour la sélection, l'évaluation et l'acquisition de contenu, et présente des méthodes pour l'analyse du web social (web crawling and mining). Par exemple, on trouve dans un de leurs livrables (Alejandro Jaimes et al., 2013) certains de leurs résultats pour extraire des informations intéressantes dans un corpus de tweets.

Ensuite, le domaine du traitement automatique du langage naturel s'intéresse aussi à Twitter et peut amener des pistes pour donner du sens à des archives de tweets. L'article de Leon Derczynski et al. (2015) donne un aperçu des recherches dans cette direction et compare les performances de méthodes de « Named Entity Recognition » sur des données Twitter pour proposer des améliorations.

Enfin, essayer d'évaluer la valeur des tweets est une autre approche qui peut être soulignée. Par exemple, Omar Alonso, Catherine C. Marshall et Mark Najork (2013), après avoir présenté d'autres études dont le cheminement est semblable, cherchent à prédire quels seront les tweets intéressants en construisant une collection de tweets « high-quality labels ».

### Bibliographie

#### *Ouvrages et articles*


CHUDNOV, Daniel, KERCHNER, Daniel, SHARMA, Ankushi, WRUBEL, Laura, 2014. Technical challenges in developing software to collect Twitter data. Code4lib [en ligne]. Vol 25. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://journal.code4lib.org/articles/10097.

DERCZYNSKI, Leon, MAYNARD, Diana, RIZZO, Giuseppe, VAN ERP, Marieke, GORRELL, Geneviève, TRONCY, Raphaël, PETRAK, Johann, BONTCHEVA, Kalina, 2015. Analysis of named entity recognition and linking for tweets. Information processing and management [en ligne]. Mars 2015. Vol 51, n° 2, pp. 32-49. [Consulté le 14 juin 2014]. Disponible à l'adresse : http://derczynski.com/sheffield/papers/ner_single.pdf.

FUNNELL, Antony, 2014. Give me a serve of data with that. Archives and manuscripts [en ligne]. 30 juillet 2014. Vol. 42, n° 2, pp. 181-183. [Consulté le 14 juin 2014]. Disponible à l'adresse : http://www.tandfonline.com/doi/abs/10.1080/01576895.2014.911682?journalCode=raam20.

HARVEY, Ross, MAHARD, Martha, 2013. Mapping the preservation landscape for twenty-first century. Preservation, digital technology & culture [en ligne]. Mars 2013. Vol 42, n° 1, pp. 5-16. [Consulté le 14 juin 2014]. Disponible à l'adresse : http://www.degruyter.com/view/j/pdtc.2013.42.issue-1/pdtc-2013-0002/pdtc-2013-0002.xml.

RAYMOND, Matt, 2010. How tweet it is ! : library acquires entire Twitter archive. Library of Congress blog [en ligne]. 14 avril 2010. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive.

RISSE, Thomas, PETERS, Wim, SENELLART, Pierre, MAYNARD, Diana, 2013. Documenting contemporary society by preserving relevant information from Twitter. In : WELLER, Katrin, BRUNS, Axel, BURGESS, Jean, MAHRT, Merja, PUSCHMANN, Cornelius (éds). Twitter and society. New York : Peter Lang, 2013, p. 207-229. Digital Formations, 89. ISBN 978-1-4331-2169-2.

SHERMAN, Sonya, 2014. People telling stories. Archives and manuscripts [en ligne]. 30 juillet 2014. Vol. 42, n° 2, pp. 204-208. [Consulté le 14 juin 2014]. Disponible à l'adresse : http://www.tandfonline.com/doi/abs/10.1080/01576895.2014.911690?journalCode=raam20.

VANDOR, Mollie, 2012. Your Twitter archive. Blog Twitter [en ligne]. 19 décembre 2012. [Consulté le 14 juin 2015]. Disponible à l'adresse : https://blog.twitter.com/2012/your-twitter-archive.

### *Conférences*

ALONSO, Omar, MARSHALL, Catherine C., NAJORK, Marc, 2013. Are some tweets more interesting than others ? #HardQuestion. In : CAPRA, Robert, FREUND, Luanne, SMITH, Catherine, SMUCKER, Mark, WHITE, Ryen. 7Th annual symposium on human-computer interaction and information retrieval, Vancouver, 3-4 octobre [en ligne]. New York : ACM, 2013. [Consulté le 14 juin 2014]. Disponible à l'adresse : http://dl.acm.org/citation.cfm?id=2528396&dl=ACM&coll=DL&CFID=683229867&CFTOKEN=23309603.

CAMPBELL, Laura E., DULABAHN, Beth, 2010. Digital preservation : the Twitter archives and NDIIPP. In : RAUBER, Andreas, KAISER, Max. 7th International conference on preservation of Digital objects, Vienna, 19-24 septembre [en ligne]. Vienna : iPRES. [Consulté le 14 juin 2014]. Disponible à l'adresse : http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/campbell-27.pdf.

COSTELLO, Kaitlin L., PRIEM, Jason, 2011. Archiving scholars' tweets. In : 4th Annual society of American archivists research forum, Chicago, 2011. Chicago : Society of Armerican archivists. [Consulté le 14 juin 2014]. Disponible à l'adresse : http://www2.archivists.org/sites/all/files/KCFinal.pdf.

NATHAN, Lisa P., SHAFFER, Elizabeth M., 2012. Preserving social media : opening a multi-disciplinary dialogue. In UNESCO. The memory of the world in the digital age : digitization and preservation : an international conference on permanent access to digital documentary heritage, Vancouver, 26-28 septembre 2012 [en ligne]. Vancouver : Sheraton Vancouver wall centre, 2012, p. 410-418. [Consulté le 14 juin 2014]. Disponible à l'adresse : http://www.unesco.org/new/en/communication-and-information/events/calendar-of-events/events-websites/the-memory-of-the-world-in-the-digital-age-digitization-and-preservation.

### *Rapports*

JAIMES, Alejandro, BARBIERI, Nicola, CAUTIS, Bogdan, DUPPLAW, David, GIANNOPOULOS, Gioros, STAVRAKAS, Yannis, SIEHNDEL, Patrick, ZENZ, Gideon, 2013. Social web-based archive contextualization V2 [en ligne]. ARCOMEM : 31 juillet 2013. Collaborative Project (ICT-2009-270239). Disponible à l'adresse : http://www.arcomem.eu/downloads/deliverables.

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION, 2013. White paper on best practices for the capture of social media records. NARA bulletin. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.archives.gov/records-mgmt/resources/socialmediacapture.pdf.

NORTH CAROLINA OFFICE OF THE GOVERNOR, 2012. Best practices for state agency social media usage in North Carolina : version 2.0. Digital records policies and guidelines. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.ncdcr.gov/Portals/26/PDF/guidelines/best_practices_socialmedia_stateagency.pdf.

### *Sites web et blogs*

ARCOMEM, 2015. ARCOMEM [en ligne]. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.arcomem.eu.

MARSHALL, Catherine C, 2014. Selected publications. Cathy Marshall [en ligne]. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.csdl.tamu.edu/~marshall/pubs-by-year.html.

ZIMMER, Michael, 2015. MichaelZimmer [en ligne]. [Consulté le 14 juin 2015]. Disponible à l'adresse : http://www.michaelzimmer.org.