UFR de Langues, Littératures et Civilisations Étrangères (LLCE)

Département Sciences du Langage

Master de Sciences du Langage

Parcours Linguistique, Informatique et Technologies du Langage

---

Mémoire de recherche Master 1

Identification et catégorisation du code-switching anglais-français dans les tweets

Andréa BLIVET

Sous la direction de :

Monsieur TANGUY Ludovic, *Maître de conférences, UT2J*

Monsieur MILETIC Filip, *Doctorant, UT2J*

Juin 2022

Remerciements

Je souhaite commencer par remercier mes directeurs de mémoire : Monsieur Ludovic TANGUY et Monsieur Filip MILETIC. Ils ont su m'accorder le temps nécessaire pour me guider, me conseiller, et me permettre d'être autonome.

Je remercie Madame Lydia Mai HO-DAC et Madame Cécile FABRE pour leurs enseignements tout au long de l'année. Elles ont su me transmettre leurs connaissances et nourrir ma réflexion pour me permettre, entre autres, d'enrichir ce mémoire.

Un grand merci à mes camarades de la promotion M1 LITL 2021-2022, l'entraide et la bonne humeur ont été les fils conducteurs de cette année scolaire grâce à eux.

Je remercie profondément mes copines Jessica, Manon, Louisa et Solène pour leur soutien et leurs encouragements concernant mes études.

Je tiens également à adresser toute ma reconnaissance à ma famille, qui me permet de m'épanouir dans mon parcours scolaire et qui croit en moi.

Enfin, je remercie sincèrement Matthias qui est ma source de motivation et mon soutien sans faille depuis des années.

Table des matières

Introduction

Le code-switching est un phénomène langagier qui est issu du contact des langues. Il se traduit par la présence simultanée de deux langues au sein d'un énoncé. La cohabitation entre les deux langues peut se faire de différentes manières avec plusieurs structures possibles. Plusieurs fonctions pragmatiques et discursives peuvent être associées au changement de langue fait par le locuteur. Le code-switching est un mélange de deux langues (ou plus), de ce fait son fonctionnement et ses contraintes ne sont pas intégrés dans les grammaires des langues impliquées, les grammaires étant focalisées sur le fonctionnement d'une langue à la fois.

Le corpus « Twitter - Montréal » (Miletic et al, 2020) rassemble des tweets anglais produits par des locuteurs de Montréal au Canada. Montréal est une ville du Québec qui abrite deux langues officielles : l'anglais et le français, avec prédominance du français. Cette région linguistique est donc soumise à un fort contact entre les deux langues, de ce fait, les tweets qui y sont produits traduisent potentiellement ce contact avec par exemple, la présence de code-switching.

De ce fait, l'objectif est d'identifier le code-switching présent dans les données. Les données utilisées appartiennent au genre textuel des *Communications Médiées par les Réseaux* (CMR). Les CMR regroupent l'ensemble des productions et des interactions en ligne, par le biais de plateformes numériques. Ces communications restent relativement récentes, les codes qui les régissent proviennent des usages et des spécificités technologiques des plateformes dans lesquelles elles s'intègrent. De ce fait, au même titre que le code-switching, le fonctionnement de ce type de communication ne répond pas à des normes établies de manière explicite et ne correspond pas toujours à la grammaire de la langue utilisée dans ce genre d'interaction.

Le travail qui suit portera donc sur un phénomène langagier dans une situation de communication donnée et il réunira deux objets d'étude observables qui ne sont pas intégrés dans les grammaires des langues. Le premier est un phénomène langagier qui résulte du contact des langues : le code-switching. Le second est un moyen de communication récent dans l'histoire du langage qui a ses propres spécificités : les tweets. De ce fait, par ce travail, l'objectif sera de proposer un traitement automatique du code-switching dans les tweets.

La problématique qui se pose est la suivante : quels sont les traitements automatiques possibles qui permettent l'identification du code-switching dans des données courtes et non-standard ?

Les objectifs de ce travail sont donc de proposer une méthode d'identification de langue suffisamment fine pour pouvoir identifier des segments d'une langue dans des contextes majoritairement d'une autre, mais également de proposer une catégorisation automatique du code-switching identifié afin d'extraire des phénomènes langagiers spécifiques du code-switching dans les tweets.

Les deux principales difficultés au bon déroulement de cette étude proviennent des données. En effet, le traitement automatique du langage est majoritairement développé et entrainé sur des données monolingues. Cependant, les données utilisées ici sont bilingues et les traitements automatiques d'identifications et de catégorisations adaptés pour une langue ne doivent pas altérer ou éclipser les parties dans une autre langue. La tâche d'identification de langue doit alors, dans un premier temps, être adaptée pour ces données et être la plus fiable possible. La deuxième difficulté découle de cette première, la tâche d'identification de langue doit également être efficace sur des données courtes et d'un genre textuel particulier, car la situation de communication (*Twitter*) contraint la longueur des énoncés qui répondent à des codes précis. L'objectif final qui motive ce travail est de pouvoir trouver la méthode et les outils les plus adaptés pour l'étude du code-switching dans les CMR. En particulier pour le repérage et l'extraction du code-switching dans les textes courts.

La première grande partie de ce travail présentera les apports théoriques nécessaires pour aborder le sujet. Dans un premier point, les définitions et les typologies du code-switching déjà établies seront répertoriées. Ensuite, je m'attarderai sur les communications que l'on peut retrouver sur Internet, à savoir, comment elles peuvent être définies et regroupées malgré leur diversité, ainsi qu'une étude de cas particulière pour le réseau social *Twitter*. Une fois nos deux principaux objets d'étude présentés, la troisième partie présentera les différentes difficultés en traitement automatique des langues liées aux données de l'étude.

Ensuite, une seconde grande partie sera consacrée à la mise en application des connaissances acquises pour l'identification et la catégorisation automatiques du code-switching. Dans un premier temps, les données utilisées seront présentées. Ensuite, la méthodologie sera détaillée selon les trois principales étapes qui la composent : l'identification de langue, l'étiquetage des données et l'identification et la catégorisation du code-switching. Enfin, la dernière partie portera sur l'évaluation de la méthodologie et sur l'analyse des résultats obtenus.

Partie I : Apports théoriques

1. Le code-switching

1.1. Définitions

Le code-switching est un phénomène langagier qui résulte du contact des langues. Le contact des langues correspond à la présence simultanée d'au moins deux langues dans une zone géographique, dans un groupe socio-linguistique ou chez un locuteur (Moreau, 1997). Le contact implique une influence mutuelle et une interaction entre ces langues qui entraînent des phénomènes langagiers spécifiques à la situation. Le code-switching est l'un de ces phénomènes. Il renvoie à un énoncé qui est composé d'au moins deux langues différentes.

La définition du code-switching a évolué à travers le temps et elle est sujette à des modifications en fonction des auteurs. Le point qui n'est pas discuté concerne le résultat linguistique : deux langues différentes dans un même énoncé. Cependant, les explications, les causes et les délimitations du code-switching entraînent différentes définitions selon le point de vue abordé. Dans la littérature, il y a donc des propositions de définition du code-switching, de la plus minimaliste à la plus maximaliste, avec des prises de positions plus ou moins différentes. Dans cette partie, je vais présenter différentes définitions du code-switching proposées par différents auteurs.

Pour Pedraza (1978) (cité par Poplack, 1980), le code-switching résulte d'un manque de contrôle du locuteur. C'est-à-dire que le locuteur n'est pas considéré comme suffisamment expérimenté dans les deux langues pour pouvoir les manier correctement et indépendamment. Cette définition peut être mise en lien avec le point de vue de Lance en 1975, qui considère que le code-switching est un phénomène aléatoire, qui ne peut être prédit et qui ne répond à aucune règle linguistique. Par ces deux auteurs, le code-switching est abordé comme étant un phénomène incontrôlable et indésirable.

Poplack propose en 1980 une définition développée du code-switching. Tout d'abord, selon elle, le code-switching est un phénomène qui intervient au niveau de l'énoncé, de la phrase ou du constituant. De plus, elle nuance la définition en précisant que pour que l'alternance des deux langues relève du code-switching, les éléments d'une des deux langues ne doivent pas être altérés par l'autre langue. C'est-à-dire que chaque élément doit respecter la phonologie, la morphologie et la syntaxe de la langue dont il est originaire. De plus, dans sa définition, Poplack (1980) contredit le point de vue de Lance (1975) en excluant le caractère hasardeux du code-switching. En effet, selon elle, quand on observe ses applications, le code-switching est régi par des règles et le hasard n'intervient pas dans ce type de production. Même si elle indique que ces règles doivent être encore précisées, développées et approfondies, elle propose tout de même deux contraintes qui lui semblent être prises en compte dans le code-switching et qui sont également présentées dans Sankoff et Poplack (1981) :

- **la contrainte du morphème libre :** cette contrainte, développée dans Sankoff et Poplack (1981), indique le fait qu'un changement de langue n'est possible entre un lexème et un morphème lié, uniquement si l'élément lexical est intégré dans la phonologie de la langue de l'élément morphologique. Le segment de l'exemple (1) n'est pas autorisé, sauf si le lexème *eat* est prononcé selon la phonologie de la langue du morphème (ici, l'espagnol).
  (1) *eat***iendo** (exemple issu de Boztepe, 2003)
- **la contrainte de l'équivalence :** elle renvoie à la grammaire et à la syntaxe des langues en jeu dans les productions avec code-switching. Sankoff et Poplack (1981) constatent que chaque segment respecte la

> grammaire de la langue dont il est issu. Autrement dit, les segments en L1 sont grammaticalement corrects d'après la grammaire de la L1, et les segments en L2 sont grammaticalement corrects selon la grammaire de la L2. De plus, le code-switching au niveau de la phrase est possible uniquement au niveau des frontières syntaxiques communes aux deux langues. Les productions sont donc contraintes par la grammaire de chaque langue en limitant les possibilités d'apparition du code-switching. Par exemple, un changement de langue de l'anglais vers l'espagnol s'avère impossible entre un nom et son adjectif, car en espagnol l'adjectif suit le nom alors qu'en anglais, il le précède. Un changement de langue impliquerait alors un non-respect de la syntaxe d'au moins une des deux langues.

Ces deux contraintes sont considérées par Poplack (1980) comme étant le résultat de fortes habiletés langagières et d'une bonne maîtrise du bilinguisme. Cette prise de position de l'auteure s'oppose à celle de Pedraza (1978) qui considère, au contraire, que le code-switching résulte d'un manque de contrôle et d'habileté du locuteur. De plus, Poplack (1980) conclut sur le fait que le code-switching n'est pas la somme de la grammaire de la L1 et de la L2, mais c'est une troisième grammaire, issue du chevauchement des deux autres, qui régit les productions de code-switching.

En 2008, Solorio et Liu proposent une approche où le code-switching renvoie à une zone sur le spectre du bilinguisme. Selon ces deux auteurs, le bilinguisme peut être représenté par un spectre, avec de chaque côté, une forme considérée comme pure et standard de chaque langue. Le terme « pure » renvoie ici à une langue ne subissant aucune influence de la part de l'autre langue. Entre les deux bouts de ce spectre, se trouve un continuum entre les deux langues, il n'y a pas de frontière stricte. C'est en se déplaçant vers le centre de ce spectre que des phénomènes de code-switching vont émerger.

Bullock et Toribio (2012) intègrent la notion d'effort et de paramètre à la définition du code-switching. Dans un premier temps, ils considèrent qu'un quelconque effort cognitif ne doit pas être à l'origine du code-switching. Cela fait lien avec l'aspect incontrôlable du code-switching proposé par Pedraza (1978), sans pour autant être justifié par un manque de maîtrise des langues par Bullock et Toribio (2012). De plus, ils précisent leur point de vue en expliquant que le code-switching résulte d'une alternance de deux langues sans modification des paramètres du locuteur. Cette notion de paramètre fait écho avec celle de Chomsky (1971), qui développe l'idée que toutes les langues ont les mêmes principes sous-jacents, mais que seuls les paramètres de ces principes sont modifiés en fonction la langue ciblée. Cela voudrait donc dire que, pour Bullock et Toribio (2012), le code-switching n'entraîne pas d'effort et de modification des paramètres cognitifs par le locuteur. De plus, pour ces deux auteurs, le code-switching répond à des modèles qui sont, certes, très variés et peu uniformes, mais qui sont en aucun cas aléatoires.

La définition de Poplack (1980) semble être désignée comme celle de référence, car elle est citée et reprise par les autres auteurs. Ces auteurs ne contredisent pas le point de vue de Poplack (1980), ils approfondissent et agrémentent cette définition. Solorio et Liu (2008) proposent une visualisation spectrale du code-switching, tandis que Bullock et Toribio (2012) intègrent la notion d'effort cognitif.

Hormis par Lance (1975), le code-switching est décrit comme répondant à des règles (Lipski, 1978 ; Poplack, 1980 ; Sankoff et Poplack, 1981 ; Bullock et Toribio, 2012). En ce qui concerne les habilités langagières requises, Pedraza (1978) considère que le code-switching est signe de mauvaise maîtrise des langues. En revanche, pour Poplack (1980) et Solorio et Liu (2008), le code-switching est possible qu'avec une bonne maîtrise des deux langues par le locuteur.

Pour résumer, le code-switching décrit un phénomène en production qui réunit deux langues différentes au sein d'un même énoncé, d'une même phrase ou d'une même proposition. Le code-switching répond à des règles et à des modèles. Il est considéré comme du code-switching uniquement lorsque les segments d'une langue répondent aux traits phonologiques, morphologiques et syntaxiques de cette même langue.

1.2. Les structures du code-switching

Indépendamment du contenu, le code-switching peut être catégorisé selon un point de vue structurel. Muysken (2000) propose de dissocier le code-switching en fonction de l'organisation des langues au sein de l'énoncé. Poplack (1980), elle, catégorise le code-switching selon l'apparition du code-switching par rapport aux frontières syntaxiques de l'énoncé. Dans ces deux types de répartition, le code-switching est catégorisé en fonction des points de changement de langue (frontières entre les deux langues).

Les exemples utilisés à partir de maintenant sont tous issus du corpus « Twitter-Canada » de Miletic et al, 2020 (ce corpus est présenté et décrit dans la partie *4. Présentation des données*). Les segments en français sont présentés en gras, et ceux en anglais en italique. Chaque exemple est suivi du numéro unique d'identification du tweet entre parenthèse (TweetId). Ce numéro permet d'accéder directement au tweet sur Twitter depuis ce lien : https://twitter.com/x/status/{tweet.id}.

1.2.1. Les types d'alternance du code-switching

Muysken (2000) propose une classification du code-switching qui repose sur l'organisation de l'alternance entre les langues. Dans cet aspect, la catégorisation repose sur la répartition des langues dans l'énoncé. Il relève deux structures possibles : la juxtaposition et l'insertion. Ces deux catégories se distinguent selon le nombre de points de changement de langue dans l'énoncé.

Juxtaposition
Le code-switching par juxtaposition est une forme d'alternance qui consiste à commencer l'énoncé dans une langue pour le terminer par une autre. Dans cette situation, le locuteur part de la L1 vers la L2 sans retourner vers la L1 par la suite. Ainsi, l'énoncé est composé d'un seul point de changement de langue. L'exemple (2) illustre ce cas de figure, où le locuteur part de l'anglais pour aller vers le français et conclure son énoncé ainsi.

> (2) *The greatest winery in #Navarra @BODEGASOCHOA @SAQCellier #OCHOA . Thank you Beatriz and Andriana for your hospitality !* **Le meilleur vignoble de Navarre . Merci pour votre hospitalité !** (TweetId : 1054102208931749888)

Insertion
Le code-switching par insertion est une forme d'alternance dans laquelle un segment dans une langue est inséré entre deux segments dans l'autre langue. Dans cette configuration, il y a donc au minimum deux points de changement de langue et le passage d'une langue vers une autre se fait dans les deux sens. L'exemple (3) est un tweet composé de code-switching par insertion. Dans cet énoncé, le locuteur passe de l'anglais vers le français, pour ensuite revenir à l'anglais et enfin retourner vers le français. Ainsi, ce tweet est composé de 3 points de changement de langue.

> (3) *It is sounding to me like you want the VIP experience* , **bel homme** . *That might cost you a pretty penny* . 😈 **Que veux-tu donner pour passer la nuit avec moi** ? (TweetId : 1131930679430795264)

1.2.2. Les types de structures syntaxiques du code-switching

La catégorisation du code-switching en fonction de sa répartition par rapport à la structure syntaxique a été initiée par Poplack (1980). Cette classification repose sur la concordance des points de changement de langue par rapport aux frontières syntaxiques de l'énoncé. Le code-switching peut être inter-phrastique ou intra-phrastique. A ces deux catégories s'ajoute le code-switching intra-mot proposé par Barman et al (2014).

Inter-phrastique

Le code-switching inter-phrastique est considéré ainsi quand des propositions monolingues et distinctes sont assemblées pour former un énoncé bilingue (Poplack, 1980). Pour que les deux propositions forment un code-switching inter-phrastique, il faut qu'elles soient indépendantes l'une de l'autre. L'exemple (4) illustre une production de code-switching de type inter-phrastique. La première proposition est une phrase monolingue en français, tandis que la seconde est également une phrase monolingue, mais en anglais. Les deux sont séparées par un signe de ponctuation (faible ou fort) qui traduit leur autonomie syntaxique, elles sont indépendantes, mais elles constituent un énoncé cohérent par leur union.

(4) **C'est ce que je me demandais** ! 😁*Wild claims based on rumours should be considered false and fake until proven true* ! (TweetId : 851204828097335299)

Intra-phrastique

Le code-switching intra-phrastique correspond à une cohabitation d'éléments de différentes langues dans une même proposition. Contrairement au code-switching inter-phrastique, les segments sont liés et dépendants l'un de l'autre, ils ne peuvent pas fonctionner en autonomie. Poplack (1980) et Lipski (1978) considèrent que ce type de code-switching est possible uniquement entre des langues qui partagent les mêmes frontières syntaxiques et que la stratégie est donc d'intégrer des fragments monolingues sur une structure syntaxique commune. Solorio et Liu (2008) partagent ce point de vue et ajoutent que les langues impliquées doivent être symétriques et que les items lexicaux peuvent être remplacés d'une langue à une autre sans impact sur la structure sous-jacente. L'exemple (5) est une manifestation du code-switching intra-phrastique. Dans ce tweet, le passage de l'anglais vers le français se fait au sein de la même proposition.

(5) *Cassie she have something* **je ne sais quoi** !!!(TweetId : 1087821140058427392)

Intra-mot ou emprunt ?

Le code-switching intra-mot apparaît au niveau du mot. L'item lexical d'une langue est intégré à la morphologie d'une autre langue (Barman et al, 2014). C'est-à-dire qu'un mot est composé d'une base lexicale en L1 et agrémenté de morphèmes en L2. L'exemple (6) illustre ce phénomène, l'item « brunch » de l'anglais est combiné avec le suffixe *-er* renvoyant à l'infinitif d'un verbe en français.

(6) **Allons** *brunch***er** ! (TweetId : 937371905123868672)

La définition du code-switching intra-mot entre en conflit avec la définition de Poplack (1980), qui, rappelons-le, considère qu'il n'y a pas de code-switching si les éléments d'une langue sont altérés par la phonologie ou la morphologie de l'autre langue. Dans Bullock et Toribio (2012), ce phénomène est décrit comme issu du contact des langues sans être du code-switching. En effet, ils considèrent ce type de phénomène comme étant un emprunt ou une assimilation, car il y a une projection des caractéristiques morphologiques et phonologiques d'une L1 sur un item lexical d'une L2. Pour ces auteurs, il y a une démarche d'intégration de l'item lexical dans le lexique d'une autre langue, mais pas d'alternance entre la L2 et la L1.

1.2.3. Classification retenue

Il est à retenir des informations présentées précédemment que les différents types de code-switching proposés sont tous basés sur le même critère : les points de changement de langue.

L'organisation de l'alternance se réfère au nombre de points de changement présents dans l'énoncé, tandis que la structure du code-switching se définie par la localisation syntaxique des points de changement.

Ces deux types de classifications ne sont pas incompatibles entre elles, au contraire, elles se complètent et leur union permet de faire émerger une nouvelle classification plus affinée. Ainsi, le code-switching inter-

phrastique peut s'organiser par insertion ou par juxtaposition, tout comme le code-switching intra-phrastique. De plus, l'organisation et la structure du code-switching ne sont pas dépendantes, de ce fait, il est possible d'étudier l'un sans prendre en compte l'autre.

En ce qui concerne le code-switching intra-mot, sa caractérisation de code-switching est débattue. Pour ce travail, j'ai décidé de ne pas le considérer comme un phénomène du code-switching car il relève davantage d'un phénomène d'emprunt ou d'assimilation.

Ainsi, la catégorisation des productions du code-switching d'un point de vue structurel retenue pour la suite du travail est présentée dans le tableau 1. La deuxième colonne du tableau illustre les différentes catégories avec l'anglais et le français et où le signe « | » représente une frontière syntaxique.

Inter-phrastique	
Juxtaposition	**EN** | *FR* ou *FR* | **EN**
Insertion	**EN** | *FR* | **EN** ou *FR* | **EN** | *FR*
Intra-phrastique	
Juxtaposition	**EN** *FR* ou *FR* **EN**
Insertion	**EN** *FR* **EN** ou *FR* **EN** *FR*

Tableau 1 : Classification des différents types de code-switching.

1.3. Les motivations et les fonctions du code-switching

Précédemment, le phénomène de code-switching a été décrit selon la forme qu'il prend au niveau syntaxique. Cependant, il peut aussi être défini selon les raisons de son utilisation et les fonctions qui lui sont associées. D'un point de vue fonctionnel, le code-switching est une stratégie de conversation. Il peut transmettre des fonctions distinctes (Begum et al, 2016).

1.3.1. Les motivations du code-switching

Le changement de langue dans un même discours peut-être expliqué selon différents niveaux. Dans un premier temps, la situation de communication peut être l'une des raisons de la présence de code-switching.

En 1982, Grosjean recense les différents facteurs qui interviennent dans le choix de la langue qui est fait par le locuteur. Parmi ces facteurs, certains sont directement reliés aux participants de la conversation (âge, langue privilégiée, statut socio-économique…). La situation de communication et le contenu du discours sont également des facteurs de changement de langue identifiés par Grosjean (1982).

Boztepe (2003) distingue le code-switching de situation du code-switching métaphorique. Le code-switching de situation est motivé par la situation de communication en général et par les participants, tandis que le code-switching métaphorique est motivé par le sujet du discours ou les changements de sujets. Ainsi, dans un premier temps, une distinction est faite entre l'environnement et le contenu de la communication.

Brasart (2011) présente différentes raisons pouvant mener à l'acte de code-switching. Selon lui, le code-switching a pour objectif de traduire des compétences linguistiques et culturelles afin d'ancrer le discours dans une réalité et de fiabiliser le propos. Il peut, par ailleurs, être utilisé pour désigner une réalité qui

n'existe pas forcément dans une autre langue, ou qui n'a pas d'équivalent. Enfin, le code-switching permet de rapporter un évènement ou des souvenirs associés à une langue pour rendre plus authentique les propos.

Plusieurs motivations sont associées au code-switching. Le code-switching est utilisé pour changer de sujet dans un discours, pour exprimer une identité linguistique et/ou culturelle, pour ajuster son discours et s'adapter à l'interlocuteur ou encore pour changer le niveau de formalité de l'interaction (Begum et al, 2016).

1.3.2. Les fonctions discursives associées au code-switching

D'un point de vue fonctionnel, le code-switching est considéré comme une stratégie conversationnelle, il permet de transmettre différentes fonctions distinctes dans un message (Begum et al, 2016).

Boztepe (2003) recense différentes fonctions discursives pouvant être transmises par le code-switching :

- **citation :** permet d'introduire du discours rapporté dans la langue de production originale.
- **spécification d'adresse :** permet d'indiquer un changement de locuteur, à qui s'adresse le message.
- **répétition :** permet de clarifier ou de reformuler le message.
- **narration / opinion :** le changement de langue indique l'implication du locuteur dans l'énoncé.
- **interjection :** permet de compléter le message avec les interjections d'une autre langue.

Begum et al (2016) ajoute également les fonctions du code-switching suivantes :

- **renforcement :** le changement de langue permet de renforcer le message exprimé.
- **sarcasme :** l'opinion sarcastique est exprimée dans une langue différente de celle du sujet.
- **traduction :** le message est traduit dans une autre langue afin d'élargir les destinataires du message.
- **cause / effet :** le changement de langue marque le passage de la cause à la conséquence.

Begum et al (2016) proposent de classer le code-switching en fonction de la relation sémantique qui unie des segments exprimés dans des langues différentes. Leur classification permet de regrouper les différentes fonctions discursives identifiées précédemment : identique, similaire, différente, contradictoire et pas de lien.

- **relation identique :** dans ce type de relation, les segments sont sémantiquement identiques. Une information produite dans une langue est ensuite traduite dans une autre.
- **relation similaire :** les segments ne correspondent pas à une traduction fidèle l'un de l'autre (contrairement à la relation identique), mais d'un point de vue sémantique, ils transmettent la même information. Les fonctions discursives qui correspondent à ce type de relation sont le renforcement ou la répétition.
- **relation différente :** le contenu des segments est différent, mais il est tout de même relié sur le plan du discours, sans s'opposer. Par exemple, un segment peut correspondre à de la narration et l'autre à l'opinion ou alors l'un exprime une cause quand l'autre exprime la conséquence.
- **relation contradictoire :** dans ce type de relation, le contenu sémantique des segments est contraire. Le sarcasme est une fonction discursive qui s'intègre dans ce type de relation.
- **pas de relation :** dans ce cas de figure, les segments n'ont aucune relation entre eux. Cela se traduit par un changement de sujet qui correspond au changement de langue.

La classification de Begum et al (2016) permet de regrouper différentes fonctions discursives du code-switching dans des catégories plus vastes et plus générales.

1.3.3. Phénomènes linguistiques spécifiques au code-switching

Le croisement des informations concernant la fonction et la structure du code-switching permet d'identifier certains phénomènes spécifiques au code-switching. Le tag-switching et le texte bilingue sont deux phénomènes identifiables en prenant en compte à la fois la structure du code-switching et la fonction associée. L'accès au sens et au contenu du message est nécessaire pour leur identification.

Tag-switching

Le tag-switching renvoie à l'intégration d'une expression courante d'une langue dans un énoncé entièrement composé dans une autre langue (Poplack, 1980). Une expression courante est une séquence lexicale généralement figée et considérée comme une même unité d'un point de vue syntaxique ou sémantique, avec également une forte fréquence d'occurrence. Le tag-switching est habituellement un mot isolé ou une courte séquence de mots produits dans l'autre langue. De plus, il renvoie principalement à la fonction d'interjection proposée par Boztepe (2003). L'exemple (7) est principalement produit en anglais, avec seulement l'expression « du coup » en français, qui peut être considérée comme courante, car fréquemment utilisée chez les locuteurs francophones. Bullock et Toribio (2012) justifient l'utilisation de tag-switching par des compétences langagières limitées dans la langue.

(7) **du coup** *pierrot is on the road to the grand slam !!* (TweetId : 863033789492203521)

Texte bilingue

Le texte bilingue est un type de code-switching proposée par Lynn et Scannell (2019). Ce type de code-switching se traduit par la présence d'une même information dans deux segments de langues différentes. Autrement dit, le segment n°1 correspond à la traduction du segment n°2, et inversement. Dans ce cas, il n'y a pas de lien discursif entre les deux segments, ils se contentent de partager le même énoncé pour apporter exactement la même information. D'un point de vue structurel, le texte bilingue est principalement construit par juxtaposition et dans une configuration inter-phrastique. L'exemple (8) est une illustration du texte bilingue.

(8) **Merci** ! *Thanks xx* ;) (TweetId : 1110646750488989699)

Cette partie permet de mieux comprendre et de mieux définir les différents types de code-switching qui peuvent être rencontrés. Les productions avec code-switching se retrouvent plus aisément dans le cadre d'interactions spontanées et informelles. De ce fait, ce phénomène est souvent associé à la modalité orale du langage. Cependant, certains types de communication écrite réunissent des caractéristiques à la fois spontanées et informelles (Begum et al, 2016).

2. Les communications digitales médiées par les réseaux

Le développement du Web et des outils technologiques ont vu émerger des nouveaux types de communications. Ces communications s'inscrivent dans des nouveaux genres qui possèdent des particularités communicatives inédites (Poudat et al, 2020). Ces communications sont décrites comme étant des Communications Médiées par les Réseaux (CMR, ou CMC - *Computer Mediated Communication* en anglais). Dans cette partie, une première section est consacrée à la présentation approfondie des CMR, ensuite, une deuxième section présente les différents types de CMR qui peuvent exister. Enfin, la troisième est dernière section détaille les particularités propres à Twitter, réseau social depuis lequel les données sont issues.

2.1. Présentation

Une vingtaine d'années après la naissance du World Wide Web en 1993, la linguistique de corpus s'est intéressée à la langue d'internet (Chanier, 2017). Ces communications écrites en ligne ont permis d'obtenir des interactions entre des locuteurs « ordinaires ». Chanier (2017) insiste sur l'aspect « ordinaire » des locuteurs, qu'il oppose aux autres locuteurs principalement issus du monde académique (journalistes, écrivains, politiques), dont les productions étaient davantage étudiées en linguistique de corpus. L'avantage du Web est qu'il permet d'accéder à une grande quantité de données écrites (Poudat et al, 2020).

Les communications médiées par les réseaux réunissent l'ensemble des interactions langagières produites à l'aide de matériel informatique (ordinateur, tablette, smartphone) et qui transitent par des réseaux (Internet, Intranet, télécommunication) (Chanier et al, 2014). Ces productions n'ont pas vocation à être imprimées, elles transitent uniquement par un canal numérique.

Les communications digitales sont diverses et variées, Crystal (2001) indique que ces communications sont trop hétérogènes pour être rassemblées dans un même genre. En effet, elles utilisent les mêmes outils, mais de manières très différentes.

Crystal (2001) propose une première catégorisation de ces nouvelles données. Cette catégorisation repose sur la nature même du message. C'est-à-dire que les genres sont décrits par le type de message ou par le dispositif de communication utilisé pour la transmission de ce message (mails, chatgroups). Les propriétés de ces communications ne sont pas pris en compte dans la classification de Crystal (2001). Dans sa classification, Crystal considère exclusivement les communications transmissent via un réseau Internet, les télécommunications ne sont pas intégrées. La classification est donc la suivante :

- Les e-mails
- Les chatgroups : les interactions produites dans les chatgroups sont divisées en deux catégories selon la synchronicité de l'espace d'interaction : synchrone et asynchrone.
- Les mondes virtuels
- Le World Wide Web

Cette classification propose un aperçu des CMR, mais elle n'est pas représentative. Depuis 2001, il y a eu des évolutions technologiques qui ont entraîné l'évolution des CMR. Désormais, les nouvelles classifications des CMR s'appuient davantage sur les propriétés techniques et linguistiques des genres de CMR.

Les nouveaux critères de classification proposés par Chanier et al (2014), Wigham et Poudat (2020) et Poudat et al (2020) s'appuient davantage sur l'espace d'interaction, le mode de la communication et la synchronisation.

L'espace d'interaction est un concept abstrait qui vise à situer la situation et la configuration des interactions en ligne (Chanier et al, 2014 et Poudat et al, 2020). L'espace d'interaction est situé dans le temps (avec un

début et une fin à l'aide de l'horodatage des données), il implique un set de participants dans un lieu défini en ligne. Le set de participants renvoie aux locuteurs qui participent à la création des interactions dans l'espace. Ce set peut se composer de membres individuels ou bien de groupes de membres. Le lieu en ligne est l'environnement qui est utilisé par les participants avec ses propriétés. L'espace d'interaction peut être rejoint ou quitter à n'importe quel moment par les participants et il peut être public ou à accès restreint (Chanier et al, 2014).

Les propriétés du lieu en ligne sont le mode de communication, la synchronisation et le type d'interaction (Wigham et Poudat, 2020).

Le mode de communication correspond aux ressources sémiotiques à disposition pour l'élaboration du message. Les communications peuvent être mono ou multimodales. Le mode verbal renvoie au textuel, le mode visuel aux images et le mode non-verbal. Dans les CMR, le mode non-verbal se traduit par le recours aux émoticônes et aux émojis qui miment le gestuel par exemple (Wigham et Poudat, 2020).

La synchronisation désigne l'obligation, ou non, de la présence simultanée des participants pour pouvoir participer à l'interaction (Wigham et Poudat, 2020).

Enfin, les communications médiées par la réseaux se démarquent selon le niveau de spontanéité ou de formalité attribué au message (Crystal, 2001).

2.2. Caractéristiques des CMR

Le modèle de Biber (1988) défini six critères linguistiques permettant de dissocier le langage écrit du langage oral (production impliquée vs. informationnelle, situation explicite vs. situation dépendante, information résumée vs. non résumée, narratif vs. non narratif, expression de la persuasion, élaboration instantanée du message). Crystal (2001) a appliqué ce modèle pour situer les CMR selon les critères linguistiques. Les résultats indiquent que les CMR n'obtiennent des scores extrêmes dans aucun des critères permettant de les associer davantage à l'écrit ou à l'oral. De plus, Crystal (2001) établit également une liste de caractéristiques discriminant l'oral de l'écrit. Cette étude montre que les CMR contiennent des caractéristiques de l'oral, comme des formes contractées ou l'utilisation de termes déictiques. Crystal (2001) les considère donc comme représentant un troisième mode de communication, hybride de l'oral et de l'écrit.

Dorlejin et Nortier (2012) caractérisent aussi les CMR comme un troisième mode de communication. Les données sont certes écrites, mais elles ne sont pas le reflet du langage écrit au sens traditionnel et conventionnel. En effet, les auteurs ont constaté que dans les CMR, il n'était pas rare de retrouver des collocations ou des caractéristiques habituellement associées au langage oral. Dorlejin et Nortier (2012) qualifient les CMR comme spontanées et informelles. Lynn et Scannell (2019) expliquent que le caractère spontané et informel des CMR provient du fait que ces productions ne sont pas soumises à un processus de correction ou d'édition. Cependant, les auteurs émettent une réserve, pour eux, les CMR ne reflètent pas forcément l'usage quotidien réel des locuteurs.

Les CMR se composent, entre autres, de données déviantes. En effet, il est possible de retrouver beaucoup de variations orthographiques : des abréviations, des transcriptions phonétiques des termes, des répétitions inattendues de lettres et des choix arbitraires entre des homophones (Charnier et al, 2014, Farzindar et Roche, 2013). De plus, Farzindar et Roche (2013) ajoutent à cela des problèmes de respect de la grammaire avec davantage de phrases fragmentées, des emplois irréguliers et des omissions de ponctuation et de majuscule. Les CMR contiennent également des formes inédites et des néologismes qui peuvent être issues des variations orthographiques (Poudat et al, 2020 et Farzindar et Roche, 2013).

De plus, les messages dans les CMR peuvent être accompagnés d'émoticônes ou d'émojis. Les émoticônes sont des pictogrammes constitués de caractères du code ASCII, tandis que les émojis sont une version stylisée et dessinée des émoticônes. Ils permettent d'ajouter une dimension visuelle au message, en

traduisant des gestuelles ou des expressions faciales qui auraient été présentes dans la version orale du message (Magué et al, 2020).

2.3. Le réseau social Twitter

Les données utilisées pour ce travail proviennent du réseau social Twitter, créé en 2006. Les réseaux sociaux font partis des CMR et ils ont des caractéristiques qui leur sont spécifiques. De plus, le fonctionnement de Twitter possède également ses spécificités par rapport aux autres réseaux sociaux.

2.3.1. Les réseaux sociaux

Farzindar et Roche (2013) définissent les réseaux sociaux comme un « *recours à des outils électroniques et à l'Internet dans le but de partager et d'échanger efficacement de l'information et des expériences* » (p. 13).

La particularité des réseaux sociaux est qu'ils sont ancrés dans le temps, avec une volonté de diffusion presque immédiate. Les réseaux sociaux ont pour objectif de favoriser les interactions entre les utilisateurs, leur accès n'est soumis à aucune restriction géographique ou temporelle (Lacaze, 2021). Ils ont également la volonté d'inciter les utilisateurs à partager leur quotidien en temps réel. Des communautés se créent et des relations se développent entre les interlocuteurs. L'aspect immédiat de ces communications privilégie le caractère spontané des interactions (Paveau, 2013).

Les réseaux sociaux sont de plus en plus exploités et démocratisés, ce qui diversifie énormément le contenu produit. En 2019, il est recensé 3.5 milliards d'utilisateurs des réseaux sociaux (Yao et Ling, 2020). Ils sont utilisés à des fins personnelles, commerciales ou encore professionnelles. Dans certaines situations, cela peut moduler le niveau de spontanéité et de formalité.

Les réseaux sociaux permettent d'étudier les comportements sociaux à travers la langue. Leur avantage est qu'ils permettent d'accéder à une grande quantité de données qui sont continuellement renouvelées (Farzindar et Roche, 2013). De plus, un autre avantage des réseaux sociaux pour les études est que, au moment de la production, le locuteur n'a pas conscience que ses messages vont être analysés à des fins linguistiques (Çetinoğlu et al, 2016).

Les données issues des réseaux sociaux se définissent par leur volume, leur variété et leur vélocité (Farzindar et Roche, 2013). Le volume correspond à la grande quantité de données produites continuellement. La variété de ces données provient du fait qu'une multitude de formats et de modes de communication sont mis à disposition des utilisateurs. Enfin, la vélocité fait référence à la présence de données bruitées, c'est-à-dire difficilement utilisables pour une étude, avec par exemple la présence de publicités, de messages produits par des robots ou alors de messages vides de mots qui sont des contenus peu pertinents qui imposent un filtrage en amont pour une utilisation de ces données (Farzindar et Roche, 2013 et Chanier, 2017).

Par le biais des réseaux sociaux, Internet est devenu un lieu social à part entière, où les interactions y sont réelles (Paveau, 2013). Farzindar et Roche (2013) considèrent que le langage sur les réseaux sociaux répond davantage à la réalité plutôt qu'à des normes linguistiques. Ils sont utilisés de manière informelle avec des messages reprenant le ton de la conversation (Farzindar et Roche, 2013).

D'après Nguyen et Dogruoz (2013), il y a plus de locuteurs multilingues que monolingues dans le monde. Dans leur comportement langagier au quotidien, les locuteurs multilingues alternent entre les langues. De ce fait, ce comportement se reflète également dans leur façon de communiquer dans un environnement en ligne.

L'anglais est la langue majoritairement utilisée sur Internet, cependant, les réseaux sociaux cherchent à réunir l'ensemble des individus, quels que soient les pays ou les langues (Das et Gambäck, 2013). De ce fait, ils sont tous traduits dans beaucoup de langues pour que chacun puisse se les approprier. Ainsi, il peut y avoir

une stratégie en fonction de la langue choisie pour un message pour des locuteurs non-natifs de l'anglais : utiliser sa langue maternelle pour s'inscrire dans une communauté, un groupe linguistique, ou, au contraire, utiliser l'anglais pour s'ouvrir à un plus grand ensemble d'utilisateurs.

Das et Gambäck (2013) indiquent que le code-switching sur les réseaux sociaux est davantage présent dans les régions multilingues, ou alors dans les régions qui ont une forte proximité géographique avec d'autres langues, ce qui est le cas du Québec. De plus, le code-switching est davantage représenté sur les réseaux sociaux que dans les écrits plus formels, car le contenu est majoritairement émotionnel (Das et Gambäck, 2013).

2.3.2. Twitter

En plus de celles attenantes aux réseaux sociaux, Twitter possède ses propres caractéristiques relatives au fonctionnement et aux utilisateurs du réseau.

Selon les caractéristiques présentées dans la partie précédente sur les CMR, Twitter propose une communication asynchrone. La communication y est multimodale, car il est possible d'utiliser du texte, des vidéos, des images, des liens ou encore des émojis (verbal, visuel et non-verbal). Enfin, l'espace d'interaction sur Twitter est public (Wigham et Poudat, 2020). Le réseau social s'inscrit dans une volonté d'*open access*, même s'il reste possible pour les utilisateurs de restreindre l'accès à leurs tweets, la plus grande majorité des tweets sont accessibles par tous (Dridi et Lapalme, 2013 et Paveau, 2013). De plus, les relations entre les utilisateurs sont asymétriques, il est possible de suivre le contenu d'un utilisateur sans accord préalable, et sans que celui-ci soit obligé de suivre le contenu de l'utilisateur qui le suit. Il n'y a pas de réciprocité entre utilisateurs, de ce fait, il est difficile de savoir à qui l'utilisateur s'adresse et qui accédera au contenu du tweet (Paveau, 2013). Ainsi, un tweet peut s'adresser à un individu, un groupe d'individus ou alors à un public plus large. De ce fait, les tweets peuvent relever de la communication publique ou privée, avec des interlocuteurs connus ou non (Magué et al, 2020).

En ce qui concerne le type d'interaction, Twitter utilise la « *délinéarisation technodiscursive* ». Technodiscursif renvoie au « *discours numériques produits sur les plateformes technologiques des réseaux sociaux* » (Lacaze, 2021). Le fait que cela soit délinéarisé signifie que les tweets peuvent être considérés comme des entrées autonomes tout en étant lié dans un fil de discours défini (Lacaze, 2021). Ainsi, le tweet peut s'inscrire dans un monologue ou dans un dialogue (Magué et al, 2020).

Les tweets se composent de quatre types de formes langagières. Dans un premier temps, le tweet peut être un texte linéaire, sans aucune caractéristique technologique comme il serait possible de trouver dans un environnement hors ligne. Ensuite, il peut contenir des hashtags et des mentions (# et @) qui permettent de situer le tweet dans un fil de discussion ou de l'adresser à un utilisateur. Les liens URL peuvent également être intégrés dans le message, tout comme les émoticônes et émojis (Paveau, 2013).

Twitter n'échappe pas au phénomène de production massive de données, d'après Fausto (2015), plus de 500 millions de tweets sont publiés par jour dans plus de 35 langues différentes. Comme pour les autres réseaux, l'anglais est la langue majoritairement représentée.

Une étude sociologique menée par Blank (2016) révèle le profil des utilisateurs de Twitter. Majoritairement, ils sont jeunes avec un niveau social et un niveau d'étude supérieurs aux autres utilisateurs d'Internet en général. De plus, les « élites » sont davantage représentées sur Twitter que dans la vie réelle. Par exemple, au Royaume-Uni, la majorité des utilisateurs ont entre 18 et 34 ans, sont diplômés et vivent en ville. Ainsi, sur Twitter les utilisateurs ne sont pas représentatifs de l'ensemble d'une population.

Les messages postés sur Twitter sont limités en caractères. Au début du réseau, ils ne pouvaient pas dépasser les 140 caractères. La contrainte des caractères était à d'origine technologique, les tweets se basaient sur les SMS limités à 160, mais dont 20 caractères étaient imputés pour le pseudo (Paveau, 2013). Même s'il n'y a plus de contraintes technologiques à l'heure actuelle, Twitter conserve sa limitation de caractères, qui est

maintenant de 280 caractères. Désormais, le microblogage s'inscrit dans un style de communication spécifique à Twitter permettant de se démarquer des autres réseaux. Ce mode de communication a été adopté par les utilisateurs et il leur permet de s'exprimer de façon brève et concise (Chanier et al, 2014).

La volonté de Twitter est d'inciter les utilisateurs à partager ce qu'ils font ou ce qu'ils pensent au moment où ils le font ou le pense. Ainsi, Twitter favorise la spontanéité de ses utilisateurs. La conception de Twitter vise une forme de blog où les utilisateurs peuvent partager leurs pensées et leurs réactions individuelles de manière immédiate et personnelle. Certaines caractéristiques des forums sont présentes avec les fils de discussion. Twitter n'intègre pas la fonctionnalité « commentaire » aux tweets, de ce fait, tous les messages ont le même statut de tweets. Twitter possède une communauté discursive, c'est-à-dire que ses utilisateurs partagent des usages avec des codes et des pratiques spécifiques au réseau. Par exemple, le vendredi est le jour où les utilisateurs peuvent partager leurs recommandations autour de la lecture (Paveau, 2013). Les tweets contiennent principalement des avis, des informations ou des témoignages (Dridi et Lapalme, 2013). La communication sur Twitter est moins normée que les écrits plus conventionnels (Magué et al, 2020).

Aussi bien sur les réseaux sociaux en général, que spécifiquement sur Twitter, les interactions sont considérées majoritairement comme étant informelles et spontanées. Ces deux caractéristiques sont également des facteurs favorisant l'apparition de code-switching. Ainsi, les données issues de Twitter sont disposées à l'étude du code-switching.

3. Traitement automatique des données

Les outils de traitement automatique de données langagières sont généralement développés à partir de données dites « traditionnelles » et monolingues (Farzindar et Roche, 2013), en opposition aux données utilisées dans ce travail.

Dans un premier temps, les données issues des réseaux sociaux représentent un challenge dans le monde du TAL, par leur proportion de données considérées comme du bruit (Lui et Baldwin, 2014). De plus, le style conversationnel qui leur est souvent attribué complique les analyses possibles sur les données, tout comme la présence accentuée de code-switching (Millour, 2020).

Segmentation

La segmentation est une tâche qui permet de délimiter les données en unités. La délimitation peut se faire à différents niveaux, comme en phrase, en syntagmes ou en tokens. Les omissions de ponctuation et de majuscule dans les données issues des CMR fragilisent la segmentation des unités (Farzindar et Roche, 2013).

Identification de langue

La tâche d'identification de langue permet de déterminer les ressources et les analyses qui sont les plus adaptées pour obtenir le traitement le plus performant sur les données. L'identification de langue se compose principalement de trois approches : une première basée sur la reconnaissance de mots-clés représentatifs, une deuxième qui analyse les n-grammes et enfin une approche basée sur l'utilisation de réseaux de neurones (Kevers, 2021). Dans ces approches, l'unité d'analyse est le texte ou le document, ainsi, Kevers (2021) souligne le fait que certains points de progrès restent à faire en ce qui concerne les documents multilingues qui nécessitent une analyse plus fine.

Les données issues de Twitter présentent plusieurs difficultés pour la tâche d'identification de langue. Tout d'abord, la taille contrainte des tweets favorise le recours à des acronymes et des abréviations qui interfèrent dans l'identification. De plus, la taille du tweet limite les données sur lesquelles s'appuie l'attribution de langue, et donc impacte la fiabilité des résultats. Enfin, le style informel, ainsi que les émoticônes et les émojis posent de réelles difficultés (Das et Gambäck, 2013).

La présence de code-switching davantage fréquente dans des données issues des CMR contraint également l'identification de la langue. Lorsque deux langues sont utilisées, cela brouille le signal pour l'identification qui peut se retrouver avec plusieurs propositions et ainsi sélectionner une langue, au détriment d'une autre pourtant bien présente, ou proposer plusieurs langues sans identifier les segments concernés par chacune (Çetinoğlu et al, 2016).

Nguyen et Dogruoz (2013) constatent que l'identification de langue des tweets produit de meilleurs résultats quand l'unité d'analyse est le token plutôt que le document et quand l'analyse se base sur une correspondance avec des dictionnaires. Le taux de réussite est de 89,5% pour une analyse au niveau du document, contre 97,6% au niveau du token. Ces performances concernent aussi bien les données monolingues que bilingues. Cependant, la fiabilité de ces performances reste soumise à la taille des textes, ceux qui sont courts contiennent moins de données et le niveau de sureté des résultats peut en être impacté. De plus, les dictionnaires utilisés ne doivent pas être des dictionnaires traditionnels, car les tweets contiennent des formes déviantes ou nouvelles (Nguyen et Dogruoz, 2013)

Les échecs des outils d'identification de langue sur des données issues des réseaux sociaux proviennent principalement du style d'écriture qui ne correspond pas aux conventions habituelles du langage écrit (Das et Gambäck, 2013).

Etiquetage morpho-syntaxique

L'étiquetage morpho-syntaxique consiste à ajouter des informations morphologiques, grammaticales et syntaxiques sur des tokens. Cette tâche peut être compromise quand elle est appliquée sur des données qui

contiennent des orthographes déviantes, des néologismes et des abréviations (Farzindar et Roche, 2013). Dans la plupart des outils, l'étiquetage se fait à partir de modèles d'une seule langue. L'intégration d'éléments dans une autre langue peut donc associer de mauvaises étiquettes ou alors à l'étiquette « X ».

Les principales difficultés de traitement automatique des données issues des réseaux sociaux sont les particularités orthographiques et le manque de normalisation. De plus, le code-switching est abordé comme étant une problématique issue de ce genre de données. Il faut retenir que dans un premier temps, la normalisation des données est l'une des solutions pour favoriser les résultats. En ce qui concerne l'identification de langue, elle doit être réalisée au niveau du token pour accroître ses performances. De ce fait, une segmentation des tweets en token est nécessaire, mais elle doit prendre en compte également le fait que les données ne respectent pas forcément les grammaires établies.

Partie II : Mise en application

4. Présentation des données

Les données utilisées pour ce travail sont issues du corpus « Twitter-Canada » constitué par Miletic et al (2020).

4.1. Le corpus Twitter-Canada

4.1.1. Constitution

Le corpus « *Twitter-Canada* » (Miletic et al, 2020) a été constitué dans le cadre de l'étude des changements sémantiques de l'anglais québécois induits par le contact du français.

L'un des critères de constitution était d'obtenir des données de l'anglais représentatives des usages réels. De plus, ces données devaient être issues de zones géographiques distinctes, dont une avec un contact fort entre l'anglais et le français, et les autres sans contact du français.

Le recours à des communications médiées par les réseaux favorise l'aspect informel et spontané des données. Cela permet de traduire au mieux les habitudes langagières réelles des locuteurs.

Un autre critère de constitution était d'avoir des productions langagières associées à des zones géographiques définies. Le réseau social *Twitter* propose un filtrage par localisation de l'auteur au moment de la publication du tweet. Cette fonctionnalité permet de récupérer des productions originaires d'une zone géographique sélectionnée.

Une partie des tweets anglais localisés à Montréal, Toronto et Vancouver a pu être récupérée auprès de Twitter. Pour garantir la fiabilité de l'origine linguistique des locuteurs, seuls les tweets dont l'auteur spécifiait explicitement qu'il était de l'une des trois villes sélectionnées étaient retenus. De plus, les contenus indésirables, les retweets et les doublons n'ont pas été inclus au corpus.

Une fois le nettoyage et le tri des données achevés, les données ont pu être rassemblées pour constituer un corpus regroupant des tweets canadiens. Ce corpus a ensuite été divisé en trois sous-corpus, chacun correspondant à l'une des trois villes retenues.

4.1.2. Contenu

Le corpus est composé de 35.164.923 tweets différents pour un total de 153.656 locuteurs. Cela fait une moyenne d'environ 229 tweets par locuteur. Tous ces tweets réunissent 628.937.204 tokens, avec un vocabulaire de 15.790.343 lemmes pour un total de 17.104.784 formes. La quantité de lemmes disproportionnée par rapport au nombre de formes s'explique par le fait que les formes non reconnues (les noms propres, les URL, les hashtags ou encore les formes avec des orthographes déviantes) sont considérées également comme leur lemme.
Le corpus a ensuite été divisé en trois sous-corpus :

- Le sous-corpus « Montréal » regroupe l'ensemble des tweets produits par des locuteurs anglophones montréalais. Il est composé de 193.228.236 tokens, ce qui représente 30,72 % des tokens du corpus global.
- Le sous-corpus « Toronto » regroupe l'ensemble des tweets produits par des locuteurs anglophones de Toronto. Il est composé de 222.508.451 tokens, soit l'équivalent de 35,37 % du nombre total des tokens du corpus.
- Le sous-corpus « Vancouver » est composé de tweets anglais produits par des locuteurs de Vancouver. Il compte 213.200.517 tokens, ce qui représente 33,89 % des tokens du corpus total.

Les trois sous-corpus correspondent chacun à environ un tiers du corpus entier. De ce fait, le corpus est équilibré en ce qui concerne la représentativité régionale des tweets.

4.2. L'échantillon du corpus

Ce mémoire porte sur le code-switching dans les tweets du Québec, de ce fait, les sous-corpus de tweets « Vancouver » et « Toronto » ne sont pas pertinents ici. Le sous-corpus « Montréal » est donc la seule source de données pour ce travail. Cependant, ce sous-corpus reste très volumineux (193.228.236 tokens) et le temps de traitement de ces données est accru. Afin de faciliter les traitements et les analyses futures, le travail porte sur un échantillon de ce sous-corpus.

L'échantillon a été généré de façon aléatoire. Dans un premier temps, 5000 utilisateurs du sous-corpus « *Montréal* » ont été sélectionnés aléatoirement, ensuite, 10 tweets par locuteur ont été retenus, toujours de manière aléatoire. L'échantillon regroupe alors 50.000 tweets. Ces données sont au format .txt, et sont organisées sous forme de colonnes à l'aide de tabulations (figure 1). La première colonne est le numéro d'identification du locuteur, la troisième est le numéro d'identification du tweet (TweetId) et la quatrième contient le tweet en question. La deuxième colonne contient un score allant de 0 à 1 qui traduit les proportions de tweets identifiés en anglais produit par utilisateur. Le score 1.0 indique que 100% des tweets publiés par l'utilisateur sont en anglais. A l'inverse, le score 0.0 indique qu'aucun tweet de l'utilisateur est publié en anglais. Cet échantillon est le corpus de travail de ce mémoire, dorénavant *le corpus* réfère à cet échantillon.

1037531	1.0	1046560939460493313	I wonder who these idiots are . Not acceptable at all .
1624587414	0.7	1102600364040294401	mais tu sais quoi , 90% of the people who told me that were non blacks
283279205	0.0	1095398349489008642	La ministre des Anciens Combattants , Jody Wilson Raybould quitte le cabinet https://t.co/QWlpxUmaLH #polcan

Figure 1 : Extrait de l'échantillon du sous-corpus

Le corpus de développement

Un extrait de l'échantillon du corpus a été constitué pour faciliter l'entrainement et la vérification des programmes et des traitements qui seront par la suite appliqués sur l'échantillon de 50.000 tweets. Cet extrait est composé de 50 tweets sélectionnés manuellement dans l'objectif de proposer des tweets monolingues anglais et des tweets avec à la fois de l'anglais et du français.

Le repérage de tweets uniquement en anglais a été facilité par la constitution du corpus ciblant majoritairement des tweets en anglais. Pour ce qui est du repérage des tweets bilingues, ils ont été repérés en observant les correspondances avec les tokens spécifiques du français « *je* » et « *tu* ». Ces deux tokens ont été choisis, car ils n'apparaissent pas dans les formes de l'anglais, mais également, car ce sont des formes fréquentes du français. Ces deux tokens ont permis d'identifier des tweets bilingues, mais aussi des tweets composés exclusivement de tokens en français qui ont été aussi intégrés au corpus de développement.

5. Identification de la langue

La détection du code-switching se fait d'abord en identifiant les langues utilisées dans un énoncé. La tâche d'identification de langue est donc une étape indispensable pour l'étude du code-switching. De plus, la qualité de cette dernière impacte directement la précision et le rappel des énoncés qui seront identifiés comme répondant à des phénomènes du code-switching.

5.1. Méthodes d'identification

Le traitement automatique des langues propose des outils permettant une identification automatique des langues. Dans un premier temps, ces outils doivent être testés afin de s'assurer qu'ils répondent au mieux aux besoins de détection du code-switching.

5.1.1. Les outils à disposition

Les outils d'identification de langue permettent de produire en sortie la langue utilisée dans un énoncé.

L'outil *langid* a été développé pour pouvoir proposer une identification de langue efficace sur les messages courts, et il a été également évalué sur des données issues de Twitter (Lui & Baldwin, 2012). *Langid* accompagne les prédictions de langue d'un score qui évalue la probabilité de sa prédiction.

Langid a été testé sur 15 tweets du corpus de développement. Les tweets monolingues anglais ont été correctement identifiés avec des scores de probabilité de 1,0 ou très proche. Certains tweets monolingues français ont été identifiés comme étant du français, mais avec des scores de probabilité légèrement inférieurs à 1,0. D'autres, n'ont pas été correctement identifiés, comme le tweet de l'exemple (9) prédit par *langid* comme étant en anglais (même résultat en enlevant l'URL). Pourtant, ce tweet contient uniquement des tokens en français. Ce défaut d'identification sur des données monolingues peut potentiellement s'expliquer par la brièveté de l'énoncé. L'impact de la longueur de l'énoncé sur l'identification de la langue est un réel problème pour une étude axée sur des données courtes.

(9) **Relance ton porc** https://t.co/FYIjgp1xbE (TweetId : 1100420443759173632)

Quant aux tweets bilingues français-anglais, la plupart d'entre eux sont identifiés comme étant en anglais. De plus, les scores de probabilité sont suffisamment élevés pour ne pas être différentiables des scores obtenus avec des tweets monolingues anglais. Par exemple, pour le tweet en exemple (10), *langid* détecte de l'anglais avec un score de probabilité de 1,0. *Langid* estime donc qu'il y a 100% de chance que ce tweet soit rédigé en anglais. Cependant, sur 32 mots qui composent le tweet, 8 sont en français.

(10) *Wanting to show a friend one of your Critical Role art , and scrolling your entire page , and just being mesmerized by everything you did* . **Je suis tellement impressionné , t'as pas idée !** (TweetId : 1084834248366981120)

La détection de langue proposée par *langid* opère sur l'ensemble de l'énoncé, de plus, elle paraît être influencée par sa taille. Il semble également que certains segments en français ne sont pas pris en compte quand ils sont dans un contexte majoritairement anglophone. Ainsi, *langid* ne peut être utilisé efficacement pour la détection de code-switching sur des données courtes. Les outils d'identification de langue fonctionnent majoritairement comme *langid*, de ce fait, l'identification de la langue doit être abordée différemment.

5.1.2. Méthodologie retenue

Le test des outils dans le point précédent a permis de mettre en avant des problèmes d'identification du code-switching : une identification au niveau de l'énoncé et une influence de la taille de ce dernier. Par conséquent, pour accroître les performances de détection du code-switching, l'identification de la langue doit être réalisée à un niveau inférieur, soit celui du token.

Pour une identification de langue au niveau du token, la première étape est la segmentation du tweet en tokens. La segmentation est une étape clé dont la qualité aura des répercussions sur la suite. De ce fait, elle doit pouvoir s'adapter aussi bien au français qu'à l'anglais, avant même de connaître la langue. L'objectif de la segmentation est d'obtenir des tokens sur lesquels vont se faire l'attribution de langue. Cette attribution va être réalisée en projetant des lexiques de l'anglais et du français sur les tokens.

Le deuxième point important de l'identification de langue envisagée, est donc la sélection des lexiques. Les lexiques sont des listes de formes pour une langue donnée. Ils ne sont pas des listes exhaustives, de plus, en fonction des données sur lesquelles ils ont été construits, ils peuvent également contenir des formes non standards pour la langue ciblée.

Une fois la segmentation et les lexiques réunis, il reste à vérifier dans quels lexiques apparaissent les tokens du corpus. Un token présent dans le lexique de l'anglais et absent du lexique du français sera alors considéré comme anglais, et inversement. Un token appartenant au lexique de l'anglais et du français sera considéré comme bilingue. Et enfin, un token apparaissant ni dans le lexique de l'anglais, ni dans le lexique du français sera considéré comme inconnu pour ces deux langues.

Une fois la vérification d'appartenance des tokens faite, la dernière étape est d'étiqueter chaque token selon ses correspondances. Ainsi, pour chaque tweet, il est possible de renvoyer la langue ou les langues qui le composent et non pas la langue majoritaire ou la plus probable pour l'ensemble du tweet.

5.2. Segmentation

L'identification de la langue par projection de lexiques sur les tokens implique une segmentation en tokens préalable. Que ce soit pour le français ou pour l'anglais, les tokens sont délimités par les espaces ou par les signes de ponctuation. Cependant, Universal Dependencies[1] alerte sur certains cas particuliers de délimitation qui différent entre l'anglais et le français.

Dans un premier temps, le traitement des apostrophes n'est pas le même entre les deux langues. Quand l'apostrophe à un rôle de séparateur, en anglais, elle appartient au token postposé à cette dernière, tandis qu'en français, elle appartient au token antéposé. Sinon, il n'y a pas de segmentation au sein du token, comme pour *aujourd'hui* en français ou la contraction de la négation *n't* en anglais.

- EN : Harry's → | Harry | 's |
- FR : l'école → | l' | école |

Ensuite, pour l'anglais, les tirets sont des séparateurs, sauf dans des cas où ils permettent de lier des affixes communs à une base (exemple : *co-ordinated* ou *e-mail*), c'est-à-dire des affixes qui n'ont pas de statut indépendant. Pour le français, les tirets sont des séparateurs lorsqu'ils impliquent un pronom clitique. Dans cas, le tiret est rattaché au clitique. Sinon, pour les mots composés, il n'y a pas de segmentation interne.

- sous-marin → | sous-marin |
- vient-il → | vient | -il |

[1] https://universaldependencies.org

En anglais, la négation se rattache à l'auxiliaire, dans cette configuration, deux tokens sont liés sans espace ou ponctuation. Ainsi, ils doivent être délimités. En revanche, quand l'apostrophe est une marque de contraction d'un token (comme dans *n't*), elle n'est pas un marqueur de délimitation.

- cannot → | can | not |
- didn't → | did | n't |

En ce qui concerne la segmentation des formes spécifiques à Twitter ou aux CMR comme les hashtags, les liens URL ou les mentions, ils ne sont pas considérés comme des formes participant à l'identification de la langue utilisée par le locuteur et ne sont donc pas segmentés.

Afin d'accroître les chances de correspondances entre les tokens des tweets et les formes des lexiques, la segmentation doit être la plus adaptée possible, à la fois pour l'anglais et pour le français.

5.2.1. Segmentation automatique

Une segmentation manuelle adaptée aux deux langues est compliquée par l'accumulation de règles et de différences entre le français et l'anglais. Elle impliquerait une liste d'exceptions très longues et difficilement représentative de l'ensemble des spécificités. Des outils proposent une segmentation des données. C'est pourquoi la stratégie est d'effectuer une première segmentation automatique, avec par la suite des ajustements réalisés manuellement.

Le corpus est normalement constitué davantage d'anglais. Donc la segmentation automatique est réalisée avec un modèle de l'anglais, ainsi, les ajustements nécessaires concernent uniquement les spécificités du français.

La segmentation a été réalisée à l'aide du segmenteur de SpaCy v3.2[2] basé sur le modèle de langue SM de l'anglais (soit le modèle de taille petite). SpaCy est une bibliothèque regroupant de logiciels Python pour le traitement automatique du langage. Le modèle de l'anglais est un modèle statistique qui a été entrainé sur des données écrites issues du web.

L'algorithme en charge de la segmentation commence dans un premier temps par segmenter l'ensemble du texte sur les espaces. Ensuite, il analyse chaque segment du texte de gauche à droite en prenant en compte le contexte des deux côtés pour établir une possible correspondance entre le segment et l'une des règles présente dans le modèle de langue utilisé.

Evaluation de la segmentation automatique

Il est très compliqué de trouver quelconques informations concernant l'évaluation du modèle de segmentation de SpaCy. De ce fait, j'ai décidé de comparer la segmentation proposée par SpaCy avec une segmentation effectuée manuellement. Cette première évaluation est réalisée avant les ajustements dans le but d'évaluer uniquement les performances de l'outil sur l'ensemble des données.

L'évaluation a été réalisée sur le corpus de développement, qui, en plus de tweets monolingues et bilingues, contient des spécificités de Twitter comme les mentions, les hashtags ou les liens URL.

Le corpus de développement a été segmenté manuellement selon les règles de segmentation présentées dans le guide en annexe 1. En parallèle, ce même extrait a été segmenté par SpaCy avec le modèle de l'anglais.

Comparaison et évaluation des annotations

La segmentation automatique (SpaCy) renvoie 838 tokens pour l'ensemble des 50 tweets. La segmentation manuelle renvoie, elle aussi, 838 tokens, mais avec des différences.

[2] *https://spacy.io*

Au total, il y a 12 points de désaccord entre les deux segmentations. Deux d'entre eux concernent la segmentation d'une suite de signes de ponctuation. La ponctuation ne déterminant pas la langue impliquée, ces désaccords n'affecteront pas l'objectif d'identification. Cinq points de désaccord portent sur la segmentation autour d'apostrophes pour le français. La segmentation manuelle est conforme aux attentes, contrairement à l'automatique. Ensuite, trois points de désaccords concernent les hashtags. Manuellement, le symbole # est rattaché à la chaîne de caractères qui suit, tandis que SpaCy segmente entre les deux. Le segment « *and/or* » a également été segmenté différemment. Automatiquement, ce segment est considéré comme un seul token, tandis que manuellement il est considéré comme trois tokens (« *and* | / | *or* »). A l'inverse, le segment « pop-ups » a été considéré comme un seul token par l'annotateur et comme trois tokens pour SpaCy (« pop | - | ups »). Cette différence de segmentation s'explique par le fait que l'adverbe *ups* a été considéré comme étant un suffixe dépendant, non soumis à une segmentation dans le guide.

Le taux d'accord entre la segmentation manuelle et la segmentation automatique a été obtenu en calculant leur nombre de points de segmentation identiques par rapport au nombre total de tokens obtenus. Le résultat est de 98,57%.

Le segmenteur de SpaCy est satisfaisant, car il s'adapte aux multiples spécificités de segmentation de l'anglais qui sont plus compliquées à répertorier dans un programme. Toutefois, certains ajustements doivent être faits sur la segmentation de SpaCy pour favoriser les correspondances au moment de la projection des lexiques, puisque pour le moment, les règles de segmentation de l'anglais sont également appliquées à des données en français.

5.2.2. Ajustements de la segmentation

Les ajustements de la segmentation ne sont pas nécessaires sur les séquences de ponctuations et les émoticônes. Le but n'étant pas de proposer une segmentation idéale, mais une segmentation qui permet une délimitation efficace des tokens qui vont permettent l'identification de langue. Par conséquent, les ajustements concernent le traitement des apostrophes et des tirets en français, ainsi que le traitement global des hashtags.

Les apostrophes

Les ajustements concernant les apostrophes doivent être réalisés uniquement sur des configurations du français pour ne pas détériorer le traitement des apostrophes de l'anglais.

Pour cela, il faut identifier dans un premier temps les structures du français dont la segmentation implique un rattachement de l'apostrophe au token postposé. L'identification de ces structures a été réalisée d'abord en répertoriant les tokens se terminant par l'apostrophe dans le lexique du français (GLAFF présenté dans la partie *5.3 Les lexiques*). La présence de l'apostrophe en position finale dans le lexique indique qu'elle doit être suivie d'un point de segmentation. De ce fait, il faut relever les occurrences avec une apostrophe finale. La séquence la plus fréquente correspond à un caractère suivi de l'apostrophe (*d', ç', l', s', c', m', t', n'*). La contraction de *que* est également représentée (*qu'*) ainsi que les formes construites depuis *que* (*lorsqu*', *jusqu*'…). Deux schémas peuvent donc être relevés : une lettre unique avec contrainte de début de mot, suivie d'une apostrophe, et le segment *qu* suivi de l'apostrophe sans contrainte de début de mot (\b\w ' et .*qu ').

Dans un deuxième temps, les deux schémas identifiés ont été observés sur Frantext[3]. Le corpus Frantext est principalement composé d'écrits littéraires, mais il contient par ailleurs des écrits plus divers, dont certains issus du Web. Il contient 264.150.870 tokens, ce qui fait de lui un corpus volumineux.

L'objectif de la recherche sur Frantext est de recenser l'ensemble des usages qui répondent aux schémas identifiés, mais qui ne seraient pas présents dans le lexique. La requête interroge l'ensemble des occurrences d'une lettre unique suivie de l'apostrophe.

[3] *https://www.frantext.fr*

Pour la séquence lettre unique suivie de l'apostrophe (\w'), 44 formes correspondent. Une frontière nette entre les fréquences des résultats s'observe. Aucun résultat n'obtient une fréquence entre 13 et 381 par million de mots. Les résultats avec une fréquence inférieure à 13/million de mots sont par conséquent exclus, car après observation des contextes, la plupart proviennent d'écrits en ancien ou moyen français ou alors de contractions stylistiques de la littérature comme « r' » utilisé pour remplacer le préfixe « re-» principalement dans les discours rapportés. Hormis la séquence « ç' » provenant de la contraction de « ça », tous les types identifiés grâce au lexique obtiennent une fréquence supérieure à 380/million de mots. Un seul type, avec aussi une fréquence supérieure à 380/million de mots, n'a pas été identifié dans le lexique, la contraction du pronom « je » en « j' » (3.149/million de mots). Etant donné sa fréquence, cette forme peut être ajoutée au lexique qui sera par la suite utilisé pour l'identification de langue.

Pour résumer, à l'aide du croisement des données de Frantext et de GLAFF, les segments en français qui nécessitent une segmentation après l'apostrophe sont les suivants : *d', ç', l', s', c', m', n', t', j' et qu'*(soit 10 schémas).

Pour ne pas détériorer la segmentation déjà établie pour l'anglais, avant d'appliquer les ajustements prévus pour le français, il faut s'assurer que ces séquences ne soient pas présentes dans les usages de l'anglais avec une segmentation attendue différente.

La vérification des usages des apostrophes de l'anglais a été faite à l'aide de l'échantillon du corpus COCA[4] (*Corpus of Contemporary American English*). Le corpus COCA est un corpus de référence pour l'anglais américain. Il se veut représentatif de la langue écrite et se développe également pour des transcriptions de l'oral. L'échantillon utilisé se compose d'extraits du corpus de 1990 à 2019. Il contient des extraits des huit genres proposés dans le corpus intégral (blog, texte académique, fiction, magazine populaire, oral, journaux, web et sous-titres de films et de séries). Au total, l'échantillon du corpus COCA contient 485.000 textes et 12.153.815 tokens.

L'étude et l'interrogation de cet échantillon a été faite sur le logiciel TXM (Heiden et al, 2010). Ainsi, les huit fichiers textes ont été regroupés en un seul corpus (le genre n'étant pas un critère pour ce travail).

Les dix schémas retenus lors de l'étude des usages des apostrophes en français ont servi de requête sur TXM pour l'échantillon du corpus COCA. Les schémas relevés pour le français sont des cas extrêmement rare en anglais (ils représentent moins de 0,0008% des cas dans le corpus COCA). De plus, les contextes montrent que ces occurrences sont généralement déjà suivies d'espaces ou alors issues de fautes certainement involontaires. La liste des occurrences des schémas du français dans le corpus de l'anglais est disponible dans l'annexe 2.

Les observations des schémas pour l'anglais confirment le fait qu'une segmentation après l'apostrophe de ces schémas est possible pour satisfaire la segmentation du français sans détériorer celle de l'anglais.

Les tirets

L'analyse des segments composés d'un tiret suit la même méthodologie que celle utilisée juste avant pour les apostrophes.

Dans le lexique, il y a 18 formes avec un tiret à l'initial : *-elle, -elles, -il, -ils, -je, -la, -le, -les, -leur, -lui, -moi, -m', -nous, -on, -t, -toi, -tu, -vous* et *-y*. Principalement des clitiques et des pronoms.

Ensuite, sur Frantext, il y a 481 résultats pour la recherche avec l'expression régulière suivante : « -\w+ ». Un seuil fréquence par millions de mots a été établi. Contrairement aux apostrophes, il n'y a pas de frontière nette entre ces fréquences, qui sont davantage continues. Arbitrairement, il a été décidé d'utiliser le même seuil que pour les apostrophes, soit >13/million de mots. Les fréquences inférieures à ce seuil correspondent en majorité à des séquences composées de numéraux ou de segments inconnus au français contemporain.

[4] *https://www.english-corpora.org/coca/*

Les segments présents dans le lexique avec une fréquence <13/million de mots sont conservés. Ceux avec une fréquence supérieure au seuil et absents du lexique doivent y être ajoutés. Les séquences supérieures au seuil et absentes du lexique sont les suivantes : *-même, -mêmes, -là, -ci, -ce, -en* et *-mesme*. Le cas de -mesme est particulier, car ses occurrences sont principalement issues d'écrits produits avant le XVIIIème siècle. Il y a seulement deux occurrences pour le XXème siècle et zéro pour le XXIème siècle. Les données utilisées pour ce travail sont des productions du XXème siècle, de ce fait, il ne parait pas pertinent de conserver cette séquence.

Pour résumer, les séquences nécessitant une segmentation avant le tiret avec rattachement du tiret à la partie postposée sont les suivantes : *-elle, -elles, -il, -ils, -je, -la, -le, -les, -leur, -lui, -moi, -m', -nous, -on, -t, -toi, -tu, -vous, -y, -même, -mêmes, -là, -ci, -ce* et *-en* (soit 26 séquences).

Contrairement aux séquences avec les apostrophes principalement composées d'une unique lettre, les séquences avec tirets sont davantage spécifiques du français. Cependant, chacune de ces séquences a été observée dans l'échantillon du corpus COCA par précaution.

Sur les 26 séquences, seulement trois ont des occurrences dans le corpus de l'anglais (*-m*, *-le* et *-on*, détail des occurrences en annexe 2). Les occurrences observées de ces séquences sont peu nombreuses, et donc rares. Ainsi, elles ne semblent pas interférer avec la segmentation déjà établie pour l'anglais.

Récapitulatif des ajustements

Les apostrophes et les tirets ne répondent pas aux mêmes règles de segmentation entre l'anglais et le français. Après une étude des usages des séquences qui en contiennent, dans les deux langues, la segmentation peut être ajustée en ajoutant des points de segmentation pour chacun des cas vus précédemment, sans risquer d'impacter la segmentation automatique faite pour l'anglais. Ainsi, une seconde segmentation peut être effectuée après les apostrophes et avant les tirets pour les cas relevés. Pour ne pas se retrouver dans des situations où les séquences retenues sont des séquences internes à des tokens en anglais, une contrainte de début de mot a été ajoutée pour la vérification des séquences avec apostrophe, et de fin de mot pour les séquences avec tiret. La seule exception concerne le cas de « qu' », qui n'a pas la contrainte de début de mot pour pouvoir prendre en compte les dérivés comme *jusqu'*, *lorsqu'*, etc. La séquence « *n'* » prend, en plus d'une contrainte de début de mot, une contrainte sur ce qui la précède. En effet, la segmentation de la séquence est effective uniquement si ce qui suit n'est pas la lettre t. La séquence « *n't* », inconnue en français (13 occurrences dans Frantext qui renvoient toutes à une contraction stylistique *de* ne et *te*), renvoie à une contraction du *not* qui peut parfois se trouver tel quel quand il a déjà été séparé de l'auxiliaire.

Les données utilisées ne sont pas standards et issues de Twitter, de ce fait, quelques variations orthographiques simples ont été prises en compte. Pour chaque segment relevé, l'alternative en majuscule ou sans accent est également pris en compte. De plus, SpaCy sépare le symbole # de la chaîne de caractère sans espace qui suit. Cependant, sur Twitter, le symbole # et les caractères qui suivent forment une seule entité qui est le hashtag. De plus, le contenu d'un hashtag ne reflète pas un choix d'utilisation d'une langue par un locuteur. De ce fait, le symbole # a été rattaché à la chaîne de caractères qui forme le hashtag pour que les lexiques ne soient pas projetés dessus.

La combinaison de la segmentation automatique de SpaCy pour l'anglais et des ajustements pour le français permet d'obtenir une segmentation plus fine et plus adaptée pour chacune des deux langues.

Selon la segmentation effectuée avec les ajustements, l'échantillon du corpus « Twitter-Canada » contient 855.700 tokens. Pour cet échantillon, les tweets contiennent en moyenne 17,11 tokens.

5.2.3. Evaluation de la segmentation après ajustement

La segmentation de SpaCy avec les ajustements a été comparée avec celle faite manuellement (cf *5.2.1. Segmentation automatique*). Dorénavant, la segmentation automatique avec ajustements renvoie 839 tokens (contre 838 avant ajustements).

Certains points de désaccord persistent malgré les ajustements. Les segmentations des segments « <<—- », « pop-ups », «): » et « and/or » restent inchangées à celles proposées avant les ajustements. Désormais, avec les ajustements, il reste plus que 4 points de désaccord (contre 12 avant ajustements). Ainsi, le taux d'accord entre les segmentations est de 99,52%, soit un point de pourcentage de plus qu'avant les ajustements.

La segmentation des segments composés uniquement de ponctuations n'aura pas d'incidence sur la tâche d'identification de langue qui n'opère pas sur les chaînes de caractères sans lettres.

En ce qui concerne « and/or » une segmentation sur la barre oblique impliquerait également une segmentation des liens URL composés également de cette barre oblique. Cette situation serait davantage problématique étant donné qu'un URL n'est pas une production réelle du locuteur, et qu'ils ne reflètent pas une habitude langagière. La décision prise est donc d'ignorer la segmentation d'occurrences du même schéma que « and/or », qui certes poseront un problème pour la suite, mais qui resteront un moins gros problème que la segmentation des URL. A titre informatif, dans l'extrait de l'échantillon du corpus, soit 50 tweets, il y a un seul schéma du type « and/or » contre 21 URL. De ce fait, je préfère conserver une erreur de segmentation dont j'ai connaissance plutôt que de l'éviter au risque d'en générer plus. En ce qui concerne la segmentation de « pop-ups », elle est due à une mauvaise interprétation de l'adverbe *up*, considéré comme un affixe dépendant à cause d'un manque de connaissance en anglais. De ce fait, la décision prise est de conserver les segmentations de SpaCy, sans intégrer d'ajustement concernant les affixes.

La segmentation automatique proposée par SpaCy agrémentée d'ajustements spécifiques aux données bilingues et aux données issues de Twitter semble être satisfaisante pour pouvoir passer à la deuxième étape du processus d'identification de langue : les lexiques. Les lexiques vont permettre d'effectuer des correspondances entre des tokens dont la langue est préalablement déterminée. A l'image de la segmentation, les lexiques sont également un point déterminant de la qualité de l'identification de langue.

5.3. Les lexiques

L'identification de langue repose sur la projection de lexiques sur les tokens du corpus. Deux lexiques sont donc nécessaires : un premier avec des formes attribuées au français et un second avec des formes attribuées à l'anglais. Pour garantir un maximum de correspondances entre les tokens et les lexiques, ces derniers doivent être le plus possible représentatifs d'une langue.

5.3.1. Lexiques de départ

Lexique du français

Le lexique choisi pour illustrer les formes du français est le lexique GLAFF (Hathout et al, 2014). L'objectif de la constitution du lexique est la représentativité. Ainsi, il regroupe une multitude de formes du vocabulaire français. Ce lexique a été constitué à partir du Wiktionnaire[5] (un dictionnaire francophone en ligne libre constitué principalement à partir de dictionnaire et collaboratif). Le lexique GLAFF est composé au total de

[5] *https://fr.wiktionary.org/wiki/Wiktionnaire:Page_d'accueil*

1.406.857 formes. Pour ces formes, des informations supplémentaires sont disponibles, comme les fréquences dans le corpus LM10[6] (constitué d'articles du journal Le Monde de 1991 à 2000).

Lexique de l'anglais

Pour représenter les usages écrits de l'anglais, le lexique ENGLAFF a été choisi. C'est un lexique flexionnel de l'anglais, extrait d'ENGLAWI (Sajous et al, 2020) qui est un dictionnaire normalisé du Wikitionnary[7] (branche anglophone du Wikitionnaire). Contrairement à GLAFF, ce lexique ne propose pas les fréquences des formes. Le lexique ENGLAFF est composé au total de 1.182.213 formes.

Les deux lexiques utilisés pour la détection de langue sont équivalents. Ils sont tous deux très volumineux et ils ont été constitués à partir de données issues du Web. De ce fait, des formes indésirables ou n'appartenant pas à la langue ciblée peuvent troubler l'identification. Ainsi, il semble indispensable d'adapter ces lexiques aux besoins de ce travail.

5.3.2. Ajustements des lexiques

Les lexiques GLAFF et ENGLAFF ont pu être ajustés, car ils sont diffusés sous une *Licence CC BY-SA 3.0*, qui permet leur transformation tant qu'ils sont crédités et rediffusés selon la même licence.

Des ajustements des lexiques doivent être réalisés afin de les adapter le plus possibles aux besoins de l'identification de langue au niveau du token. La segmentation adaptée au français a fait émerger de nouvelles formes qui doivent être présentes dans le lexique GLAFF. Enfin, les formes présentes simultanément dans les deux lexiques doivent être étudiées pour limiter au mieux leur présence qui interfère dans la tâche d'identification de langue.

Adaptations préalables

La première étape consiste à nettoyer les lexiques des informations non pertinentes pour une tâche d'identification de langue de tokens. Ainsi, pour les deux lexiques, les formes ont été conservées, et pour le lexique GLAFF les fréquences relatives des formes dans LM10 ont également été sauvegardées (la fréquence relative est une information qui s'avère importante pour les prochaines étapes d'ajustement des lexiques).

Ensuite, les majuscules présentes dans les formes ont été converties en minuscules, cela permet de favoriser les chances de correspondances avec les tokens qui seront vérifiés en minuscules pour éviter les variations de casse.

Les entrées identiques ont été fusionnées pour ne renvoyer qu'une seule forme. Pour le lexique du français GLAFF, les fréquences relatives des doublons ont été additionnées. Par exemple, à la base, le lexique GLAFF contient deux entrées *son*, une pour le déterminant avec une fréquence relative de 3745.28, et une autre pour le nom avec une fréquence relative de 19.59. Désormais, le lexique contient plus qu'une seule entrée son, avec une fréquence relative de 3764,87. Cette étape est possible, car le sens et la catégorie morpho-syntaxique des formes ne sont pas des critères qui vont être utilisés par la suite. Ainsi, la distinction entre deux formes identiques n'est pas nécessaire, puisque la seule information attendue est la langue à laquelle elles appartiennent, et pour GLAFF, la fréquence totale, peu importe le type d'emploi. De plus, les formes débutant par un chiffre ont été exclues des lexiques, elles sont très nombreuses et ne participent pas à l'attribution de langue.

Le processus d'adaptation de la segmentation pour le français produit des formes composées d'apostrophes ou de tirets. Les formes relevées dans la partie *5.2.2 Ajustement de la segmentation* absentes de GLAFF y ont été ajoutées afin de garantir la correspondance des formes dont les usages ont été rattachés uniquement au français, soit les formes suivantes : *j'*, *-même*, *-mêmes*, *-là*, *-ci* et *-en*. Les formes contenant des caractères

[6] *http://redac.univ-tlse2.fr/voisinsdelemonde/infos/apropos.jsp*

[7] *https://en.wiktionary.org/wiki/Wiktionary:Main_Page*

accentuées ont été ajoutées également dans une version sans accent pour accroître les correspondances. Ce sont donc 9 formes qui ont été intégrées dans le lexique.

Ensuite, les apostrophes ont été harmonisées entre les deux lexiques et avec le corpus. Les jeux de caractères des fichiers informatiques proposent deux transcriptions de l'apostrophe : dactylographiée (') et typographiée (') (*Office québécois de la langue française*[8]). Cette différence, bien que difficilement visible, est discriminante pour la mise en correspondance de chaînes de caractères. Ainsi, par précaution, toutes les apostrophes typographiées ont été remplacées par une apostrophe dactylographiée. Cette conversion a également été effectuée sur le corpus pour les mêmes raisons.

Détection des homographes

Les homographes (ici) désignent des formes graphiques identiques (sans prise en compte du sens ou de la catégorie morpho-syntaxique) qui appartiennent au deux langues selon les lexiques. Ces homographes génèrent une ambiguïté au moment de l'identification de langue qui ne prend pas en considération le contexte du token. La présence des homographes dans les lexiques s'explique dans un premier temps par l'intégration de termes anglais dans le lexique du français (ex : *hangover*) ou inversement (ex : *deux*). De plus, dans les homographes se trouve des cognats, c'est-à-dire des tokens qui ont le même sens et les mêmes usages entre les deux langues (ex : *image*) et des faux-amis qui sont des tokens sont graphiquement similaires entre les deux langues, mais leur sens et leurs usages diffèrent (ex : *for*).

L'objectif est de lever le maximum les ambiguïtés au moment de l'attribution de langue. De ce fait, il est important de limiter le nombre d'homographes. Pour cela, il faut définir la langue d'usage la plus probable pour les homographes.

Dans un premier temps, les formes présentes à la fois dans GLAFF et dans ENGLAFF ont été extraites afin de recenser l'ensemble des formes homographiques. Au total, 33.297 formes sont communes aux deux lexiques. Le choix a ensuite été fait d'étudier uniquement les formes présentes dans le corpus pour diminuer leur nombre et faciliter leur analyse, sans pour autant perdre en qualité de détection des tokens du corpus. Sur les 33.297 homographes initiaux, 5315 sont présents dans le corpus. Ce sont ces 5315 formes qui vont être étudiées par la suite.

Catégorisation des homographes

Une fois la liste des homographes du corpus obtenue, leurs usages doivent être étudiés afin de déterminer s'ils sont employés équitablement entre les deux langues ou s'ils sont davantage représentatifs d'une langue en particulier. Pour déterminer la langue d'usage d'un homographe, la méthode la plus fiable est d'observer son contexte. Cependant, cette méthode, certes plus fiable, est plus longue et fastidieuse. C'est pourquoi, l'étude des usages des homographes qui suit porte sur le critère de fréquence et non de contexte. Ainsi, l'objectif de l'étude de ces homographes est de pouvoir les répartir selon trois catégories : les homographes du français, les homographes bilingues et les homographes de l'anglais.

Dans les catégories des homographes monolingues, se retrouvent les formes du lexique qui appartiennent à l'autre langue et qui ne sont pas intégrées dans la langue cible du lexique, ainsi que les formes existant dans les deux langues, mais dont les usages sont suffisamment déséquilibrés pour pouvoir les attribuer à la langue la plus probable avec la plus petite marge d'erreur. Cette approche a déjà été proposé dans l'étude de Nguyen et Doğruöz (2013), où là également, la fréquence des usages permet de définir la probabilité de la langue.

Récupération des fréquences relatives des homographes

Désormais, l'objectif est de récupérer les fréquences relatives des homographes depuis des corpus ciblant la représentativité de l'anglais et du français. Cette représentativité est un critère important pour garantir le reflet le plus réel des usages des formes par les locuteurs.

Le corpus COCA a été de nouveau utilisé pour cette étape (présentation du corpus dans la partie *5.2.2. Ajustements de la segmentation*). La diversité des genres qui le compose et son nombre de tokens permet

[8] *http://bdl.oqlf.gouv.qc.ca/bdl/gabarit_bdl.asp?id=5173*

d'espérer une fréquence relative fiable. La table de fréquence absolue des formes présentent dans COCA a été extraite. Ces formes ont toutes été converties en minuscule également. Pour les formes identiques, suite à la mise en minuscule par exemple, les fréquences ont été additionnées. Ensuite, les fréquences absolues des formes ont été transformées en fréquence relative en divisant la fréquence absolue d'une forme par la taille du corpus (soit. 12.153.815 pour COCA) et en multipliant ce résultat par un million (pour faciliter la lecture et la comparaison).

Ensuite, la table des fréquences a été croisée avec la liste des homographes établie précédemment. Sur les 5315 homographes du corpus, 4856 d'entre eux sont présents dans COCA. L'absence de 459 homographes de la liste peut être interprétée comme une absence d'usage de ces derniers en anglais. De ce fait, pour ces 459 homographes, une fréquence relative de 1 pour 100 millions leur a été attribuée. Ce nombre permet d'illustrer l'absence d'une forme dans le corpus COCA sans utiliser simplement zéro qui bloquerait de futurs calculs (la division par zéro étant impossible).

Désormais, il reste à faire la même chose pour le français. La même démarche a été utilisée pour obtenir les fréquences relatives des homographes du corpus. Dans un premier temps, la récupération des fréquences a été réalisée à l'aide du corpus DIV-706. Les fréquences proposées dans GLAFF n'ont pas été utilisées comme premier choix car elles sont issues de corpus moins comparables à COCA utilisé pour l'anglais. Le corpus DIV-706 est composé de différents genres, se rapprochant davantage de ceux présents dans COCA. L'objectif est donc d'observer les fréquences dans les deux langues dans des corpus les plus comparables possible.

Le corpus DIV-706 se compose de textes issus de 14 genres différents et variés (articles de presse, articles scientifiques, articles Wikipédia, compte-rendus médicaux, critiques de films, discours politiques, discussions Wikipédia, entretiens, exposés, littérature jeunesse, profils de coach surfers, résumés de films, sous-titres de films et textes réglementaires - détail des sources des données en annexe 4). Ce corpus a été réalisé pour une utilisation pédagogique et n'est pas publié à l'heure actuelle. Il a été choisi parce qu'il est composé de textes non-standards et variés.

Sur 5315 homographes du corpus, 2925 d'entre eux sont présents dans le corpus DIV-706. Contrairement à l'anglais, les formes absentes du corpus DIV-706 n'ont pas tout de suite reçu la fréquence de 1 pour 100 millions. Le corpus DIV-706 est composé de 951.073 tokens, soit 11 millions de moins que COCA. Pour que la taille du corpus DIV-706 ne soit pas une répercussion directe sur les fréquences, pour les homographes absents de ce corpus, la fréquence relative dans le corpus LM10 a été retenue (information disponible directement depuis le lexique GLAFF). La fréquence dans LM10 a été choisie parce que les données issues de ce corpus journalistique sont garanties comme étant conventionnelles à l'usage du français et le recours aux formes non-francophones est limité. Cependant, ces fréquences n'ont pas été utilisées en premier choix, car elles sont spécifiques au genre journalistique et manquent donc de représentativité (par exemple, les marques d'adresse directe comme *ta* ou *vos* ont une fréquence de 0 dans LM10). Certaines formes absentes du corpus DIV-706 ont obtenu une fréquence de zéro dans LM10. Là aussi, les fréquences de zéros ont été transformés en une fréquence de 1 pour cent millions pour les futurs calculs.

Pour faciliter les comparaisons entre les langues et entre les homographes, les fréquences sont converties en pourcentages qui illustrent la proportion des usages dans chaque langue. Pour cela, la fréquence relative dans chaque langue est divisée par la somme des fréquences relatives dans les deux langues, puis le résultat est multiplié par 100.

Attribution de langue des homographes

La répartition en pourcentage permet d'établir une échelle commune à tous les homographes qui définie leur langue d'usage la plus fréquente. Trois niveaux de répartition sont attendus : les formes qui doivent être rattachées uniquement au français, celles qui doivent être rattachées aux deux langues et enfin celle qui doivent être rattachées à l'anglais. Il faut désormais définir les limites de ces niveaux qui vont organiser la catégorisation des homographes. La première proposition de délimitation est de diviser l'échelle des catégories en trois parties équitables, soit avec des changements de catégorie à 33% et à 66%. Après observation de quelques données, cette répartition arbitraire semble correspondre avec les pourcentages de la part des fréquences en anglais des homographes observés :

- Inférieur à 33% → formes du français → 1558 formes
- Supérieur ou égal à 33% et inférieur à 66% → formes bilingues → 1047 formes
- Supérieur ou égal à 66% → formes de l'anglais → 2710 formes

Sur les 5315 homographes présents dans le corpus, 1558 sont considérées comme étant des formes du français, 1047 comme étant des formes appartenant à l'anglais et au français et 2710 comme étant des formes de l'anglais. Un peu plus de la moitié des homographes initiaux sont attribués à l'anglais.

De ce fait, les formes qui ont été considérées comme ayant le plus de probabilités d'appartenir au français qu'à l'anglais ont été supprimées du lexique de l'anglais, et inversement. Cette méthode n'est pas infaillible, mais elle est à mon sens la plus adaptée pour le traitement automatique d'un grand nombre d'homographes.

5.3.3. Lexiques finaux

Les lexiques ont subi des ajustements pour répondre au mieux aux critères de précision et de fiabilité que nécessite la tâche d'identification de langue au niveau du token. Le tableau 2 récapitule l'ensemble des étapes d'ajustement dans l'ordre chronologique pour chaque lexique.

		GLAFF	ENGLAFF
1	Extraction des formes	✓	✓
2	Extraction des fréquences relatives	✓	✗
3	Conversion des majuscules	✓	✓
4	Suppression des doublons et des chiffres	✓	✓
5	Ajout des formes issues de la segmentation adaptée	✓	✗
6	Harmonisation des apostrophes	✓	✓
7	Traitement des homographes	✓	✓

Tableau 2 : Etapes d'ajustement des lexiques GLAFF et ENGLAFF

Ces étapes d'ajustement ont entrainé des modifications sur les lexiques initiaux. Ainsi, le lexique GLAFF comporte désormais 1.019.977 tokens et le lexique ENGLAFF, 611.732 tokens (tableau 3). Ce sont ces lexiques après ajustement qui vont être utilisé pour la suite de ce travail.

	GLAFF	ENGLAFF
Nombre de tokens de départ	1.406.857	1.182.213
Nombre de tokens après ajustements	1.019.977	611.732

Tableau 3 : Taille des lexiques avant et après ajustements

6. Etiquetage des langues du corpus

La segmentation et les lexiques permettent de réaliser la tâche d'identification de langue prévue au niveau du token. Il est nécessaire d'étiqueter chaque token du corpus en fonction de la langue à laquelle il se rattache ou de son rôle dans le tweet.

6.1. Jeu d'étiquettes

L'objectif est de proposer un jeu d'étiquettes qui permet de couvrir l'ensemble des configurations. Ce jeu d'étiquettes doit également rendre compte des besoins d'analyses futurs, c'est-à-dire que les étiquettes doivent être liées à l'identification de langue et aux spécificités langagières de Twitter.

Dans un premier temps, les étiquettes doivent refléter l'identification de la langue effectuée à l'aide des lexiques. L'étude du code-switching porte uniquement sur l'anglais et le français. Ainsi, il faut des étiquettes pour le français, l'anglais et le bilingue.

La ponctuation des messages écrits n'est pas un facteur discriminant pour l'identification de la langue, en revanche, elle permet d'accéder à l'organisation du message. Les signes de ponctuation sont regroupés sous la même étiquette.

Dans les tweets, il y a des mentions et des hashtags. Ces segments directement intégrés dans le message sont spécifiques aux productions issues de Twitter. Ils permettent l'intégration du message dans un ensemble (hashtag) ou alors de cibler le destinataire du message ou la personne sur qui porte le message (mention). Ces parties de discours ne rentrent pas dans l'identification de la langue, car elles participent principalement à la contextualisation du message. Les hashtags et les mentions doivent donc être étiquetés en fonction de leur nature.

Les tweets peuvent contenir des liens URL. Ces liens sont une extension du message, mais ils ne reflètent pas une production langagière. De plus, les liens renvoient à du contenu qui peut être spécifique à une langue, en revanche, le lien en lui-même n'est pas un marqueur utile pour l'identification de langue. De ce fait, les liens sont regroupés sous une même étiquette afin de pouvoir aisément les mettre de côté si besoin.

Enfin, une dernière étiquette regroupe l'ensemble des tokens qui ne peuvent recevoir aucune des autres étiquettes.

Pour résumer, le jeu d'étiquettes choisi pour couvrir le plus fidèlement l'ensemble des tokens du corpus contient 8 étiquettes :

- F → tokens qui relèvent du français
- E → tokens qui relèvent de l'anglais
- B → tokens qui relèvent de l'anglais et du français
- P → tokens de ponctuation
- M → tokens dont le rôle est de mentionner un individu dans le message
- H → hashtags
- U → liens URL
- I → inconnu, autre

6.2. Attribution des étiquettes

Les étiquettes spécifiques à l'identification de langue (F, E, B) ont été attribuées en vérifiant les correspondances des tokens avec les lexiques de l'anglais et du français. L'étiquetage de la ponctuation s'est fait à partir d'une liste pré-définie de différents signes. Les tokens qui entre en correspondance avec cette liste sont étiquetés P. Pour les étiquettes adaptées au genre des données (M, H, U), elles ont été attribuées uniquement en se basant sur le début du token. Par exemple, les tokens commençant par @ ont été étiquetés M, par # ont été étiquetés H et enfin ceux commençant par *http* ont été étiquetés U. Enfin, les tokens qui ne possèdent aucune des étiquettes précédentes ont été étiquetés I. Le groupe de tokens étiquetés I est celui sur lequel il n'y a pas de contrôle. De ce fait, l'ensemble des tokens de ce groupe ont été extraits afin de pouvoir comprendre ce qui compose ce groupe. Un extrait du corpus étiqueté est disponible en annexe 3.

6.3. Répartition générale des étiquettes

Suite à l'étiquetage de l'ensemble des tokens, il est possible d'avoir une vue générale sur la constitution du corpus concernant la répartition des langues ou la présence des spécificités liées à Twitter. Au total, le corpus contient 855.770 tokens. Avec la représentation graphique des données en pourcentage (figure 2), un premier déséquilibre est mis en avant. En effet, les tokens de l'anglais sont davantage représentés (72,89% des tokens). Cette sur-représentation de l'anglais n'est pas inattendue, car le corpus a été constitué dans le but d'être un corpus de l'anglais. La ponctuation est la deuxième catégorie la plus représentée, loin derrière l'anglais, avec 9,72% des tokens du corpus. La catégorie des tokens inconnus arrive en troisième position en ordre décroissant avec 6,02% des tokens, suivie des hashtags (3,20%), puis des tokens bilingues (3,05%). Les mentions et les liens URL arrivent ensuite avec 1,94% et 1,65% de couverture des tokens du corpus. La dernière catégorie, la moins représentée, est celle des tokens du français avec seulement 1,54% des tokens. Ce pourcentage est assez conséquent pour un corpus qui se veut être composé uniquement d'anglais. A présent, les étiquettes assignées vont pouvoir être utilisées pour étudier les contextes et les structures dans lesquelles les tokens qu'elles représentent s'inscrivent dans les tweets.

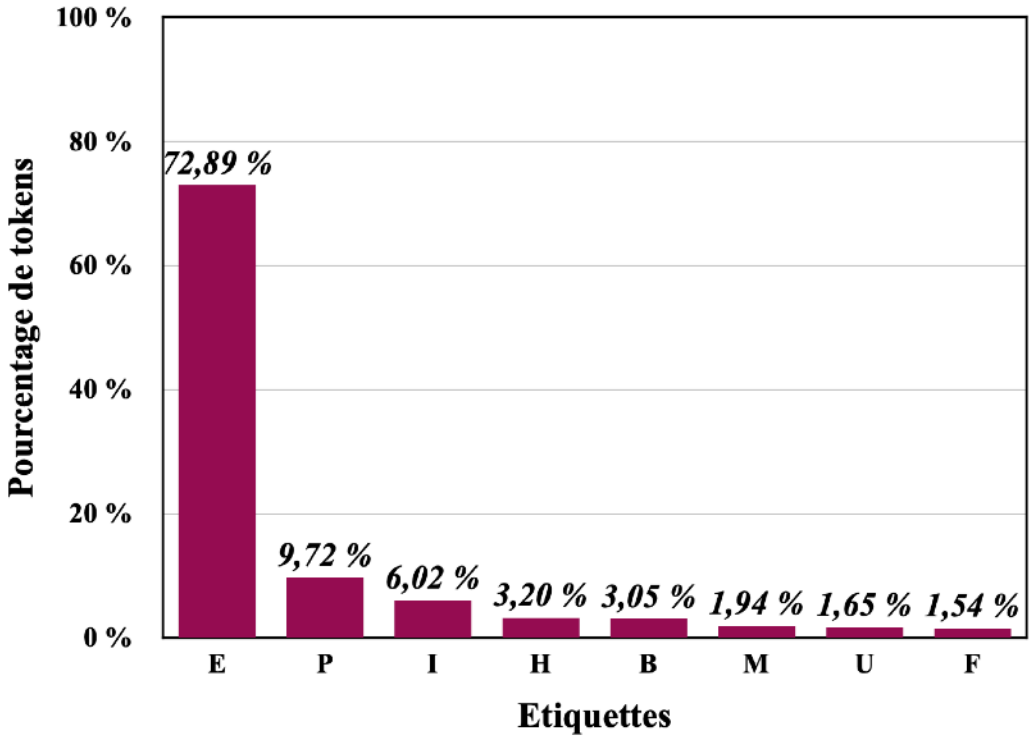

Figure 2 : Répartition en pourcentage des tokens du corpus par étiquette

6.4. Observation des tokens

Les tokens bilingues
La table de fréquence des tokens bilingues montre que les tokens bilingues les plus fréquents sont des mots grammaticaux en anglais et en français. La préposition de l'anglais *on*, qui est aussi un pronom du français, est le token bilingue le plus fréquent avec 5813 occurrences. Il est suivi du token *me* avec 2998 occurrences et de *as* avec 1925 occurrences. Ces tokens sont fréquents aussi bien en anglais qu'en français.

Les tokens français
En ce qui concerne les tokens du français, le plus fréquent est la préposition *de* avec 278 occurrences. Dans un premier temps, il s'avère que les fréquences de tokens en français sont nettement inférieures par rapport aux fréquences de tokens bilingue. Le deuxième token le plus fréquent est le nom *article* avec 183 occurrences. Cependant, ce terme existe en anglais également, de ce fait, il aurait dû être considéré comme bilingue. Sa fréquence aussi élevée ne traduit donc pas les usages en français, car une grande partie de ses occurrences doit être issue de tweets en anglais.

Les tokens inconnus
Les tokens dont l'étiquette I (inconnu) a été attribuée ont été extraits et observés afin de comprendre ce qui compose ce groupe d'inconnus à l'aide de la table des fréquences.

Les émoticônes, les émojis et certains symboles sont également considérés comme inconnus. Ils sont les tokens inconnus les plus fréquents. Les répertorier pour les étiqueter différemment du reste serait une tâche longue et peu utile ultérieurement, car leurs distinctions ne participent pas à l'identification de langue ou à l'observation des structures du code-switching. Ensuite, il y a les chiffres qui ne participent pas à la tâche d'identification de langue. De ce fait, l'assignation de l'étiquette « *inconnu* » ne trouble pas les traitements qui vont par la suite être réalisés, comme l'extraction du code-switching.

Dans les tokens inconnus, se trouvent certains noms propres et certaines entités nommées. *Montréal* est l'entité nommée la plus fréquente, avec 109 occurrences. Certaines d'entre elles peuvent être davantage associées à une culture ou une zone géographique, mais cette association n'est pas suffisamment fiable pour les intégrer dans le processus d'attribution de langue.

Enfin, dans la liste des tokens inconnus du corpus, se trouve les tokens avec orthographe déviante (*cf. 2.2. Caractéristiques des CMR*). Certains tokens correspondent à des abréviations (exemple (11) : btw → by the way, exemple (12) : OMG → oh my god), à des allongements (exemple (13) : timeeeeeee → time), à des fautes d'orthographe (exemple (14) : Unbelliveable → unbelievable) ou à des fautes de frappe (exemple (15) : htat → that). Ces tokens relèvent d'une langue précise généralement (l'anglais pour les exemples précédents). Cependant, ces variations orthographiques ne sont pas incluses dans le lexique. Adapter le lexique à l'ensemble des variations possibles serait une tâche très longue qui aurait peu de chance d'aboutir étant donné la multitude de variations possibles pour chaque token. De ce fait, ce problème se règle plus facilement en proposant une normalisation des tokens avant l'identification de langue.

(11) *manga spoiler ? btw what's this* (TweetId : 942111015147266049)

(12) *Haha .. OMG awesome* !! (TweetId : 933899501374492673)

(13) *Casting timeeeeeeee !! Let's pick some models* 👀 #nemracstyle #casting #fashion #models #search … https://t.co/ttfTrNUAcC (TweetId : 704066092851818496)

(14) *Amazing Look Miami ! Unbelliveable how You are Sexy !* 🎉😘 (TweetId. : 780445392764829696)

(15) *Logged into the azure portal for the first time to setup an AD for SAML . Not htat this is a bad thing but ... IDK what I am doing ...* #Azure (TweetId : 1086267732545413126)

7. Identification et catégorisation du code-switching

L'identification et l'analyse du code-switching peut difficilement se faire avec l'étiquetage actuel. En effet, les multiples possibilités de répartition et d'organisation des étiquettes ne permettent pas d'en dégager des schémas. Les schémas des tweets, qui permettront l'identification et l'analyse du code-switching, sont obtenues en observant les signatures sur différents niveaux.

7.1. Les signatures

7.1.1. Présentation

L'objectif des signatures est de proposer une représentation schématique des données. L'accès au contenu d'un tweet ne se fait plus par ses tokens, mais par les étiquettes attribuées à chaque token qui le compose (exemple ligne N0 du tableau 4).

Ce type de représentation des tweets permet de faire des regroupements et des catégorisations plus largement. De plus, il est possible de quantifier les différents schémas qui seront obtenus à l'aide de ces signatures.

Les signatures des tweets s'obtiennent en extrayant les étiquettes des tokens et en conservant leur ordre. Différents traitements peuvent être appliqués sur ces signatures pour cibler les informations qui seront utiles par la suite.

La représentation du tweet en signature permet d'accéder plus facilement aux informations concernant les langues utilisées, mais également leur organisation au sein du message contenu dans le tweet. En fonction de ce qui est recherché, les signatures peuvent être exprimées selon différents niveaux.

7.1.2. Les différents niveaux de signatures

Les signatures qui vont servir pour l'étude et l'analyse du code-switching peuvent prendre différentes formes. Elles s'organisent selon plusieurs niveaux en fonction des potentiels besoins futurs. Chaque niveau se construit à l'aide d'un niveau précédent.
Le tableau 4 propose un schéma représentant les différentes formes des signatures d'un tweet selon les quatre niveaux choisis qui vont être présentés.

	@Qc511_Mtl	*seriously*	*big*	*problems*	*exit*	*13*	*south*	*Gouin*	*to*	*take*	*13*	*north*	*!*	*Vous*	*devez*	*trouver*	*une*	*solution*	*!*	*!*	*!*	*!*	*!*
N0	M	E	E	E	E	I	E	**B**	E	E	I	E	P	F	F	F	F	F	P	P	P	P	P
		↓	↓	↓	↓		↓	↓	↓	↓		↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
N1		E	E	E	E		E	**E**	E	E		E	P	F	F	F	F	F	P	P	P	P	P
N2		E	E	E	E		E	E	E	E		E		F	F	F	F	F					
N3		E											**P**			F							
N4	**E**												**F**										

Tableau 4 : Schéma de représentation d'un tweet selon les différents niveaux de signatures (TweetId : 850086089310208000)

Niveau 0

Le niveau 0 est le niveau de représentation fidèle du tweet, ce sont simplement les étiquettes de chaque token qui sont conservées et alignées. Les résultats de ce niveau facilement interprétables, en revanche, il est compliqué d'obtenir des séquences identiques entre différents tweets à regrouper pour permettre une catégorisation. En effet, le niveau 0 comptabilise 35.863 signatures différentes (figure 3), cela signifie qu'une grande partie des signatures n'ont qu'une seule occurrence. La signature la plus fréquente de ce niveau, EEE, compte 494 occurrences, ainsi, 494 tweets du corpus sont composés uniquement de trois tokens en anglais et rien d'autre. Toutes les autres configurations représentent donc moins de 494 tweets à la fois. Ce niveau ne permet pas d'obtenir une représentation lisible du corpus.

Niveau 1

Le niveau propose une version lissée du niveau 0. Dans ce niveau, seules les étiquettes des tokens participant à l'identification de langue sont conservés, tout comme les étiquettes de ponctuation. Ainsi, les étiquettes relevant des spécificités de Twitter ou de données médiées par les réseaux (H, M et U) sont écartées, tout comme les étiquettes faisant référence aux mots inconnus (I). L'objectif est d'accéder simplement aux informations utiles pour l'identification et la catégorisation du code-switching. Sur ce principe, les ambiguïtés alimentées par les étiquettes B (bilingues) doivent être levées. Pour cela, elles ne sont pas prises en compte, la résolution d'un token bilingue est possible en fonction de son contexte. Alors, de cette manière, si une séquence d'étiquettes B est prise entre deux étiquettes E, elle peut être considérée comme E également, de même si elle est prise entre une étiquette E et une étiquette F, le fait qu'elle soit F ou E n'impacte pas l'organisation ou la structure du code-switching et elle peut ainsi s'effacer. Cette méthode permet de lever toutes les ambiguïtés possibles d'identification de langue, en revanche cela entraîne une perte dans le nombre de tokens des tweets. Les étiquettes de ponctuation (P), quant à elles, sont conservées. Ce niveau de signature permet d'accéder aux mots d'une langue isolés dans une séquence d'une autre langue, mais aussi, à l'aide de la ponctuation conservée, d'étudier de potentielles structures du texte bilingue.

Pour ce niveau, seulement 49.838 tweets sont conservés, les 162 tweets perdus sont ceux qui sont composés uniquement de mentions, hashtags, liens URL ou de tokens inconnus. Le niveau 1 permet de représenter l'ensemble du corpus avec 16.298 signatures différentes (figure 3). Moins de signatures sont nécessaires par rapport au niveau 0, mais il reste tout de même une grande quantité difficilement observable en l'état. Tout comme pour le niveau 0, la signature la plus fréquente est EEE avec 1589 occurrences. Seulement, dans ce cas, cela signifie que 1589 tweets sont composés de trois tokens en anglais et potentiellement de liens URL, de mentions, de hashtags ou de tokens inconnus.

Niveau 2

Le niveau 2 est une variante du niveau 1 dont les étiquettes de ponctuation ont été effacées. L'utilité de ce niveau est de pouvoir comptabiliser la répartition de chaque langue dans un tweet, et ainsi de proposer un score de présence du français ou de l'anglais pour chaque tweet. Le niveau 2 s'applique sur 49.789 tweets, soit 49 de moins que le niveau 1. Les 49 tweets qui ne sont pas pris en compte sont ceux qui étaient composés uniquement de tokens étiquetés P au niveau 1. 2538 signatures différentes permettent de représenter l'ensemble des tweets restants (figure 3). La signature la plus fréquente (EEEE) regroupe 2615 tweets. Avec ce niveau, la signature la plus fréquente devient celle qui regroupe quatre tokens en anglais, ni plus, ni moins, dans un tweet. Cela signifie que, dans les signatures précédentes, il y avait déjà ce nombre d'occurrences de tweets de quatre tokens en anglais, mais ce n'était pas aussi visible, car les ponctuations généraient plusieurs signatures différentes pour ces tweets.

Niveau 3

Le niveau 3 est construit à partir du niveau 1. Pour ce niveau, dans un premier temps, les étiquettes identiques contiguës sont regroupées pour former plus qu'une seule unité. Par exemple, une suite de trois étiquettes E devient une seule séquence E. A ce niveau, les étiquettes deviennent des séquences. Le terme *étiquette* renvoie à un token, tandis que l'utilisation du terme *séquence* implique un possible regroupement de tokens. De plus, ce niveau implique un filtrage des séquences de ponctuation (P), seules certaines ponctuations sont conservées. Les étiquettes P prises entre deux étiquettes de langue identiques (EPE ou FPF) sont effacées. De même pour les étiquettes P situées en fin ou début de signature. En revanche, si

l'étiquette P n'a pas la même séquence de langue entre son contexte droit et son contexte gauche, alors elle est conservée. Le niveau 3 permet de rendre compte de la ponctuation uniquement quand elle coïncide avec un point de changement de langue. Cela permet de potentielles distinctions entre les structures de code-switching intra-phrastiques et inter-phrastiques. Pour ce niveau, 137 signatures permettent de représenter l'ensemble des tweets (figure 3). Il est désormais plus facile d'observer la distribution des signatures du corpus. La signature la plus fréquente (E) renvoie aux tweets composés exclusivement de tokens en anglais, avec 41.200 occurrences. La proportion de tweets monolingue anglais est accessible plus facilement à ce niveau. Les tweets avec code-switching intra-phrastiques par juxtaposition EFE sont en deuxième position avec 3731 occurrences.

Niveau 4

Le niveau 4 correspond à la forme la plus canonique et la plus réduite de la structure du tweet, elle correspond au niveau 3 sans les séquences P. Dans la version de ce niveau, seules les séquences informant sur les langues utilisées sont conservées (E ou F). Cette signature est la plus minimale possible et elle permet de rendre compte de l'organisation de l'alternance des langues (juxtaposition ou insertion). Désormais, seulement 27 signatures sont nécessaires pour rendre compte de l'ensemble des tweets (figure 3). La signature la plus fréquente reste celle qui renvoie aux tweets monolingues anglais avec toujours 41.200 occurrences. La deuxième signature est également EFE mais cette fois avec 5044 occurrences, cela signifie que les tweets par insertion avec uniquement deux points de changement de langue ont été regroupés, peu importe la concordance de ces points avec la ponctuation.

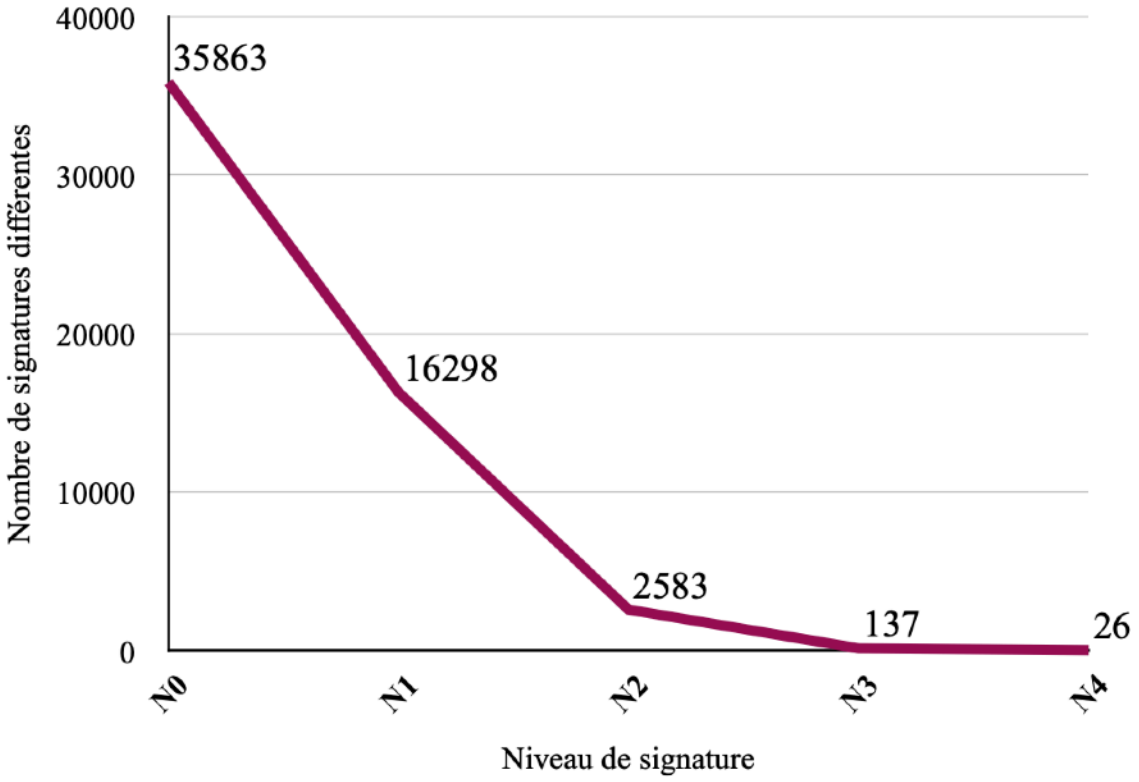

Figure 3 : Nombre de signatures différentes par niveau

Les mentions, les hashtags et les liens URL traduisent un type de données : les communications médiées par le réseau social Twitter. Cependant, ces informations ne sont pas impliquées dans le processus d'identification de la langue, et par la suite, dans le repérage du code-switching et dans l'étude des phénomènes langagiers qui en découlent. C'est pour cette raison que ces informations sont exclues des signatures dès le premier niveau, elles ne seront pas exploitées plus tard pour la détection et l'analyse du code-switching, tout comme les tokens inconnus.

Les signatures de niveaux 3 et 4 sont celles qui permettent davantage la catégorisation et la quantification plus générale des phénomènes. Néanmoins, les niveaux inférieurs restent nécessaires pour la construction des niveaux supérieurs, mais également pour une analyse et une étude plus fine des contenus des tweets par la suite.

7.2. Identification du code-switching

Les signatures de différents niveaux permettent de quantifier les tweets selon la présence ou l'absence de code-switching ainsi que le type de code-switching qu'ils contiennent.

7.2.1. Couverture du code-switching

Dans un premier temps, l'information à obtenir est la proportion de tweets présentant du code-switching. Il se repère en identifiant les tweets qui contiennent à la fois des étiquettes E et F. La table de fréquence des signatures de niveau 4 (Annexe 5) indique que 41.390 tweets sont composés d'une seule langue. Parmi eux, 41.200 tweets sont monolingues en anglais et 190 tweets sont monolingues en français. Puis, 8399 tweets qui se composent à la fois de français et d'anglais. Ainsi, sur les 50.000 tweets, 82,78% d'entre eux sont monolingues et 16,8% des messages contiennent à la fois de l'anglais et du français.

Sur la totalité des tweets, 0,42% d'entre eux ne contiennent ni une étiquette F, ni une étiquette E. C'est le cas pour le tweet de l'exemple (16) qui contient uniquement une interjection et un emoji. De ce fait, ces tweets ne seront pas pris en compte par la suite.

(16) Woah 😻 (TweetId : 1027275380141555712)

7.2.2. Proportion des langues par tweet

Les signatures de niveau 2 permettent de calculer la proportion d'anglais et de français dans les tweets. Seules les étiquettes des tokens participant à l'identification de langue sont comptabilisées (E ou F). Ainsi, la proportion d'anglais n'est pas impactée par la quantité de signes de ponctuation ou de mentions par exemple.

Le taux d'anglais dans un tweet s'étend de 0,00 à 1,00. Le plus haut score, 1,00, indique que le tweet est composé uniquement de tokens en anglais. A l'inverse, le score 0,00 renvoie à un tweet produit exclusivement en français. Un score de 0,50 indique que le tweet est composé d'autant de tokens en anglais qu'en français.

Désormais, l'en-tête de chaque tweet contient un score qui reflète la proportion d'anglais, en plus de l'identifiant de l'utilisateur et du tweet et du score de proportion de tweet en anglais de l'utilisateur.

La figure 4 représente le pourcentage de tweets par proportion d'anglais qu'ils contiennent. Les tweets monolingues n'ont pas été intégrés puisque l'objectif est d'observer la répartition pour les tweets avec code-switching.

Environ 77% des tweets bilingues sont constitués majoritairement d'anglais, tandis que moins d'1% contiennent majoritairement du français. Avec la figure 4, il apparaît que plus la part d'anglais dans un tweet bilingue augmente, plus le nombre de tweets concernés augmente également.

Ainsi, les langues dans les tweets avec code-switching ne sont pas réparties équitablement. En effet, l'anglais domine largement le français dans les énoncés bilingues. Cependant, cette conclusion ne peut être généralisée, car la répartition générale entre l'anglais et le français est largement déséquilibrée dans le corpus de base.

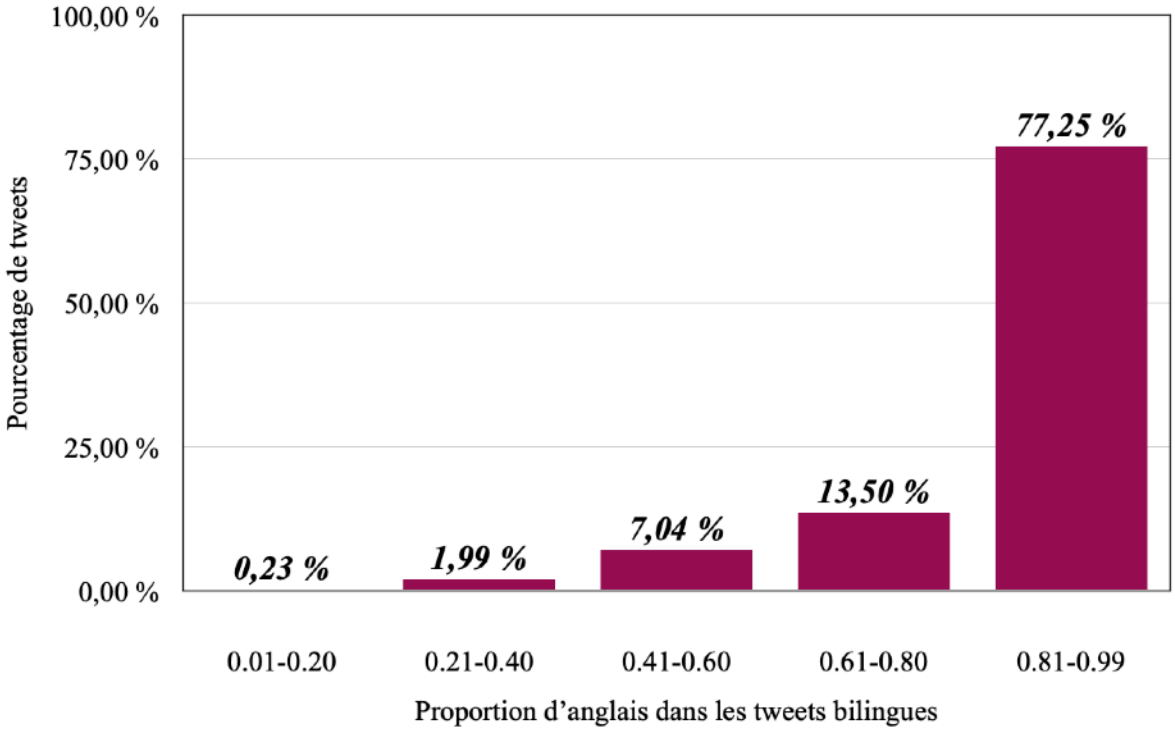

Figure 4 : Répartition des tweets par proportion d'anglais

7.3. Catégorisation du code-switching

Le code-switching, une fois identifié, peut être caractérisé en fonction de ses points de changement de langue dont l'analyse permet d'établir la catégorisation des différentes formes qu'il peut prendre.

7.3.1. Organisation de l'alternance du code-switching

Les signatures de niveau 4 permettent d'observer l'organisation d'alternance des langues dans les tweets contenant du code-switching. Ce pan d'analyse ne prend pas en compte la répartition de la ponctuation. Pour l'étude de l'alternance des langues, la seule information pertinente est le nombre de points de changement. Deux possibilités d'organisation d'alternance ont été relevées dans le point *1.2.1 Les types d'alternance du code-switching* : la juxtaposition ou l'insertion.

Juxtaposition
L'alternance peut se faire par juxtaposition, où dans cette configuration, il y a un seul point de changement de langue. Cela se traduit par des tweets amorcés dans une langue et terminés dans une autre, sans retour à la langue d'amorce. Ces structures correspondent à des signatures EF ou FE pour le niveau 4, c'est-à-dire à des signatures composées uniquement de deux séquences.

D'après la table de fréquence des signatures de niveau 4, 1980 tweets contiennent seulement un point de changement de langue. Ils représentent 24% des tweets avec code-switching. Parmi eux, 835 tweets sont amorcés en français pour être ensuite terminés en anglais, contre 1145 tweets amorcés en anglais et terminés en français. Alors, dans les tweets avec code-switching par juxtaposition, 42,17% d'entre eux vont du français vers l'anglais (FE) et 57,83% de l'anglais vers le français (EF).

Insertion
L'insertion est une stratégie d'alternance qui consiste à insérer une langue entre deux séquences d'une autre langue. Ce type de code-switching par insertion contient donc au minimum deux points de changement. L'alternance par insertion peut se répéter plusieurs fois dans un même énoncé. Les signatures de niveau 4 composées de plus de trois séquences réfèrent à des tweets avec un code-switching par insertion.

La table de fréquence des signatures de niveau 4 indique que pour 6419 tweets bilingues le code-switching est inséré. Cela représente alors 76% des tweets avec code-switching. Pour les tweets sans répétition de l'insertion (c'est-à-dire avec que deux points de changement, FEF ou EFE), le français est davantage inséré entre deux séquences en anglais. Dans les structures avec deux points de changement, 99% d'entre elles sont de type EFE. Ainsi, contrairement au code-switching par juxtaposition, il y a davantage de déséquilibres concernant la répartition du français et de l'anglais.

7.3.2. Structure du code-switching

La structure du code-switching correspond aux potentiels liens des points de changement avec la ponctuation. Dans le point *1.2.2 Les types de structures syntaxiques du code-switching*, deux possibilités sont énoncées. La première est la concordance d'un point de changement de langue avec une ponctuation (inter-phrastique). La ponctuation sert de frontière entre des éléments syntaxiques, de ce fait, le point de changement de langue coïncide avec un changement de phrase ou de segment. La seconde possibilité est un code-switching intra-phrastique. Dans cette configuration, le point de changement de langue ne concorde pas avec la ponctuation du tweet, et donc avec un potentiel changement de segment. La structure du code-switching s'observe à l'aide des signatures de niveau 3 qui conservent les séquences de ponctuation lorsqu'elles sont prises entre deux séquences différentes. Néanmoins, les différentes structures du code-switching ne s'excluent pas. En effet, il est possible de retrouver du code-switching inter-phrastique et intra-phrastique dans un même tweet.

Code-switching inter-phrastique
Les signatures qui contiennent encore des séquences P (de ponctuation) regroupent les tweets dont le changement de langue est balisé avec une marque de ponctuation. Pour 526 tweets, le changement de langue se fait uniquement sur des ponctuations. Donc, le code-switching exclusivement inter-phrastique se produit dans 6% des cas.

Code-switching intra-phrastique
Il y a 73% des tweets bilingues dont le changement de langue survient exclusivement sur une marque de ponctuation (soit 6113 tweets). Ces tweets sont ceux dont les signatures de niveau 4 ne contiennent aucune séquence P.

Code-switching hybride
Plusieurs points de changement de langue dans un même tweet peuvent ne pas être marqués de la même manière. Par exemple, le premier point de changement de langue peut coïncider avec la ponctuation, mais pas le deuxième. Sur l'ensemble des tweets bilingues, 1760 d'entre eux se composent de changements de

langue sans marquage et de changements sur une marque de ponctuation. Cela signifie que 21% des tweets contiennent du code-switching intra-phrastique et inter-phrastique. La variété des points de changement dans un tweet avec un code-switching hybride implique au minimum deux points de changement, de ce fait, les langues de ces tweets sont forcément organisées par insertion.

Répartition des points de changement

La possibilité de code-switching hybride ne permet pas d'avoir une représentation fiable de la part de changement intra- ou inter-phrastique. De plus, plusieurs points de changement identiques peuvent survenir dans un même tweet. De ce fait, le calcul de la proportion des structures doit se faire non pas sur le nombre de tweets concernés, mais sur le nombre de points de changement. 17584 points de changement ont été comptabilisés pour l'ensemble des tweets avec code-switching. Sur ce total, 2687 d'entre eux coïncident avec une marque de ponctuation, et 14897 ne coïncident pas. Ainsi, 85% des changements de langue dans les tweets se font en dehors d'une marque de ponctuation.

D'après le corpus et l'étiquetage qui lui a été assigné, le code-switching intra-phrastique est majoritaire. Or, la proportion des structures du code-switching doit être nuancée. Les données non-standard (comme ici, issues de Twitter) peuvent ne pas refléter un usage conventionnel de la ponctuation (*cf 2.2 Caractéristiques des CMR*).

Pour conclure, le code-switching peut se catégoriser en fonction de sa structure et de son organisation. Pour cela, la catégorisation repose sur l'analyse des points de changement de langue. Le nombre de points de changement de langue dans un énoncé permet d'accéder au type d'alternance des langues. La position des points de changement (sur une ponctuation) indique la structure sur laquelle s'organise le code-switching. Le nombre et la position des points sont des informations qui peuvent être croisées afin d'affiner la catégorisation.

Le tableau 5 répertorie les proportions de chaque catégorie du code-switching. Il est indiqué, par exemple, que 66% des structures inter-phrastiques ont une alternance par juxtaposition. Sur l'ensemble des possibilités, le code-switching inter-phrastique par juxtaposition représente 4% des cas. Enfin, le code-switching intra-phrastique par insertion est celui qui est le plus représenté dans le corpus (53%), il occupe un peu plus de la moitié du code-switching en général. En revanche, le code-switching inter-phrastique par insertion est le moins représenté avec seulement 2% des tweets bilingues.

	Proportion des structures	**Type d'alternance**	**Proportion de l'alternance par structure**	**Proportion totale**
Intra-phrastique	73 %	Insertion	73 %	**53 %**
		Juxtaposition	27 %	**19 %**
Inter-phrastique	6 %	Insertion	34 %	**2 %**
		Juxtaposition	66 %	**4 %**
Hybride	21 %	Insertion	100 %	**21 %**

Tableau 5 : Répartition des différentes formes de code-switching

8. Résultats : analyse et évaluation

La méthodologie présentée précédemment a pour objectif primaire de détecter et de catégoriser le code-switching, pour ensuite, d'un point de vue plus large, accéder à certains phénomènes linguistiques découlant du code-switching.

Dans un premier temps, une évaluation de la méthodologie est proposée afin d'estimer une potentielle fiabilité des résultats obtenus. Ensuite, ces résultats seront utilisés pour vérifier la possibilité d'une étude ciblée sur un phénomène linguistique : le texte bilingue.

8.1. Evaluation de la détection et de la catégorisation du code-switching

L'évaluation du code-switching consiste à vérifier si le tweet contient effectivement du français et de l'anglais. Dans l'ensemble, cette tâche est peu compliquée quand on maîtrise le français et l'anglais. Cependant, certains cas sont particuliers, et ils peuvent être plus difficiles à juger.

Par exemple, il peut être parfois compliqué d'identifier les entités nommées comme telles, et non pas comme des parties du discours en français. Dans l'exemple (17), le contexte et la majuscule à Colombe permettent de le considérer comme une entité nommée, or, sans ces indices, il est plus compliqué de se décider quand l'entité nommée est construites à partir de formes de la langue reconnues.

> (17) La Colombe *is the ultimate dining experience . Must do when in Cape Town* #foodie #inheaven (TweetId : 784102060861992960)

De plus, certains tweets contiennent du code-switching en impliquant d'autres langues que juste le français et l'anglais. Par exemple, le tweet (18) contient du code-switching espagnol-anglais. La discrimination de l'espagnol par rapport au français se fait uniquement sur le verbe *atenta*. De ce fait, il y a effectivement du code-switching, mais il ne peut être comptabilisé comme un résultat satisfaisant, car les langues ne correspondent pas à celles attendues.

> (18) Frida atenta a la final de #handball . *Frida the #bulldog watching the handball game ….* https://t.co/T6TUqTlIXy (TweetId : 825757777050091520)

Enfin, certains tweets ne contiennent pas suffisamment de contenu pour pouvoir décider des langues impliquées (exemple (19)).

> (19) Bruhh 😨 (TweetId : 1162287097098600448)

8.1.1. Evaluation de la détection de code-switching

L'identification et la catégorisation du code-switching, d'après la méthode développée dans les points précédents, ont été évaluées.

100 tweets identifiés comme contenant du code-switching par la méthode utilisée ont été extraits. Tout d'abord, 10 tweets par signature de niveau 4 avec une fréquence supérieure à 50 ont été sélectionnés. Cela représente donc 8 signatures (EFE - EF - FE - EFEFE - FEFE - EFEF - EFEFEFE - FEF). Ensuite, 20 autres tweets identifiés bilingues ont été extraits indépendamment de leur signature. La sélection des tweets s'est faite de manière totalement aléatoire.

Dans un premier temps, l'évaluation a porté sur l'identification du code-switching. Les 100 tweets sélectionnés ont tous été identifiés comme contenant du code-switching. Ils ont tous été vérifiés manuellement. Il s'est avéré que seulement 38 d'entre eux contenant réellement du code-switching selon mon jugement.

Sur les 38 correctement identifiés, ce sont ensuite les signatures qui ont été évaluées. Finalement, 23 tweets sur les 100 initiaux ont été correctement identifiés et possèdent une signature qui correspond. De ce fait, sur les 38 tweets corrects, 61% d'entre eux sont illustrés par une signature fidèle. Les erreurs ont été analysées et commentées dans la partie *8.2. Analyse des erreurs*.

8.1.2. Evaluation de la catégorisation du code-switching

En plus de l'évaluation de la détection du code-switching sur 100 nouveaux tweets, le type de code-switching identifié a également été évalué. Le jeu de données à évaluer est composé de tweets catégorisés par juxtaposition, insertion, inter-phrastique, inter-phrastique et hybride. Il y a 20 tweets par catégorie, soit 100 tweets. Dans un premier temps, la vérification a porté sur la présence réelle de code-switching. La vérification du type de code-switching assigné a été réalisée dans un second temps. Sur les 100 tweets, 33 sont réellement composés de code-switching. Sur ces 33, 28 d'entre eux ont été correctement catégorisés. De ce fait, 85% des tweets correctement identifiés bilingues sont catégorisés de manière adéquate.

Sur l'ensemble des jeux de données utilisés pour l'évaluation (200 tweets), 35,5% des tweets contiennent réellement du code-switching et 25,5% ont été correctement catégorisés. Sur l'ensemble des tweets dont la détection du code-switching a été réussie, 72% correspondent à la catégorisation qui leur a été attribuée.

Les déséquilibres de performances entre les deux tâches indiquent que l'identification du code-switching est la partie la moins performante. La tâche de catégorisation est directement liée à l'étape antérieure d'identification. Quand les données sont correctement identifiées, la catégorisation montre de meilleures performances. Cela laisse à penser que les améliorations des performances doivent se concentrer sur la partie de l'identification. Pour pouvoir améliorer l'automatisation de la détection et de la catégorisation du code-switching, il faut comprendre d'où proviennent les erreurs.

8.2. Analyse des erreurs

Les tweets ont été examinés afin de recenser les différentes erreurs qui ont engendré l'identification, à tort, de code-switching. Les erreurs d'identification du code-switching des deux évaluations précédentes sont issues d'erreurs d'identification de langue. Ces erreurs d'étiquetage peuvent avoir été provoquées à différents niveaux du traitement. Dans un premier temps, les erreurs sont présentées en fonction de leurs origines, ensuite, elles sont quantifiées afin d'établir leur part de responsabilité dans la mauvaise identification de langue, et donc de code-switching.

8.2.1. Erreurs dues à la segmentation

La première étape pour l'identification de langue a été la segmentation des données. Cependant, certains points de segmentation ont créé des tokens indépendants qui existent peu ou pas.

Par exemple, *gotta* a été divisé en *got* et *ta*. Mais, *ta* est une marque de contraction de *got to*, qui est spécifique de l'oral. La forme *ta* est présent dans le lexique de l'anglais et du français, ses usages ont donc été étudiés pour pouvoir l'attribuer potentiellement à une langue en particulier. Dans 75% des cas, il est utilisé dans des contextes en français, il a été considéré comme du français. Son manque de représentativité en anglais s'explique par le fait que ce soit une forme principalement associée à l'oral. Ainsi, la segmentation

de *gotta* a entraîné l'identification d'un token en français. Cependant, la forme *gotta* est absente d'ENGLAFF, alors, cette erreur peut être résolue si la forme *gotta* n'est pas segmentée et si elle est ajoutée au lexique de l'anglais. Cela peut s'appliquer également sur d'autres formes qui observent le même phénomène comme *gonna* ou *wanna*.

Le même cas de figure se présente également pour *ain't* segmenté *ai* et *n't*. La partie portant la négation *n't* est correctement identifiée comme de l'anglais, car cette forme apparaît dans le lexique. En revanche, la partie *ai*, présente à l'origine dans les deux lexiques, a été attribuée au français, puisqu'elle correspond à 98% à des usages du français. Là aussi, la partie *ai* a été identifiée comme du français. Mais *ain't* apparait dans le lexique de l'anglais, de ce fait, la conservation de *ain't* en un seul token aurait évité des erreurs.

Ces deux exemples informent sur le fait que la segmentation ne doit pas générer de nouveaux tokens qui existent peu ou pas dans une langue, et qui donc ont peu de chances d'être intégrés dans un lexique.

8.2.2. Erreurs dues aux lexiques

Sur la segmentation effectuée, des lexiques ont été projetés pour déterminer la langue des tokens. Cependant, les lexiques peuvent contenir des erreurs qui entraînent de mauvaises attributions de langue.

La répartition des homographes dans les lexiques est à l'origine de certaines erreurs d'identification de langue. Certains homographes ont été attribués à une seule langue d'après les fréquences entre les deux langues, mais cela a entraîné un retrait de certains tokens d'un lexique qui génère des erreurs. Par exemple, le token *a*, qui s'utilise dans les deux langues, a été attribué à l'anglais, car dans 73% de ses usages il apparaît dans des contextes en anglais. La différence de proportion des usages de ce token s'explique par sa nature différente entre les deux langues. En anglais, *a* est un déterminant, tandis qu'en français, c'est une forme conjuguée du verbe *avoir*. La fréquence d'un déterminant a plus de chance d'être élevée que celle d'une forme conjuguée, même si cette dernière possède une très haute fréquence en français. Le problème est que cette différence a provoqué le retrait de *a* du lexique français, attribuant toutes les occurrences du token à l'anglais, même si ce token est également fréquent en français. Ainsi, pour l'exemple (20), le code-switching a bien été détecté, mais en tant que code-switching hybride avec insertion, alors que la lecture du tweet indique plutôt un code-switching inter-phrastique par juxtaposition.

> (20) " *Nobody loves it better than me .* " " *I'm the least racist person you know .* " " *Nobody's better at the military than I am .* " **Il y a une tendance** . (TweetId : 835154264116822016)

Les erreurs d'identification de langue engendrent également des erreurs de catégorisation du code-switching.

8.2.3. Erreurs dues aux spécificités des CMR

Pour certains cas, l'erreur d'identification de la langue provient de la nature des données. Les données produites dans le cadre des communications médiées par les réseaux peuvent être davantage composées d'orthographes déviantes que les écrits standards. Certaines formes peuvent avoir des orthographes déviantes qui correspondent à une forme reconnue dans une autre langue, et donc être attribuées à la mauvaise langue. L'exemple (21) illustre ce phénomène avec un tweet qui est produit entièrement en anglais, mais dont le token *ans* est considéré comme du français. En effet, *ans* est une forme reconnue du français et absente de l'anglais. Cependant, dans le contexte, il semble que *ans* est en fait le déterminant de l'anglais *and* avec une faute de frappe.

> (21) *Nothing quite like setting up a new website to bring up nostalgic memories of your first one . It was written in pure html back in the day , ans now I feel old .* (TweetId : 1082315521759084544)

Les données non-standards ne respectent pas toujours l'accentuation attendue pour certains tokens. Même si cela impacte peu la compréhension, cela peut contribuer à des erreurs d'identification de langue. Par exemple, le token *déjà* est présent dans le lexique du français et de l'anglais. En revanche, le token *deja* (équivalent de *déjà* sans l'accentuation) se trouve uniquement dans le lexique de l'anglais. Le lexique de l'anglais recense une version non accentuée du token car, l'anglais n'utilise pas les accents. Mais il intègre également la version avec accents, puisque c'est la forme originelle du token qui est un emprunt du français issu de l'expression figée « *déjà vu* ». Le token deja n'est pas dans le lexique français quant à lui. De cette manière, les occurrences de *deja* sont considérées comme de l'anglais, même quand elles sont dans des contextes français comme dans l'exemple (22).

(22) *Montreal got those Good Vibes & Great People tonight for the second edition* ! **J'ai deja hâte d'entendre le** ... https://t.co/0dfk9EXWzY (TweetId : 708056489102065664)

Un autre type d'erreur spécifique aux données provient de la présence d'abréviations. La limite de caractères ou l'aspect informel de ces communications favorisent l'utilisation des abréviations. Par exemple, dans l'exemple (23), *conf* est l'abréviation de *conference* en anglais. Cependant, ce token est identifié en français qui utilise cette même abréviation. Ce problème est également lié aux lexiques, car le lexique du français est davantage diversifié en forme que celui de l'anglais, beaucoup plus restreint.

(23) *Don't miss this #AI & it's fuel , #datascience / #bigdata conf in Toronto in June ! cc .* @gaessaki @DesjardinsLab (TweetId : 985858720474124288)

Certains tokens n'ont été attribués à aucune langue (inconnus). Le fait que ces tokens restent sans langue ne trouble pas la détection de code-switching. Cependant, cela impacte la précision de la tâche d'identification de langue. Une partie des tokens inconnus sont des tokens connus, mais avec des variations orthographiques (cf *6.4. Observation des tokens*). Ces variations, qui ne sont pas intégrées dans les lexiques, ne sont donc pas reconnues. Dans l'exemple (24), le token *shiiiiit* est sans conteste le token *shit* avec un allongement de la voyelle. Ce tweet reste tout de même bien identifié en anglais, mais il traduit tout de même un problème d'identification qui peut engendrer des mauvaises identifications et catégorisations du code-switching dans d'autres circonstances.

(24) *HOLY SHIIIIIT* (TweetId : 752981713014312964)

Les communications médiées par les réseaux ont la particularité de se trouver entre l'oral et l'écrit (*cf 2.2. Caractéristiques des CMR*). De ce fait, certains tweets ont recours aux onomatopées afin de reproduire certains sons qui accompagneraient le message dans la modalité orale. Certaines onomatopées sont associées principalement à une langue (comme *miaou* en français vs *meow* en anglais), tandis que d'autres, non. L'onomatopée *haha* est utilisé dans les tweets pour simuler un rire franc joint au message. Sans contexte, cette onomatopée n'est pas attribuable à une langue. Or, elle est intégrée dans le lexique de l'anglais, mais pas du français. Ainsi, dans l'exemple (25) et (26), *haha* est étiqueté en anglais. Cependant, dans l'exemple (25) *haha* est utilisé dans une production entièrement en anglais tandis que dans l'exemple (26) il est utilisé dans une production en français. Du fait de l'étiquetage, l'exemple (26) est donc considéré, à tort, comme du code-switching.

(25) *I thought they were good friends haha* (TweetId : 1187452360701833216)

(26) **Haha grave** ^^ (TweetId : 1051258812357890049)

8.2.4. Erreurs dues aux entités nommées

D'autres erreurs d'attribution de langue sont dues à la présence d'entités nommées. Ces entités nommées peuvent se composer d'un ou plusieurs tokens qui sont des formes reconnues d'une langue. L'exemple (27) illustre ce problème, le locuteur s'exprime en anglais, mais en faisant référence à des lieux précis en français. L'ensemble des tokens de « *Salon Urbain - Place des Arts* » ont été identifié comme du français.

Techniquement, l'identification de langue est correcte, mais le contexte ne permet pas de considérer ce tweet comme bilingue.

> (27) *Wedding story from the Salon Urbain – Place des Arts* https://t.co/PMFqgyi5S1
> (TweetId : 753623002462363648)

Ce problème est d'autant plus dérangeant dans le type de données qui est utilisé. Les tweets ont été produits par des locuteurs de Montréal, une ville majoritairement francophone qui favorise la présence d'entités nommées construites depuis le français. Les entités nommées en anglais dans une production en français ne sont pas considérées comme du code-switching, et inversement, car elles ne reflètent pas un choix langagier fait de la part du locuteur.

8.2.5. Quantification des erreurs

Sur les 200 tweets vérifiés manuellement lors de l'étape précédente d'évaluation, 129 ne contiennent pas de code-switching. Sur les 129 tweets, 161 erreurs ont été identifiées. Il y a plus d'erreurs que de tweets, car certains tweets en contiennent plusieurs.

Pour la segmentation, seulement 4 erreurs ont été comptabilisées, elles représentent alors, 2% des erreurs globales (figure 5). Ensuite, 12 erreurs ont été attribuées aux spécificités de langage des CMR, soit 7%. En ce qui concerne les erreurs d'entités nommées, elles correspondent à 23% du total des erreurs, soit 37 erreurs. Enfin, les erreurs dues aux lexiques sont celles qui ont été le plus identifiées. Il y en a 108 au total, et elles représentent donc 67% des erreurs d'identification en général (figure 5).

Les lexiques sont donc responsables de presque trois quarts des erreurs recensées. Ils constituent alors un point majeur d'amélioration qui permettrait d'accroître la fiabilité de la méthodologie. Le traitement des entités nommées, au même titre que tous les autres tokens dans l'attribution de langue, a également une part importante de responsabilité dans les erreurs d'identification. En effet, les entités nommées constituent près d'un quart des erreurs. En ce qui concerne les erreurs dues à la nature langagière des données, leur part paraît faible en comparaison aux deux types d'erreurs mentionnées juste avant. Cependant, elles sont quand même responsables de quasiment un dixième des erreurs. Enfin, les erreurs de segmentation sont les moins représentées.

L'ensemble des erreurs identifiées doit être pris en compte pour de futures améliorations, avec une priorité sur celles de lexiques et d'entités nommées qui sont responsables de plus de 90% des erreurs.

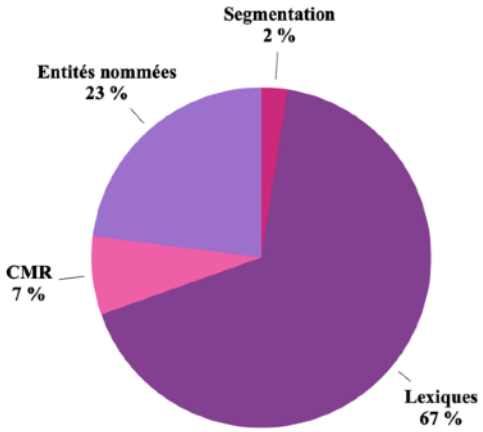

Figure 5 : Répartition des différentes erreurs d'identification

8.3. Pistes d'amélioration

L'identification de langue est une étape déterminante pour les performances de détection et de catégorisation du code-switching. Différents types d'erreurs entraînent des erreurs d'identification de la langue. Certaines de ces erreurs peuvent être évitées en adaptant la méthodologie utilisée.

Dans un premier temps, certaines erreurs dues aux lexiques viennent du fait qu'ils contiennent à la fois trop d'informations et en même temps, pas assez. Les étapes d'ajustement des lexiques ont pour objectif d'enlever les formes indésirables et d'ajouter les formes manquantes. Pour cela, la principale méthode utilisée a été le traitement des homographes qui a généré de nouvelles erreurs.

Le traitement des tokens bilingues peut être amélioré en incluant le contexte dans leur attribution de langue. Jusqu'à présent, la langue des tokens bilingues a été attribuée en étudiant les fréquences dans des corpus généraux du français et de l'anglais. Ainsi, des tokens ont été cantonnés à une seule langue en dépit d'usages avérés dans l'autre langue (exemple de *a* dans *8.2.2 Erreurs dues aux lexiques*). Ce genre de traitement semble être trop clivant et brutal pour être efficace. Plusieurs possibilités d'adaptations sont possibles pour permettre une catégorisation des tokens bilingues plus fine. Une première possibilité est de modifier le seuil de catégorisation des homographes. Actuellement, un token qui est utilisé dans moins de 33% des cas dans une langue est considéré comme appartenant à l'autre langue. Diminuer ce seuil augmenterait le nombre de tokens qui resteraient bilingues et éviterait certaines erreurs d'identification de langue. Une autre possibilité consisterait à établir un seuil de fréquence relative au-dessus du quel il n'y aurait pas d'attribution à une langue uniquement. De cette façon, les usages seraient pris en compte que dans une seule langue, sans influence des usages d'une autre.

Ensuite, l'identification de la langue a également été impactée par la présence d'entités nommées. Ce problème peut être résolu en proposant dans un premier temps une identification des entités nommées pour qu'elles ne soient pas intégrées dans la partie de l'identification de langue. L'outil SpaCy, utilisé pour la segmentation, propose aussi une reconnaissance des entités nommées, mais comme pour la segmentation, basée sur le modèle d'une seule langue. Sinon, sans utilisation d'outils, l'interférence des entités nommées peut être diminuée en n'identifiant pas la langue des tokens qui commencent par une majuscule et qui ne se situent pas en début de phrase.

La nature des données ne peut être modifiée, cependant, il est possible de proposer une normalisation de ces données afin de les rendre plus conventionnelles. Par exemple, la succession de caractères identiques peut être remplacée par une seule lettre. Il faut rester prudent avec ce genre de transformation pour ne pas altérer les données et engendrer davantage d'erreurs. Par exemple, cette transformation peut s'appliquer qu'à partir de trois caractères (hors ponctuation) identiques à la suite. Ensuite, une correction orthographique peut s'appliquer sur les tokens (après segmentation) qui sont des formes non reconnues. Même si la correction ne permet pas de régulariser toutes les formes, elle peut permettre de limiter le nombre de tokens inconnus.

Enfin, la segmentation peut être adaptée en s'effectuant que dans le cas où des tokens existants sont générés. De cette manière, la segmentation ne s'applique pas si le résultat de cette dernière provoque des tokens qui n'existent pas selon les grammaires des langues. De cette façon, les morphèmes liés ne sont pas considérés comme un token.

8.4. Utilisation des résultats pour la détection d'un phénomène en particulier : le texte bilingue

Le texte bilingue est une production avec code-switching dont la fonction est de proposer un même énoncé traduit entre deux langues, ou plus (*cf 1.3.3. Phénomènes linguistiques spécifiques au code-switching*).

L'objectif de cette partie est d'utiliser les annotations (signatures et étiquetages) faites sur les tweets pour vérifier si elles permettent d'accéder à un phénomène ciblé, mais également de vérifier si ce phénomène est annoté comme attendu.

Il est attendu que le texte bilingue corresponde à une structure inter-phrastique par juxtaposition du code-switching. Les signatures de niveau 3 permettent d'accéder à l'ensemble des informations concernant la structure et l'alternance du code-switching. Ainsi, il est attendu des signatures de type EPF ou FPE qu'elles renvoient en partie à des tweets avec du texte bilingue.

Une stratégie complémentaire pour accéder potentiellement à du texte bilingue est de s'appuyer sur la proportion de langue des tweets (*cf. 7.2.2. Proportion des langues par tweet*). Il est attendu d'un tweet avec une même information dans les deux langues, qu'il contienne à peu près la même proportion d'anglais et de français. De ce fait, le taux d'anglais se situerait autour de 0,5.

Désormais, il reste à vérifier que le croisement des signatures avec le taux d'anglais permet d'accéder aux résultats attendus, à savoir des textes bilingues.

Il y a 66 tweets qui ont à la fois une structure inter-phrastique par juxtaposition et une proportion d'anglais comprise entre 0,4 et 0,6. Après vérification, 53% d'entre eux contiennent effectivement du code-switching et 30% sont des textes bilingues. Lorsque la détection du code-switching est correcte, 57% des tweets répondant au schéma sont effectivement des textes bilingues.

Dans un premier temps, la précision des tweets avec code-switching est plus élevé que par rapport à la précision général (35,5%). Cette augmentation de la précision est certainement due à la prise en compte de la proportion des langues dans la récupération des tweets. Cela limite ainsi les tweets contenant uniquement un ou deux tokens identifiés dans une autre langue et qui ne le sont pas réellement.

Les exemples suivants ont été récupérés en croisant la structure et la part d'anglais pour correspondre aux pré-requis du texte bilingue. L'exemple (28) correspond effectivement à un texte bilingue. La première partie du tweet, en français, est traduite dans la seconde partie en anglais. Les deux parties sont de tailles équivalentes (taux d'anglais de 0,5) et l'unique point de changement de langue est marqué par une ponctuation.

> (28) **Nouveau en magasin pour vois aider dans vos projets** . *New at the store to help you with your projexts* . https://t.co/2pLTLBcgoJ (TweetId : 777148468519378944)

L'exemple (29), quant à lui, correspond au schéma du texte bilingue, mais n'en est pas un. Ce résultat n'est pas pour autant dérangeant, car même si le schéma permet d'identifier le texte bilingue, il n'est pas réservé à ce phénomène.

> (29) **Au gratin** , *of course* . (TweetId : 917239931780640768)

En revanche, l'exemple (30), est un résultat indésirable. Il ne contient pas de code-switching mais a été identifié ainsi à cause de la présence d'entités nommées (un film) en anglais. Ce problème renvoie aux erreurs déjà identifiées dans la partie *8.2.3. Erreurs dues aux spécificités des CMR*.

> (30) Jurassic World Fallen Kingdom *:* **rutilant , mais pas renversant** https://t.co/bdOFHYQBNV (TweetId : 1004292900061175808)

Dans l'autre sens, des tweets correspondants à du texte bilingue n'ont pas été récupérés par le schéma. Ces tweets ont été identifiés par hasard dans de précédentes étapes comme le corpus de développement ou l'évaluation des résultats.

Le tweet de l'exemple (31) n'a pas été ramené par le schéma à cause de son taux d'anglais de 0,71. Ce taux s'explique par l'attribution de l'étiquette bilingue pour *parties* et *d'* et de l'étiquette inconnu pour *aujourd'hui*. Les étiquettes B et I ne sont pas prises en compte dans la comptabilisation des tokens de l'anglais ou du français. De ce fait, un déséquilibre inexistant entre les deux langues est renvoyé par le taux de code-switching.

> (31) **Voici votre classement suite aux parties d'aujourd'hui** / *Here are your updated standings following today's games* (TweetId : 1103121459193499648)

L'exemple (32) n'est pas récupéré non plus par le schéma du texte bilingue. Son taux d'anglais (0,6) est compris dans l'intervalle défini, mais c'est la structure du code-switching qui ne correspond pas. Ce tweet est considéré avec un code-switching intra-phrastique par juxtaposition. Or, cette catégorisation n'est pas

correcte et elle a été induite par le mauvais étiquetage des points de suspension en anglais. Les points de suspension ont pourtant été intégrés dans la liste des marques de ponctuation à étiqueter P. Cependant, les points de suspension sont présents dans le lexique de l'anglais, et malheureusement, la présence dans le lexique est prioritaire sur la présence dans la liste des ponctuations au moment de l'étiquetage. Cette défaillance a été constatée au moment de l'étude du tweet de l'exemple (32).

> (32) **C'est ce qui se passe quand tu as trop attendu pour acheter tes billets et que tu manques WordCamp Montréal** ... #wcmtl - - - - - - - - *That's what happens when you wait too long to buy your tickets and you miss WordCamp Montreal 2019* ... #WordCamp (TweetId : 1156240032383954944)

La précision de récupération de tweets avec du texte bilingue est un peu plus élevée que celle des tweets avec code-switching en général. Cependant, cette précision est soumise aux mêmes problématiques soulevées qui impactent l'identification de langue, et donc par extension la détection et la catégorisation du code-switching.

Conclusion

Bilan

Tout d'abord, l'état de l'art sur le code-switching permet de comprendre ce que c'est réellement et quelles sont les conditions qui permettent de le caractériser. De plus, les différents types de code-switching ont été présentés, que ce soit en fonction des structures ou des fonctions. Le code-switching est donc un phénomène linguistique qui implique la cohabitation de deux langues au sein d'un même énoncé. Pour qu'il y ait du code-switching, il faut que les différentes parties du message respectent la grammaire, la syntaxe, la morphologie et la phonologie de la langue dont elles sont issues. Les entités nommées d'une langue qui sont intégrées dans un message dans une autre langue ne sont pas considérées comme du code-switching car elles ne reflètent pas un choix langagier.

Le code-switching se catégorise en fonction de la forme qu'il prend ou des fonctions qu'on lui attribue. En ce qui concerne la forme, les points de changement de langue sont les informations qui permettent de dissocier différents types de code-switching. Le nombre de points de changement de langue permet d'obtenir l'organisation de l'alternance des langues : par juxtaposition ou par insertion. La position syntaxique des points de changement de langue permet d'accéder à la structure du code-switching : intra-phrastique ou inter-phrastique.

Ensuite, la deuxième point de ce mémoire a servi à recenser les différentes caractéristiques linguistiques des communications médiées sur les réseaux, et en particulier sur le réseau social Twitter. Ce type de données est propice à l'insertion de contenu avec du code-switching par son aspect informel et spontané et par sa situation ambiguë entre l'oral et l'écrit. Il est important de connaître les spécificités de ce genre de communication pour pouvoir ensuite proposer un traitement des données le plus adapté possible.

Le traitement automatique des langues doit s'adapter aux types de données qui doivent être traitées. Dans le cadre de ce travail, le bilinguisme des données ainsi que leur caractère non standard motivent des adaptations. Le traitement doit prendre en compte les spécificités de deux langues et non pas se baser uniquement sur un seul modèle de langue. De plus, il faut que les spécificités linguistiques et technologiques des données soient prise en compte afin de ne pas dégrader les résultats du traitement.

Enfin, après l'étude théorique du sujet, les connaissances ont été mises en application à travers la proposition d'un traitement automatique visant à identifier les phénomènes linguistiques du code-switching.

L'objectif de l'automatisation de la détection et de la catégorisation du code-switching (avec l'identification de langue) est de proposer des manières d'y parvenir pour ensuite étudier les failles de la méthode pour les résoudre et, par la suite, améliorer le traitement.

Pour pouvoir étudier des phénomènes linguistiques propres au code-switching dans les tweets, il faut dans un premier temps détecter et catégoriser le code-switching. La détection est possible suite à une identification de langue précise sur les tokens qui composent l'énoncé.

Le code-switching intra-phrastique par insertion est celui qui a été plus identifié dans les tweets. Cependant, le taux de précision de détection du code-switching est assez faible, de ce fait la représentation des structures intra-phrastiques par juxtaposition doit être nuancée.

La fiabilité de la détection du code-switching est augmentée lorsque que la proportion des langues dans les tweets est prise en compte dans la récupération des données. Ainsi, pour accéder au mieux aux tweets avec du code-switching, il faut définir un seuil maximal de proportion d'anglais lors de la récupération des tweets, afin d'éviter le plus possible la récupération de tweets contenant des erreurs d'identification.

Ce travail permet également de souligner l'importance du contexte des tokens pour optimiser la fiabilité de la détection de langue. En effet, une simple identification au niveau du token, sans contexte, ne permet pas de lever les ambiguïtés possibles.

Limites

Certaines limites ont été constatées pendant le déroulement de ce travail. Dans un premier temps, le déséquilibre entre le lexique du français et de l'anglais. Le lexique du français contient presque le double de formes que celui de l'anglais, mais le corpus contient moins de 2% de formes reconnues du français. Avec la méthodologie adoptée, la répartition des langues ne peut être connue à l'avance et les lexiques ne peuvent être proportionnels à la répartition. Cependant, il semble important que les lexiques soient équilibrés en nombre de formes et en diversités de celle-ci. Par exemple, le lexique du français intègre des abréviations contrairement au lexique l'anglais qui parait en contenir moins.

De plus, le recours aux lexiques n'est pas ce qu'il y a de plus adapté pour des données non-standards qui génèrent des formes inédites, qui ont donc peu de chance d'être intégrées dans un lexique.

Les performances peu élevées de l'identification de langue compliquent l'étude de certains phénomènes comme le code-switching des mots isolés ou l'identification des emprunts. En effet, la récupération du code-switching composé d'un seul mot risque davantage de ramener des erreurs d'étiquetage.

L'intégration du taux d'anglais dans la tâche de récupération des tweets bilingues augmente la précision. En revanche, les phénomènes d'emprunts ou de mots isolés du code-switching peuvent être facilement manqués avec une contrainte sur le taux.

Perspectives

Ce travail offre des perspectives d'améliorations et de continuité. Tout d'abord, du point de vue des améliorations, l'identification de langue doit être améliorée avec principalement une prise en compte des entités nommées et un recours à des lexiques plus équilibrés. Ce qui résulte également de ce travail est la difficulté de faire de l'identification de langue sans prendre un minimum en compte le contexte. De ce fait, l'identification de langue sur des données courtes doit effectivement se réaliser à un niveau inférieur à l'énoncé, comme le token, mais elle ne doit pas écarter le contexte des éléments qui peut être décisif dans le bon choix de langue (notamment dans les cas d'ambiguïté).

De plus, lors de l'étude du texte bilingue, le taux de code-switching s'est avéré être une information utile. Cette information mériterait d'être davantage intégrée et utilisée dans la détection et l'analyse du code-switching. Par exemple, il pourrait être intéressant de calculer la moyenne des proportions d'anglais pour chaque structure afin de savoir si, dans le cas du code-switching français-anglais, la part de code-switching d'une langue est liée à la structure utilisée.

Plusieurs phénomènes provenant du code-switching ont été identifiés dans la partie *I. Apports théoriques*, comme le tag-switching ou l'emprunt. A l'image du texte bilingue, il serait intéressant d'étudier les potentiels schémas favorisant leur identification.

Enfin, ce travail a abordé principalement l'aspect fonctionnel et organisationnel des messages. Néanmoins, il reste à étudier les relations entre l'aspect fonctionnel et l'aspect structurel.

Bibliographie

- Barman U., Das A., Wagner, J. & Foster J. (2014). Code Mixing : A Challenge for Language Identification in the Language of Social Media. *Proceedings of The First Workshop on Computational Approaches to Code Switching*, 13-23. Doha, Qatar.
- Begum R., Bali K., Choudhury M., Rudra K. & Ganguly N. (2016). Functions of Code-Switching in tweets : An annotation scheme and some initial experiments. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1644-1650. Portorož, Slovénie.
- Biber D. (1988). *Variation across Speech and Writing*. Cambridge University Press.
- Blank G. (2016). The Digital Divide Among Twitter Users and Its Implications for Social Research. *Social Science Computer Review*, 35(6), 679-697.
- Boztepe E. (2003). Issues in Code-Switching: Competing Theories and Models. *Studies in Applied Linguistics & TESOL,* 3(2), 1-27.
- Brasart C. (2011). Code-switching, co-texte, contexte : une analyse du jeu de langue dans les conversations bilingues. *Etudes de stylistique anglaise*, 3, 107-122.
- Bullock B.E. & Toribio A.J. (2012). Themes in the study of code-switching. Dans Bullock B.E. & Toribio A.J. (eds), *The Cambridge Handbook of Linguistic Code-switching,* 1-17. Cambridge University Press.
- Çetinoğlu O., Schulz S. & Vu N.T. (2016). Challenges of Computational Processing of Code-Switching. *Proceedings of the Second Workshop on Computational Approaches to Code Switching,* 1-11. Austin, Texas, USA.
- Chanier T., Poudat C., Sagot B., Antoniadis G., Wigham C., Hriba L., Longhi J. & Seddah D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for language technology and computational linguistics*, 29 (2), 1-30.
- Chanier T. (2017). Saisir la parole du citoyen / usager / apprenant en interaction sur les réseaux. Dans Wigham C. & Ledegen G. (eds.), *Corpus de communication médiée par les réseaux,* 211-222. L'Harmattan.
- Chomsky N. (1971). *Aspects de la théorie syntaxique*. Paris, France: Seuil.
- Crystal D. (2001). *Language and the Internet*. Cambridge University Press.
- Das A. & Gambäck B. (2013). Code-Mixing in Social Media Text. *TAL*, 54(3), 41-64.
- Dorlejin M. & Nortier J. (2012). Code-switching and the internet. Dans Bullock B.E. & Toribio A.J. (eds), *The Cambridge Handbook of Linguistic Code-switching,* 127-141. Cambridge University Press.

- Dridi H. E. & Lapalme G. (2013). Détection d'évènements à partir de Twitter. *TAL*, 53(3), 17-39.
- Farzindar A. & Roche M. (2013). Les défis de l'analyse des réseaux sociaux pour le traitement automatique des langues. *TAL*, 54(3), 7-16.
- Fausto S. & Aventurier P. (2015). Scientific literature on Twitter as subject research : preliminary findings based on bibliometric analysis. *Twitter for Research 2015*, 1. Lyon, France.
- Grosjean F. (1982). *Life with two languages: An introduction to bilingualism*. Cambridge, MA: Harvard University Press.
- Hathout N., Sajous F. & Calderone B. (2014). GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1007-1012. Reykjavik, Island.
- Heiden S., Magué J.-P. & Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data,* 1-12. Rome, Italie.
- Kevers L. (2021). L'identification de langue, un outil au service du corse et de l'évaluation des ressources linguistiques. *TAL*, 62(3), 13-37.
- Lacaze G. (2021). Renouvellement des formes langagières dans la communication sociale sur Twitter. Dans Dufiet J.P. & Jullion M.C. (eds.), *Les nouveaux langages au tournant du XXI siècle,* 107-128. LED, Edizioni Universitarie di Lettere Economia Diritto.
- Lance D. (1975). Spanish-English code-switching. Dans Hernandez-Chavez E. et al. (eds), *El lenguaje de los chicanos*. Arlington: Center for Applied Linguistics.
- Lipski J. (1978). Code-switching and the problem of bilingual competence. Dans Paradis M. (ed), *Aspects of bilingualism*, 250-264. Hornbeam.
- Lui M. & Baldwin T. (2012). langid. py: An off-the- shelf language identification tool. *Proceedings of the ACL 2012 system demonstrations*, 25-30. Jeju Island, Corée du Sud.
- Lui M. & Baldwin T. (2014). Accurate Language Identification of Twitter Messages. *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM) 2014*, 17-25, Gothenburg, Suède.
- Lynn T. & Scannell K. (2019). Code-switching in Irish tweets : A preliminary analysis. *Third Celtic Language Technology Workshop*, 32-40. Dublin, Irlande.
- Magué J. P., Rossi-Gensane N. & Halté P. (2020). De la segmentation dans les tweets : signes de ponctuation, connecteurs, émoticônes et émojis. *Corpus*, 20.
- Miletic F., Przewozny-Desriaux A. & Tanguy L. (2020). Collecting Tweets to Investigate Regional Variation in Canadian *English*. *12th Conference on Language Resources and Evaluation* (LREC 2020), 6255-6264. Marseille, France.

- Millour A. (2020). *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées*. Thèse de doctorat, Informatique et langage, Sorbonne Université.
- Moreau M.L. (1997), *Sociolinguistique, Concepts de base*. Bruxelles: Mardaga.
- Muysken P. (2000). *Bilingual Speech, a typology of code-mixing*. Cambridge : Cambridge University Press.
- Nguyen D. & Doğruöz A.S. (2013). Word Level Language Identification in Online Multilingual Communication. P*roceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 857–862, Seattle, Washington, USA.
- Paveau M. A. (2013). Genre de discours et technologie discursive. *Pratiques*, 157-158, 7-30.
- Pedraza P. (1978). *Ethnographic observations of language use in El Barrio*. Unpublished ms.
- Poplack S. (1980). Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL : toward a typology of code-switching. *Linguistics*, 18(7-8), 581-618.
- Poudat C., Wigham C. R. & Liégeois L. (2020). Les corpus de la communication médiée par les réseaux : une introduction. *Corpus*, 20, 1-8.
- Sajous F., Calderone B. & Hathout N. (2020). ENGLAWI: From Human to Machine-Readable Wiktionary. *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, 3016-3026. Marseille, France.
- Sankoff D. & Poplack S. (1981). A formal grammar for code-switching. *Paper in Linguistics*, 14(1), 3-45.
- Solorio T. & Liu Y. (2008). Learning to predict code-switching points. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 973-981. Honolulu, Hawaii.
- Wigham C. R. & Poudat C. (2020). Corpus complexes et standards : un retour sur le projet CoMeRe. *Corpus*, 20, 1-18.
- Yao M. Z. & Ling R. (2020). "What Is Computer-Mediated Communication ? "—An Introduction to the Special Issue. Journal of Computer-Mediated Communication, 25(1), 4-8.

Annexes

Annexe 1 : Guide de segmentation

La segmentation consiste à isoler les tokens indépendants dans un texte. Les tokens sont tout ce qui compose un texte : les mots et les ponctuations. Pour cela, il faut identifier les frontières entre les tokens. Les espaces sont considérés comme des frontières. Les marques de ponctuations sont considérées comme des tokens à eux seul dans la plupart des cas. Les liens URL, les hashtags et les mentions ne sont pas à segmenter. En revanche, l'apostrophe et le tiret sont des tokens que dans certains cas, et ils peuvent parfois être considérés en même temps comme des frontières.

Si le contenu est en anglais :

Pour l'apostrophe :
La segmentation se fait avant l'apostrophe et l'apostrophe est intégrée au token qui suit :
- Harry's → | Harry | ' |
- I'll → | I | 'll |

Quand l'apostrophe est la marque d'une contraction à l'intérieur d'un même token, il n'y a pas de segmentation :
- n't → | n't |

Pour le tiret :
La segmentation se fait de part et d'autre du tiret sauf si la partie antérieure au tiret correspond à un affixe commun qui n'a pas de statut indépendant comme dans *e-mail* ou *co-ordinated*.

Cas de la négation :
Une segmentation doit être effectuée afin de séparer la négation du verbe :
- didn't → | did | n't |

Si le contenu est en français :

Pour l'apostrophe :
Quand les deux éléments de part et d'autre de l'apostrophe ne peuvent exister sans l'autre, il n'y a pas de segmentation :
- aujourd'hui → | aujourd'hui |

Sinon, il y a segmentation après l'apostrophe et l'apostrophe est intégrée au token antérieur :
- l'école → | l' | école |

Pour le tiret :
Les mots composés avec un tiret ne sont pas segmentés :
- sous-marin → | sous-marin |

Sinon, il y a segmentation avant le tiret et le tiret est intégré au token qui suit :
- vient-il → | vient | -il |

Annexe 2 : Ajustement de la segmentation

Occurrences des séquences « .’ » et « .*qu’ » COCA :

- ***d’*** :
  - 23 occurrences
  - Se retrouve dans des contextes en français, ou en abréviation de *do*, suivie d’un espace.
  - Segmentation après l’apostrophe possible

- ***ç’*** :
  - 0 occurrences
  - Segmentation après l’apostrophe possible

- ***l’*** :
  - 51 occurrences
  - En contexte : remplace un *i* majuscule dans des contractions de l’auxiliaire précédées du pronom personnel *I*.
  - Segmentation après l’apostrophe possible

- ***s’*** :
  - 80 occurrences
  - En contexte : contraction de *is* immédiatement suivie de l’apostrophe en tant que marque de citation.
  - Segmentation après l’apostrophe possible

- ***c’*** :
  - 3 occurrences
  - En contexte : marqueur du possessif pour la séquence « *vitamin c’s* »
  - Segmentation après l’apostrophe possible

- ***m’*** :
  - 2 occurrences
  - En contexte : contraction de am suivi de l’apostrophe de contraction du mot suivant
  - Segmentation après l’apostrophe possible

- ***n’*** :
  - 32 occurrences
  - En contexte : appartient à l’expression figée « *rock n’ roll* », ou contraction de *and*. Dans les deux cas, l’apostrophe est suivi d’un espace.
  - Segmentation après l’apostrophe possible

- ***t’*** :
  - 3 occurrences
  - En contexte : possiblement une contraction de *to*, ou espace imprévu avant le t dans des séquences « it’s → i t’s »
  - Segmentation après l’apostrophe possible

- ***j’*** :
  - 0 occurrence
  - Segmentation après l’apostrophe possible

- ***qu’*** :
  - 0 occurrence
  - Segmentation après l’apostrophe possible

Occurrences des séquences *-elle*, *-elles*, *-il*, *-ils*, *-je*, *-la*, *-le*, *-les*, *-leur*, *-lui*, *-moi*, *-m'*, *-nous*, *-on*, *-t*, *-toi*, *-tu*, *-vous*, *-y*, *-même*, *-mêmes*, *-là*, *-ci*, *-ce* et *-en* dans COCA :

- ***-elle, -elles, -il, -ils, -je, -la, -les, -leur, -lui, -moi, -nous, -t, -toi, -tu, -vous, -y, -même, -mêmes, -là, -ci, -ce* et *-en*** *:*
  - 0 occurrence
- ***-m*** :
  - 3 occurrences
  - En contexte : uniquement en système de notation médical dans les textes académiques
- ***-le*** :
  - 1 occurrence
  - En contexte : tiret de citation suivi d'un segment en français
- ***-on*** :
  - 20 occurrences
  - En contexte : préposition « *on* » (EN) à l'initial d'une incise

Annexe 3 : Extrait du corpus étiqueté

User: 1037531		**Score: 1.0**	**Tweet: 1007975472943321089**
Yup	E		
.	P		
@QueerEye	M		
is	E		
my	E		
ASMR	E		
.	P		
All	E		
the	E		
feels	E		
!	P		
User: 992197110064930820		**Score: 0.77**	**Tweet: 1087821140058427392**
Cassie	B		
she	E		
have	E		
something	E		
je	F		
ne	F		
sais	F		
quoi	F		
!	P		
!	P		
!	P		
User: 27177826		**Score: 0.0**	**Tweet: 877871269030514688**
Cogir	I		
et	F		
DevMcGill	I		
se	F		
marient	F		
https://t.co/2iJDcsQQ1E	U		
#immobilier	H		

Annexe 4 : Composition du corpus DIV-706

Genres textuels	Sources
Articles de presse	98 articles d'information tirés des sites de Canal + et de Libération
Articles scientifiques	13 articles scientifiques issus des actes d'une conférence de linguistique
Articles Wikipedia	12 articles de l'encyclopédie en ligne Wikipedia
Compte-rendus médicaux	638 compte-rendus d'un service de réanimation chirurgicale
Critiques de films	96 critiques de films
Discours politiques	24 discours politiques prononcés par des ministres et des présidents en exercice
Discussions Wikipedia	1231 discussions sur le forum Wikipedia
Entretiens	Entretiens réalisés dans le cadre d'un projet de recherche (projet PFC)
Exposés	78 exposés sollicités dans le cadre d'un projet de recherche (projet PFC) : le sujet raconte son parcours biographique
Littérature de jeunesse	17 textes issus de la littérature de jeunesse
Profils de coach surfers	315 profils
Résumés de films	131 résumés de films (sites Canal + et Libération)
Sous-titres de films	16 épisodes transcrits de la série télé Desperate Housewives
Textes réglementaires	14 réglements intérieurs de collèges, lycées, associations

Tableau 6 : Sources des données par genre textuel pour le corpus DIV-706. Réalisé par M^me^ Fabre (échange personnel)

Annexe 5 : Table de fréquence des signatures de niveau 4

Signature	Fréquence
E	41200
EFE	5044
EF	1145
FE	835
EFEFE	675
FEFE	195
F	190
EFEF	188
EFEFEFE	112
FEF	66
EFEFEF	36
FEFEFE	36
EFEFEFEFE	23
FEFEF	16
FEFEFEFE	7
EFEFEFEFEFE	5
EFEFEFEF	4
FEFEFEFEFE	2
FEFEFEF	2
FEFEFEFEFEFE	2
EFEFEFEFEF	1
EFEFEFEFEFEFEFE	1
EFEFEFEFEFEF	1
EFEFEFEFEFEFE	1
EFEFEFEFEFEFEFEFE	1
FEFEFEFEFEFEFEF	1

Tableau 7 : Table des fréquences des signatures de niveau 4 dans le corpus

UNIVERSITÉ TOULOUSE Jean Jaurès

SCIENCES DU LANGAGE

Déclaration sur l'honneur de non-plagiat

(à joindre au mémoire à la fin du document)

Je soussigné.e,

Nom, Prénom : BLIVET Andréa
Régulièrement inscrit.e à l'Université de Toulouse II Jean Jaurès
N° étudiant : 21808046

Année universitaire : 2021 - 2022

certifie que le document joint à la présente déclaration est un travail original, que je n'ai ni recopié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations intégrales sont signalées entre guillemets.

Fait à : Toulouse

Le : 01.06.2022

Signature :