Faculté des **Langues**

Université de Strasbourg

2017-2019

Étude de la vitalité des mots du français régional

Estelle ROBERT

MASTER Technologies des langues

Projet encadré par :
Delphine Bernhard et Julie Glikman

N° étudiant : **21708303**

Table des matières

Introduction 5

1. État de l'art 6

1.1 Contexte de l'étude 6

1.2 Définition des notions de variation linguistique 9

1.2.2 Origine des régionalismes et dialectes 9

1.2.3 Types de régionalismes 10

1.2.4 Variation interne et externe 10

1.2.5 Bon usage et norme 12

1.2.6 Vitalité 14

1.2.7 Les fondements de la linguistique variationniste 15

1.3 Recueils de régionalismes 16

1.3.1 Atlas linguistique 16

1.3.2 Dictionnaires de régionalismes 17

1.4 Comparaison entre la presse et Twitter 20

1.5 Étude de la variation régionale sur Twitter 21

1.5.1 Comment constituer un corpus sur Twitter ? 21

1.5.2 Les travaux de Jack Grieve 22

1.5.3 Les régionalismes aux États-Unis 24

1.5.4 Inconvénients de Twitter 27

1.5.5 Études similaires 28

1.6 Prise en compte des entités nommées et du métadiscours 29

1.6.1 Cas du métadiscours 29

1.6.2 Cas des entités nommées 30

1.7 Conclusion 31

2. Données et méthode 32

2.1 Introduction à la classification automatique 32

2.2 Élaboration du corpus 33

2.2.1 Sources de régionalismes 33

2.2.2 Critères de sélection des régionalismes 33

2.2.3 Collecte de tweets 35

2.2.4 Annotation 36

2.2.5 Régionalismes retenus pour l'apprentissage 38

2.3 Pré-traitements 48

2.3.1 Importation des données 48
2.3.2 Choix des librairies 50
2.3.3 Annotation du régionalisme 51
2.3.4 Traitements textuels 54
2.3.5 Tokenisation 54
2.3.6 Vectorisation 55
2.4 Apprentissage 56
2.4.1 Liste des modèles 58
2.5 Conclusions pré-traitements 61
2.5.1 Traitement du tweet 61
2.5.2 Entités 62
2.5.3 Entités nommées 63
2.5.4 Métadiscours 63
2.6 Paramétrage des modèles 63
2.6.1 Résultats du paramétrage 65
3. Comparaison avec étude d'Evangelia Balouta 66
4. Visualisation des tweets 67
4.1 Prédiction et génération de fichiers 67
4.2 Qlik Sense Cloud 68
4.3 Étude de la vitalité et de l'expansion 69
4.3.1 Régionalismes du Nord (Belgique et Nord de la France métropolitaine) 70
4.3.2 Régionalismes du Sud 71
4.3.3 Comparaison de régionalismes 73
Conclusion 75
Bibliographie 77

Table des illustrations

Figure 1 : Dénominations du sac plastique (Avanzi et al. 2016) 8
Figure 2 : Représentation graphique de l'emploi sur Twitter des mots « man », « dude », « guy » et « bro » aux États-Unis 23
Figure 3 : Carte lissée de l'alternance « dad/father » (Ibid.)..... 27
Figure 4 : Exemple de tweet contenant un emploi réel de régionalisme 29
Figure 5 : Exemple de tweet représentant un métadiscours 30
Figure 6 : Exemple de tweet contenant une entité nommée 30
Figure 7 : Nombre de tweets par régionalisme et type d'annotation..... 47
Figure 8 : Répartition des classes dans le corpus d'apprentissage 49
Figure 9 : Moyenne d'entités par classe cible 50
Figure 10 : : Explication du paramétrage des arbres (« Complete Guide to Parameter Tuning in Gradient Boosting (GBM) in Python » 2016)..... 64
Figure 11 : Score d'exactitude et rapport de classification (Random Forest) 65
Figure 12 : Matrice de confusion (Random Forest)..... 65
Figure 13 : Exemple de visualisation de données sur Qlik Cloud 69
Figure 14 : Répartition géographique du régionalisme « drache »..... 70
Figure 15 : Répartition géographique du régionalisme « babache » 70
Figure 16 : Répartition géographique du régionalisme « chicon » (à droite), « clenche » et « clanche » (à gauche)..... 71
Figure 17 : Répartition géographique du régionalisme « dégun »..... 71
Figure 18 : Répartition géographique du régionalisme « tarpin » 72
Figure 19 : Répartition géographique des régionalismes « döner » et « dürüm » 73
Figure 20 : Répartition géographique des régionalismes « ducasse » et « feria » 73
Figure 21 : Répartition géographique des régionalismes « gsm » et « natel » 74

Table des tableaux

Tableau 1 : Paramètres impactant la vitalité d'une langue (Carli 2007) (version originale et traduite pour ce mémoire) 14
Tableau 2 : Répartition des tweets après l'annotation (tancarville) 39
Tableau 3 : Répartition des tweets après l'annotation (étendage et étandage) 40
Tableau 4 : Répartition des tweets après l'annotation (chicon) 40
Tableau 5 : Répartition des tweets après l'annotation (chocolatine) 41
Tableau 6 : Répartition des tweets après l'annotation (clanche et clenche) 42
Tableau 7 : Répartition des tweets après l'annotation (drache) 42
Tableau 8 : Répartition des tweets après l'annotation (carabistouille) 43
Tableau 9 : Répartition des tweets après l'annotation (ducasse) 44
Tableau 10 : Répartition des tweets après l'annotation (frairie) 44
Tableau 11 : Répartition des tweets après l'annotation (petiot) 45
Tableau 12 : Répartition des tweets après l'annotation (minot) 46
Tableau 13 : Répartition des tweets après l'annotation (tchot) 46
Tableau 14 : Répartition des tweets après l'annotation (petitou) 46
Tableau 15 : Répartition des tweets après l'annotation (gsm) 47
Tableau 16 : Répartition des tweets après l'annotation (natel) 47
Tableau 17 : Taux d'exactitude des modèles sur trois plis 61

Introduction

Les régionalismes sont des particularités linguistiques caractérisées par leur usage géographiquement restreint. Ce sont des écarts par rapport aux standards d'un pays ou d'une langue pouvant être de forme lexicale, phonétique ou syntaxique. Ce projet de recherche a pour but d'étudier la vitalité des expressions du français régional à travers le réseau social Twitter. Il fait suite à l'étude d'Evangelia Balouta (2017) qui a analysé les régionalismes d'un corpus composé d'articles de presse régionale. Il est important de préciser que ces sujets portent tous deux sur les formes lexicales et régionales de France, de Suisse romande et de Belgique francophone.

La première partie de ce projet de recherche se focalise principalement sur la réalisation de l'état de l'art. Nous parcourrons ainsi les différentes études qui ont été publiées sur le sujet des régionalismes. Nous verrons tout d'abord le contexte de l'étude et le projet auquel elle se rapporte. Dans un second temps, nous définirons les notions de base de la variation linguistique ; les dialectes, les facteurs de variations, le bon usage et la vitalité pour finir sur un exemple d'étude fondatrice. Nous observerons alors différents types de recueils de variations linguistiques. Dans cette première partie, nous analyserons ensuite les avantages et inconvénients de Twitter au travers d'études linguistiques portant sur ce réseau social. Nous finirons par une introduction à la problématique de la reconnaissance du métadiscours, c'est-à-dire le cas où une personne utilise un régionalisme tout en le décrivant, et des entités nommées pour une analyse plus profonde des régionalismes sur Twitter.

Notre deuxième partie se concentrera sur la mise en place d'un outil de classification automatique, destiné à différencier les utilisations réelles de régionalismes et les cas de métadiscours ou d'entités nommées. Tout en définissant les mots clés de la classification automatique, nous présenterons le corpus qui permettra de réaliser l'apprentissage du classifieur et les tâches nécessaires à sa mise en place. Puis, nous expliquerons les différents pré-traitements réalisés en amont de notre classification. Nous présenterons ensuite quelles méthodes ont été les plus efficaces pour l'apprentissage automatique et comment nous avons amélioré nos résultats. Nous décrirons ensuite ces derniers et les comparerons avec ceux obtenus par Evangelia Balouta. Pour finir, nous analyserons la répartition géographique de régionalismes classés tout en évaluant leur vitalité.

1. État de l'art

1.1 Contexte de l'étude

Ce mémoire s'inscrit dans le projet Français de nos régions (FNR - https://francaisdenosregions.com/). C'est une collaboration qui vise à rendre compte de la vitalité des régionalismes français. Le projet se concentrait à l'origine sur la France, la Belgique francophone et la Suisse romande, mais il s'est récemment étendu au Canada. Les études de ce projet consistent à interroger des internautes grâce à des enquêtes en ligne, dans le but d'analyser leur utilisation de variantes lexicales, phonétiques et parfois syntaxiques. C'est un projet innovant dans le sens qu'il réalise ses enquêtes grâce au crowdsourcing.

« La méthode dite du crowdsourcing (Cooke et al. 2013)*, soit littéralement « approvisionnement par la foule » ou « production participative », permet par exemple, sur une période relativement courte, de réaliser des sondages auprès de milliers de participants, présentant des profils variés (notamment en regard de leur origine géographique, de leur âge et de leur statut socio-éducatif) »* (Avanzi et al. 2016, p. 2).

Cette méthode se distingue des enquêtes classiques en linguistique diatopique qui étaient constituées de questionnaires oraux ou écrits, renseignés en laboratoire ou en contexte réel. En effet, le crowdsourcing permet de rassembler un échantillon assez important et représentatif de locuteurs en peu de temps et présente donc un avantage non négligeable. Les questionnaires peuvent toutefois provoquer des biais cognitifs si le participant est en insécurité linguistique, c'est-à-dire qu'il peut dire qu'il connaît une expression, mais il n'avouera pas qu'il l'emploie à cause des jugements dont il a été victime par le passé. Le crowdsourcing peut remédier en partie à ce problème puisque les questionnaires sont anonymes et ne demandent pas à la personne d'être physiquement présente. Cependant, il restera toujours une part de subjectivité, qui est une source d'erreur, puisque l'on demande au participant de répondre sur sa propre utilisation du langage.

En 2017, la plateforme avait déjà recueilli le témoignage de 50 000 internautes. Les résultats des enquêtes ont ensuite été publiés sur le site

« francaisdenosregions.com » et ont récemment fait l'objet d'un atlas linguistique : *Atlas du français de nos régions* (Avanzi 2017b).

Le projet a également publié un article en 2016 (Présentation d'une enquête pour l'étude des régionalismes du français) écrit par M. Avanzi, C. Barbet, J. Glikman et J. Peuvergne. Ils y décrivent leur première enquête, « Quel français régional parlez-vous ? », « auprès des collègues des universités partenaires, parallèlement via les réseaux sociaux informatisés et certains médias locaux au mois de mai 2015. En septembre 2015, plus de 12'000 participants avaient pris part à l'enquête » (Avanzi et al. 2016, p. 4).

Cette étude portait ainsi sur la vitalité et l'aire d'extension de plusieurs régionalismes de France, Suisse romande et Belgique. L'enquête s'est présentée ainsi : les internautes étaient invités à répondre par la forme qui leur semblait la plus naturelle et non pas par la forme qui serait la plus répandue ou grammaticale. Les participants devaient tout d'abord « répondre à une quinzaine de questions relatives à leur statut socioéconomique et sociolinguistique (âge, sexe, plus haut diplôme obtenu, métier, pays et localité où ils ont passé la plus grande partie de leur enfance, pays et localités(s) où ils ont passé une ou plusieurs grandes parties de leur vie) » (*Ibid.*, p. 3). Les participants devaient ensuite choisir la variante lexicale ou grammaticale la plus appropriée dans une liste fermée ou indiquer leur préférence sur une échelle de Likert (échelle allant de « tout à fait d'accord » à « pas du tout d'accord » pour évaluer la façon de penser des participants).

Des cartes ont ensuite été réalisées à partir des résultats ; celle que nous présentons ci-dessous est « une carte à point où un point correspond à un code postal » (*Ibid.*). Notons qu'il s'agit du « code postal de la localité d'enfance que chaque participant a renseigné » (*Ibid.*). Cela permet de représenter la répartition géographique de l'emploi des différentes variantes.

Figure 1 : Dénominations du sac plastique (Avanzi et al. 2016)

La carte ci-dessus a été réalisée grâce aux réponses des participants à la question « Comment appelez-vous cet objet ? », accompagnée d'une image de l'objet en question.

Les résultats obtenus ont ensuite été comparés aux zones décrites par d'autres études et par les dictionnaires et atlas régionaux déjà existants. Les auteurs ont ainsi pu constater que certaines variantes étaient utilisées sur une aire plus large, ou au contraire étaient de moins en moins employées. Nous pouvons prendre comme exemple « sachet », que le DRF ou Dictionnaire des régionalismes de France (Rézeau (ed.) 2001) considère comme le terme standard, tandis que cette étude démontre que le terme « sac » est bien plus utilisé de manière générale.

Cette étude n'est cependant pas complète puisque les données concernant le statut socioéconomique et les déménagements renseignés par les participants n'ont pas encore été traités. Une étude complémentaire permettrait de compléter ces analyses. En parallèle, l'équipe du projet FNR réalise sans cesse de nouvelles enquêtes afin d'enrichir sa base de données.

1.2 Définition des notions de variation linguistique

1.2.2 Origine des régionalismes et dialectes

L'origine des régionalismes est un sujet qui a pris de plus en plus d'ampleur en linguistique diachronique, grâce à l'intérêt croissant pour les variétés régionales rencontré ces dernières décennies. Selon Glessgen et Thibault (2005, p. 12) :

« La régionalité nait de facteurs variés qu'on peut ramener à trois catégories : archaïsmes (par rapport à la langue standard), néologismes (forgés à partir de celle-ci), et emprunts aux langues en contact (qu'il s'agisse de dialecte primaire, de langues alloglottes [langues différentes de celle du pays considéré], ou même d'une langue internationale comme l'anglais ou l'espagnol) ».

Intuitivement, nous pourrions penser que les emprunts sont majoritaires parmi ces trois catégories. Une étude a cependant mis en évidence qu' « un quart des entrées du DRF seulement représentent des emprunts » (*Ibid.*, chap. Éva Buchi, Les emprunts dans le Dictionnaire des régionalismes de France). Cette même étude précise que 56% des emprunts proviennent de l'occitan, les deux sources suivantes étant l'allemand/alsacien et les dialectes d'oïl.

La variation régionale est une notion complexe qui a pris de l'ampleur suite à l'extinction progressive des dialectes. Elle se distingue de ces derniers de par sa genèse et sa dimension restreinte puisqu'elle s'ajoute à une langue standardisée. En effet, la variation régionale ne représente pas une langue à part entière et reste compréhensible. Les Français de Paris comprennent, par exemple, un Français de Lyon malgré ses expressions et tournures syntaxiques atypiques. Afin d'éclaircir cette différence entre régionalité et dialectes, nous allons définir les notions de dialectes primaires et secondaires, selon le point de vue de Glessgen et Thibault (2005).

Les dialectes primaires de France sont des vestiges du latin et d'autres langues de proximité. Ce sont tout d'abord des langues parlées qui ont continué de se développer même après l'apparition des langues romanes écrites, et qui ont peu à peu souffert de l'expansion de ces langues « neutralisées ». Le picard et le breton sont des exemples de dialectes primaires. Ils s'opposent à la régionalité linguistique dans le sens que cette dernière ne peut pas « être antérieure à la formation des premières langues

élaborées et […] au développement de l'écrit » (*Ibid.*) car les régionalismes reposent justement sur les langues neutralisées et se développent grâce à ces dernières. Les dialectes secondaires sont quant à eux définis comme une « variation ultérieure d'une langue déjà constituée » (*Ibid.*). Ils sont le résultat de la différenciation des dialectes primaires. Nous pouvons citer comme exemple l'espagnol d'Andalousie ou le français parlé au Québec. Ils sont cependant trop éloignés de leur langue d'origine pour être considérés comme de simples régionalismes.

1.2.3 Types de régionalismes

Intuitivement, une personne lambda a tendance à associer les régionalismes avec des expressions familières. Cela est dû à l'influence des grammairiens qui a, pendant longtemps, poussé les locuteurs à dissimuler leur accent ou leurs expressions atypiques aux yeux des locuteurs d'autres régions plus « centrales ». Ces régionalismes étaient, et le sont toujours, réservés en grande partie à une sphère familiale ou locale. Cependant, un régionalisme, également appelé « diatopisme », peut également être un mot courant ou même soutenu ; « les diatopismes peuvent appartenir à n'importe quel niveau de langue (soutenue ou relâchée), voire au vocabulaire institutionnel et officiel (statalismes), et particulièrement de plein droit à la configuration de la langue standard. » (*Ibid.*, p. 8). Ces statalismes peuvent être employés dans n'importe quel contexte, sans être connotés négativement. Nous pouvons prendre comme exemple le mot « bourgmestre » qui est l'équivalent du maire en français de Belgique. Certains régionalismes peuvent également représenter une spécialité culinaire qui n'a pas d'équivalent en langue de référence. Le locuteur n'aura ainsi pas d'autres choix que d'utiliser ce mot, notamment pour partager sa culture.

1.2.4 Variation interne et externe

Nous continuons, dans notre souhait de définir les aspects de la variation régionale, vers une explication des facteurs internes ou externes de la variation. Cela nous aidera dans notre analyse des régionalismes.

Pour ce qui concerne la variation interne, il existe la variation phonique, la variation grammaticale et la variation lexicale. C'est sur cette dernière que nous allons nous concentrer. Chevillet (1991) évoque la variation lexicale de quatre manières. Tout d'abord, un même mot peut être employé dans deux régions distinctes pour désigner deux choses différentes. Le « déjeuner » fait par exemple référence au repas du matin en Belgique et au repas du midi en France. Le mot peut également avoir le même sens dans deux aires à l'exception que l'une d'entre elles aura un sens supplémentaire, une définition un peu plus spécifique. Nous pouvons citer comme exemple « poche » qui désigne un contenant situé sur un vêtement. Ce mot est en plus considéré comme un sac plastique dans le Centre-Ouest et Sud-Ouest de la France. Ensuite, un même mot peut être utilisé à des fréquences différentes, si nous reprenons l'exemple précédent il s'agirait de comparer les fréquences d'utilisation de « poche » dans plusieurs régions données. Pour finir, un même concept peut être évoqué par deux mots différents comme par exemple « soixante-dix » et « septante ». Ces deux dernières catégories sont sans doute celles que nous traiterons le plus dans ce projet car elles présentent des avantages. En effet, un même terme désignant deux concepts différents ne sera pas facile à analyser sur Twitter puisqu'il faudrait d'abord extraire le sens de chaque tweet pour savoir à quoi ce terme fait référence. Il faut également noter que certains régionalismes n'ont pas d'équivalent dans la langue de référence, nous comptons dans cette catégorie les statalismes ou encore les spécialités culinaires. Ils ne nous permettent donc pas de comparer leur fréquence d'apparition avec d'autres termes.

D'un autre côté, nous avons la variation externe. D'après Blanchet et Bulot (2013a, p. 48), la variation externe s'organise autour de cinq dimensions : diachronique, diatopique, diastratique, diaphasique et diamésique. Tout d'abord, la variation diatopique se réfère à une zone géographique spécifique. La variation diachronique, quant à elle, prend en compte la dimension temporelle. La variation diastratique reflète les couches sociales et dépend donc de l'individu et de son ressenti d'appartenir à telle classe sociale. Elle est souvent associée à la variation diaphasique qui rend compte des usages variant d'une situation de communication à une autre, le langage variera alors d'un registre soutenu à un registre familier en fonction du contexte. Pour finir, la variation diamésique reflète le support du langage, le canal utilisé. Elle permet notamment de mettre en lumière les écarts entre l'oral et l'écrit et de « prendre en

charge les formes liées aux nouvelles technologies » (Blanchet et Bulot 2013b). Le langage utilisé dans une conférence orale ou un article de journal écrit ne sera alors pas le même, nous reviendrons sur ce phénomène afin d'en voir les spécificités.

1.2.5 Bon usage et norme

« Le français pris en considération dans l'enseignement, les grammaires, les travaux de recherche en linguistique est le plus souvent une image de synthèse monochrome, dont la variation géographique est pratiquement exclue. » (Rézeau 2001)

Il reste encore à définir les notions de langue standard, de bon usage et de norme. Cette citation de Pierre Réseau résume parfaitement la vision qu'ont les institutions et l'éducation par rapport aux régionalismes. Pendant longtemps, le point de vue des grammairiens était le point de référence pour l'usage d'un bon français. Leur idée principale est associée au fait que le français standard est à dissocier des variations, qu'elles soient d'ordre sociologique ou régional. De plus, le système éducatif a aidé à renforcer cette idée en condamnant ce qui ne fait pas partie du bon usage. Les variations font pourtant partie intégrante de la langue standard, elles l'enrichissent et sont le reflet de notre héritage linguistique.

Dans le « Dictionnaire du bon usage au service du sens et de la nuance » P.-R. Ambrogi (2015), défendeur du bon usage, ne condamne pas les emprunts faits à d'autres langues ni les néologismes ; « tout ce qui contribue, par un effort maîtrisé, à l'enrichissement de la langue française, doit être encouragé » (*Ibid.*, p. 18). Il reconnait l'importance de l'évolution d'une langue et des échanges entre culture, cependant cela se fait en respectant certaines règles ; « Débarrassons-le [français] d'altérations accumulées, d'emprunts indignes, de constructions hasardeuses pour redécouvrir et entretenir le « bon usage » » (*Ibid.*). Cela permettrait ainsi d'entretenir une intercompréhension et une « unité linguistique » du français. De plus, il explique que les grammairiens luttent contre l'appauvrissement de la langue, provoqué par les erreurs de français transmises par les médias ou par ceux qui n'ont pas réussi dans le système scolaire. L'usage de la langue peut cependant être enrichi par les « pressions imposées par les modifications juridiques, technologiques ou sociales » (*Ibid.*, p. 19).

Cependant ce système de bon usage reste instable et peu pertinent pour étudier les régionalismes. Ce sont des notions souvent trop subjectives ou abstraites et ne rendent pas compte de la réalité d'une langue. Une manière de relever des régionalismes serait par exemple de définir un français central, ou neutralisé qui serait le point de référence pour analyser les régionalismes. Voici ainsi quelques définitions possibles :

Selon J.Dubois (1965, p. 5) « La norme se définit par l'intercompréhension la plus étendue. Le français étudié est alors dit neutralisé », ce qui voudrait dire que logiquement, la norme devrait être représentée par les générations moyennes, les « groupes sociaux urbains, dont le volume de communications est le plus grand relativement à l'ensemble linguistique français » (*Ibid.*). Warnant L. (1973) s'est alors attardé sur la question de français neutralisé, et sur comment l'on pourrait mieux le définir afin de relever les usages régionaux par rapport à « l'usage central ».

« Il faudrait, au départ, considérer qu'il n'existe aucun dialecte privilégié et recueillir un nombre suffisant de corpus régionaux, les corpus des dialectes de Paris, de Lille, de Marseille, de Liège, de Genève, de Bordeaux » (*Ibid.*, p. 107)**.**

Le français neutralisé serait alors composé des « traits qui sont communs à tous les dialectes » et « des traits non communs, ceux qui sont utilisés par le plus grand nombre des locuteurs » (*Ibid.*), cela permettrait d'obtenir une majorité de traits du français central. Selon cet auteur, les expressions qui représentent une réalité non nommée en français central, devraient également être ajoutées. Un exemple serait le mot « *drève*, employé en Belgique pour désigner une allée bordée d'arbres » (*Ibid.*, p. 108). Puisque ce mot n'a pas d'équivalent, il doit « être considéré comme le plus fréquemment utilisé et faire partie du français neutralisé » (*Ibid.*).

Une autre possibilité serait de définir un français neutralisé distinct pour la France, la Suisse romande et la Belgique francophone. Cela permettrait de rendre compte des différentes variations au sein de ces zones géographiques et de bien traiter les régionalismes en tant que tels. Cependant, nous avons développé ces notions à titre informatif et nous ne prendrons pas partie pour l'un de ces points de vue dans ce mémoire.

1.2.6 Vitalité

Cette étude a pour but d'étudier la vitalité des régionalismes. Le concept de vitalité est cependant complexe à définir dans ce contexte puisqu'il s'utilise généralement pour parler de vitalité des langues. Ce concept est alors utile pour estimer combien de temps une langue restera vivante. Un degré de vitalité ou de menace est alors évalué selon différents paramètres. Tout d'abord, ce degré d'extinction est en grande partie lié à la communauté de locuteurs et à la transmission intergénérationnelle. Nous pouvons le voir sur ce tableau qui classe « les paramètres que l'UNESCO considère comme indispensables afin d'évaluer le degré de vitalité d'une langue donnée ou de menace, surtout pour une langue minoritaire » (Carli 2007, p. 5) :

1. Intergenerational language trasmission	6. Materials for language education & literacy
2. Absolute numbers of speakers	7. Govern.-institut. language attitudes/policies
3.Proportion of speakers within the total population	8. Community members' attitudes towards their own language
4. Loss of existing language domains	9. Amount and quality of documentation
5. Response to new domains / new media	10. *Others?*

1. Transmission intergénérationnelle de la langue	6. Ressources pour l'enseignement de la langue et l'alphabétisation
2. Nombre absolu de locuteurs	7. Dispositions et politiques langagières des gouvernements et institutions
3. Proportion de locuteurs dans la population totale	8. Attitude des membres de la communauté envers leur propre langue
4. Perte de domaines langagiers existants	9. Quantité et qualité de la documentation
5. Réponses à de nouveaux domaines/de nouveaux média	10. Autres ?

Tableau 1 : Paramètres impactant la vitalité d'une langue (Carli 2007) (version originale et traduite pour ce mémoire)

Nous pouvons alors noter le nombre de paramètres entrant en compte et la difficulté à évaluer cette vitalité. Dans notre contexte de régionalismes, nous n'aurons pas à

prendre en compte tous ces éléments, mais ils nous permettront sans doute d'émettre certains commentaires quant à la répartition des items relevés. De plus, les expressions régionales évoluent selon des facteurs que nous pouvons mettre en relation avec ceux cités ci-dessus (tab. 1). Tout d'abord, les régionalismes peuvent se former à partir de néologismes, en réponse à un besoin ou à un nouveau concept, par exemple ; « cagouille 'escargot' est utilisé pour nommer l'arobase chez les charentais » (Rézeau (ed.) 2001). De plus, « certains régionalismes deviennent des mots utilisés dans le français populaire (ex : les emplois argotiques qui ne l'étaient pas dans leur région d'origine – castagne, morfler) » (*Ibid.*). Dans ce cas, la vitalité diminue, au profit de la langue standard et le nombre de personnes employant ce régionalisme en tant que tel diminue également. De nos jours, les écarts régionaux sont connotés de moins en moins négativement et la communication médiatique sur le sujet semble provoquer, notamment chez les jeunes, l'envie d'affirmer leurs origines. Cela se traduit par exemple par des « hashtags » sur Twitter ou des groupes sur Facebook, comme « Tu sais que tu es de Tours si tu dis… ». Ce phénomène fait augmenter le nombre de locuteurs et consolide la transmission intergénérationnelle. Pour donner un dernier exemple, de nombreux régionalismes perdent tout simplement leur vitalité, de par leur expansion trop restreinte ou leur manque de transmission effective vers la jeune génération. En prenant en compte tous ces changements, nous comprenons qu'il est difficile d'évaluer la vitalité des régionalismes, ce mémoire en ressort ainsi plus pertinent. En effet, l'étude des expressions régionales via les réseaux sociaux permettra sans doute d'obtenir des résultats permettant de comparer la résistance des régionalismes les uns par rapport aux autres.

1.2.7 Les fondements de la linguistique variationniste

Les variations linguistiques ont pendant longtemps été mises de côté par les linguistes, qui considéraient celles-ci comme appartenant au domaine psychologique ou sociologique. La linguistique variationniste trouve ses principes fondamentaux aux Etats-Unis dans les travaux de William Labov. L'une de ces études consiste à étudier le côté social des variations et la conscience qu'ont les locuteurs de ces variantes.

En effet, en 1963, il va effectuer une enquête dans les grands magasins de New York afin d'étudier la variation du phonème /r/ selon la classe sociale des locuteurs (Labov

2006). Pour cela il va mener son enquête dans trois magasins : Saks Fifth Avenue, Macy's et Klein, qui représentent respectivement les classes sociales supérieures, moyennes et inférieures. Il émet l'hypothèse que les classes supérieures auront plus tendance à utiliser le /r/ rétroflexe et qu'inversement, les classes plus modestes utiliseront peu ce /r/ rétroflexe. Pour rappel, un /r/ rétroflexe est produit avec l'apex de la langue positionné vers le haut dans la région alvéolaire.

Sa méthodologie est la suivante : un ou plusieurs enquêteurs doivent poser certaines questions aux employés des trois magasins. Ils doivent tout d'abord demander : « Excuse me, where are the women's shoes ? » (Excusez-moi, où sont les chaussures pour femmes ?). Puis, une fois arrivés au 4ème étage ils devaient poser la question : « Excuse me, what floor is this ? » (Excusez-moi, à quel étage sommes-nous ?) Cela permettait d'obtenir des usages de /r/ différents selon le contexte et la place du /r/ dans la réponse « fourth floor ».

Ce type de méthodologie a pour avantage de ne pas provoquer de biais cognitif, dû à l'insécurité linguistique de la personne interrogée puisque celle-ci n'a pas conscience de participer à une enquête. Comme nous le disions, cette étude a eu un fort impact international sur la recherche en variation linguistique. De nombreux chercheurs se sont alors intéressés à ce domaine et ont tenté de relever et analyser les régionalismes de différentes manières.

1.3 Recueils de régionalismes

1.3.1 Atlas linguistique

Maintenant que nous avons défini les différents aspects de la variation, nous pouvons expliquer de quelle manière cette dernière est répertoriée.

Les atlas linguistiques sont une manière de représenter géographiquement la répartition des variétés régionales. Tout d'abord, les premières cartes se focalisaient sur l'oral puisque c'est ainsi que les régionalismes se manifestent en majorité. Les

premiers atlas représentaient donc sur une carte les différentes prononciations qu'un mot pouvait avoir dans chaque dialecte, à l'aide d'une notation phonétique. Les volumes de l'Atlas linguistique de la France (ALF) publiés entre 1902 et 1910 en sont un exemple. L'auteur J. Gilléron et le dialectologue E. Edmont ont ainsi parcouru les campagnes de France, Suisse romande et Belgique dans le but de relever les différentes formes dialectales encore présentes.

« Gilliéron établit un questionnaire de 1 400 articles, sélectionnés selon des critères divers, et Edmont, de 1897 à 1901, visita 639 localités du domaine gallo-roman ; à chaque point du réseau, il cherchait un informateur qui eût le loisir, la bonne grâce et la compétence de traduire dans son vernaculaire les mots français qu'on lui soumettait » (Séguy 1973, p. 67).

L'Atlas de 1920 cartes rencontra à sa sortie un fort succès et fut à l'origine de nombreuses études en dialectologie. Cependant, il avait plusieurs lacunes, causées par cette méthodologie innovante mais fastidieuse. Des auteurs comme Dauzat tentèrent alors d'élaborer leur propre atlas linguistique. Ce dernier réalisa alors un questionnaire « Le Nouvel Atlas linguistique de la France par régions ». Celui-ci fut remarqué par le CNRS et c'est ainsi qu'est né le projet des « Atlas linguistique de la France par régions », qui a regroupé de nombreux dialectologues au fil du temps.

Aujourd'hui, nous pouvons également évoquer l'Atlas du français de nos régions de Mathieu Avanzi (2017b), que nous avons déjà abordé plus haut, et qui a l'avantage d'être adapté à un public néophyte en linguistique.

1.3.2 Dictionnaires de régionalismes

Les régionalismes peuvent être représentés dans certains dictionnaires standards par une annotation « rég. », mais la principale source pour ces phénomènes reste les dictionnaires de régionalismes. Nous allons nous concentrer sur trois d'entre eux : le dictionnaire des belgicismes (DB, Francard 2015), le dictionnaire des régionalismes de France (DRF, Rézeau 2001) et le dictionnaire suisse romand (DSR, Thibault et Knecht 2004). Ces ouvrages ne sont bien sûr pas l'œuvre d'un seul auteur, ils sont le fruit de longues collaborations entre des chercheurs et universitaires de différents domaines de spécialité.

Dans sa préface, l'auteur du DB décrit ce recueil ainsi ; « Il s'agit […] de formes, de sens, de constructions grammaticales et d'expressions dont la connaissance et l'usage sont répandus tant en Wallonie qu'à Bruxelles » (Francard 2015, p. 7). Le DSR a pour but de cerner les « particularités lexicales du français contemporain en Suisse romande ». Le DRF, quant à lui,

« décrit, sous l'angle lexical, les principales variétés géographiques du français, observées dans la seconde moitié du 20^{e} siècle en France (dans ses limites continentales) et qui témoignent de la diversité des réalisations de la langue selon les régions. » (Rézeau 2001).

En d'autres mots, ces ouvrages constituent une liste non exhaustive des régionalismes de France, Belgique francophone et Suisse romande. Chaque item peut être associé à une ou plusieurs langues en contact, pour mettre en évidence l'origine du régionalisme, à des informations chronologiques et géographiques, à des références quant à l'usage francophone et à des sources.

Le DSR précise qu'il n'est pas fait pour imposer une norme, il n'est que descriptif, c'est-à-dire qu'il « enregistre un usage ». Il est important de préciser ce fait car la majorité des recueils de régionalismes réalisés le siècle dernier avaient pour but de condamner les « barbarismes » du français régional, tout en étudiant en profondeur ces particularités lexicales. Nous pouvons citer dans cette catégorie l'ouvrage « Chasse aux Belgicismes » et le « Nouveau dictionnaire des difficultés du français moderne » de Joseph Hanse et al., respectivement publiés pour la première fois en 1971 et 1983. Ces ouvrages étaient justifiés par le fait que « ces mots et expressions [étaient] considérés pendant longtemps tout simplement comme « fautes de français » » (Thibault et Knecht 2004, p. 9). Cela n'est pas étonnant si l'on considère l'influence des grammairiens qui a subsisté pendant plusieurs siècles. À cette époque, les régionalismes et les dialectes demeuraient mal perçus par les grammairiens et représentaient « un vocabulaire non légitimé par la norme, dont l'accès est souvent verrouillé par l'autocensure ». Cependant, les grammairiens n'étaient pas les seuls fautifs puisqu'en Belgique comme en Suisse, la population était soumise aux jugements de leur voisin français. Cela a donc pendant longtemps engendré une « population soucieuse de ne pas se singulariser linguistiquement par rapport aux Français » (Thibault et Knecht 2004).

Aujourd'hui, Bruxelles et la Wallonie sont majoritairement francophones mais la population de Bruxelles était principalement flamande au XIXe siècle, tandis que la Wallonie est francophone depuis plusieurs siècles. Le DB relève donc de nombreuses différences dans l'appellation d'un même concept (par exemple ; maquée/plattekeis - fromage blanc) ou des différences de diffusion dans les termes. Cependant, ces divergences restent minoritaires face aux belgicismes qui se sont étendus sur ces deux territoires. La Suisse romande est, quant à elle, l'héritage de nombreux évènements historiques. Le français parlé et écrit sur ce territoire est une « variété issue du domaine d'oïl », ce qui explique les ressemblances que l'on peut trouver avec les autres régions françaises limitrophes.

Les ouvrages qui ont précédemment tenté d'énumérer les régionalismes se basaient essentiellement sur les observations subjectives des auteurs et sur les quelques ressources textuelles disponibles. Ici, dans le Dictionnaire des belgicismes, « la nomenclature (formes et sens retenus) est fondée sur une enquête préalable, menée à partir de 2000 auprès d'une centaine d'informateurs de Wallonie et de Bruxelles » (Francard (ed.) 2015, p. 10). Les auteurs ont ainsi présenté des belgicismes décrits dans des ouvrages publiés antérieurement et ont retenu 2000 items « qui étaient compris pour au moins 50% de [leurs] informateurs et utilisés effectivement par au moins 30% d'entre eux » (*Ibid.*), cela leur a permis de ne retenir que les mots et les sens représentatifs du parler de la Belgique francophone et de décrire les usages réels.

Nous retrouvons la même méthodologie dans le DSR et le DRF. Ce dernier décrit cependant une dernière étape de comparaison des résultats avec des sources écrites, orales et lexicographiques et mentionne également les bases de données de l'ATILF (anciennement INALF) de Nancy. Cependant, des études postérieures mettent en évidence le manque de ressources électroniques et textuelles du DRF et décrit le fait qu'il aurait « souffert aussi de la grave pénurie documentaire qui affecte les périodes allant du 16^{e} au 19^{e} siècle » (Glessgen et Thibault 2005, p. 15), ce qui est sans doute un problème commun à de nombreuses autres études.

1.4 Comparaison entre la presse et Twitter

Le but premier de ce mémoire est d'analyser la vitalité des régionalismes. Il serait donc plus pertinent de se concentrer sur les régionalismes familiers ou courants, plutôt que sur les statalismes. En effet, ces derniers sont employés dans les textes de lois, administratifs ou officiels et font donc sans doute partie du vocabulaire de toute la population belge ou de Suisse romande. Nous pourrons les relever à titre comparatif mais notre étude portera principalement sur les régionalismes moins bien attestés et dont on ne connait pas l'expansion exacte.

Ce mémoire avait tout d'abord pour but de relever les régionalismes dans la presse régionale. Nous nous sommes alors posé des questions quant à la pertinence de cette source. En effet, la presse révèle deux principaux problèmes : son caractère formel et l'absence de dialogue. Nous avons abordé le premier point précédemment, en relevant le fait que les régionalismes que nous voulons étudier font principalement partie du domaine courant et familier. Nous pensons donc qu'un journal ne pourra pas remplir cet objectif. Certains journaux ont bien sûr un ton léger qui permet l'emploi de régionalismes familiers mais ils ne sont pas majoritaires. De plus, nous savons que le « français écrit par des personnes d'un certain niveau social ne diffère pas essentiellement d'une région à l'autre du domaine linguistique français. C'est dans le langage oral que se marquent plus spécialement les « traits régionaux » » (Warnant 1973, p. 116). Un dialogue, même écrit, se rapproche d'un langage oral. Cependant, ce médium ne représente pas un dialogue entre le lecteur et le journaliste. Il existe une certaine distance entre ces deux acteurs. Seules les lettres et annonces des lecteurs pourraient permettre d'obtenir un discours authentique et sans retenue.

Nous avons alors suivi une nouvelle direction ; celle du réseau social Twitter. Cette plateforme permet le dialogue entre les utilisateurs, ces derniers se sentent libres de parler comme ils le souhaitent, que ce soit pour répondre à des connaissances ou publier leurs idées. Cela permet donc de recueillir un langage au registre courant ou familier, qui se rapproche d'une conversation orale ; « microblogging services provide us with real-time communication among users that […] tend to employ an oral speech» (les services de microbloggage nous fournissent une communication en temps réel

entre des utilisateurs qui […] tendent à employer le langage oral) (Donoso et Sanchez 2017, p. 9).

Nous pouvons évoquer une première critique sur le réseau social Twitter, nous développerons les inconvénients et avantages de ce site plus loin dans ce projet. L'un des inconvénients est le fait que les utilisateurs de ce réseau social ne s'adressent pas à une communauté bien délimitée. Ils ont bien une communauté d'abonnés mais bien plus de personnes peuvent voir leurs tweets avec le principe de « retweet » (partager un tweet sur son propre profil). Ainsi, contrairement à la presse régionale qui s'adresse à un public spécifique, les utilisateurs de Twitter ne peuvent pas utiliser de régionalismes s'ils veulent être compris par tous les lecteurs potentiels. Nous pourrons donc craindre un emploi de variations réservé aux utilisateurs ayant un nombre d'abonnés restreint ou qui s'adressent à un public spécifique.

1.5 Étude de la variation régionale sur Twitter

1.5.1 Comment constituer un corpus sur Twitter ?

Avant de présenter différentes études portant sur l'analyse des régionalismes sur Twitter, nous allons rapidement expliquer comment il est possible de rassembler des tweets afin de constituer un corpus.

Le réseau social Twitter met à disposition des APIs (Application Programming Interface – Interface de programmation applicative), utiles pour les développeurs ou ceux qui souhaitent collecter des données. Certaines APIs permettent de créer des publicités ou de suivre l'activité de son compte Twitter et de ses abonnés. D'autres sont utiles pour la recherche de tweets selon certains critères (mot(s) recherché(s), dates…). La « Search API » est celle que nous utiliserons pour constituer notre corpus de tweets (https://developer.twitter.com/en/docs/tweets/search/overview). Après s'être abonné à cet outil, il est possible de traiter les données recueillies à l'aide des langages de programmation Python ou R.

Certains tweets sont géolocalisés, c'est-à-dire que l'utilisateur a accepté de divulguer sa position à Twitter lorsqu'il a écrit son message. Ces coordonnées

spatiales sont ainsi récupérables grâce à ces APIs, cela nous permettra d'observer la répartition spatiale des régionalismes recherchés. Le nombre de tweets géolocalisés n'est malheureusement pas très élevé (2 à 3%), même si ce chiffre peut varier selon les pays et tend à augmenter avec le temps.

1.5.2 Les travaux de Jack Grieve

Jack Grieve est un linguiste spécialisé dans l'analyse de corpus, la dialectologie, la variation et les changements du langage. Il travaille également avec le système judiciaire pour analyser des preuves orales ou écrites. Ses recherches se focalisent sur la variation grammaticale et lexicale de l'anglais. De plus, il développe des méthodes d'analyse linguistique quantitative, notamment pour l'étude de corpus sur Twitter (https://sites.google.com/view/grievejw/home).

Avec l'aide de Andrea Nini et Diansheng Guo, Jack Grieve parvient à développer l'application « Word Mapper » en 2016. Elle permet aux utilisateurs de « cartographier la fréquence relative des 10 000 premiers mots les plus communs dans un grand corpus basé sur Twitter couvrant l'ensemble des États-Unis »[1] (http://languagelog.ldc.upenn.edu/nll/?p=29824). L'application a ensuite été améliorée jusqu'à proposer 97 246 mots apparaissant au moins 500 fois dans le corpus. Ce dernier comprend 890 millions de Tweets récupérés entre Octobre 2013 et Novembre 2014. Les cartes présentées sur le site ne montrent pas la simple répartition des données brutes. Elles ont été lissées (« regional smooting »), c'est-à-dire que les aires ayant des résultats similaires ont été regroupées afin d'obtenir des tendances géographiques.

[1] « map the relative frequencies of the 10,000 most common words in a big Twitter-based corpus covering the contiguous United States » (notre traduction)

Cet outil permet à des utilisateurs néophytes en linguistique d'observer l'utilisation d'un mot choisi à travers tout le pays. Ainsi, certaines personnes pourront réaliser une recherche comme celle présentée ci-dessous très rapidement.

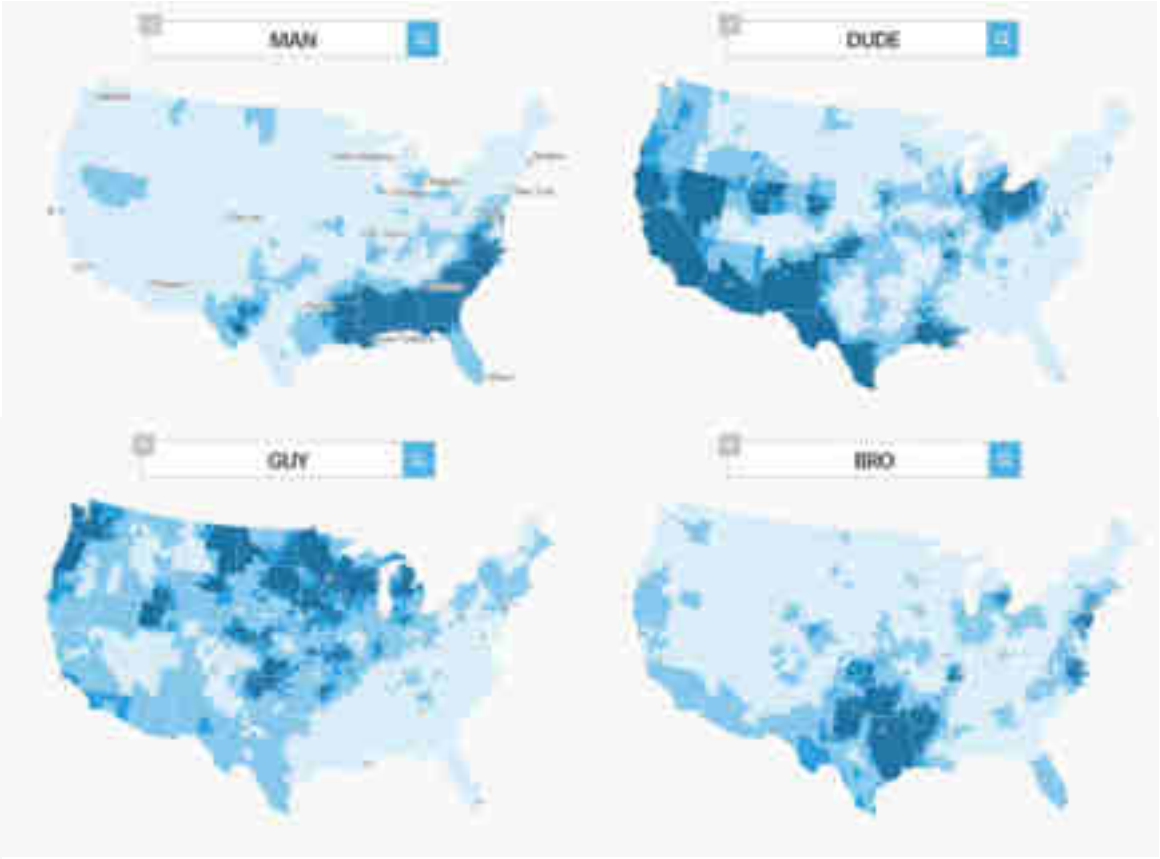

Figure 2 : Représentation graphique de l'emploi sur Twitter des mots « man », « dude », « guy » et « bro » aux États-Unis

Cette carte (fig. 2), générée à partir du lien https://qz.com/862325/the-great-american-word-mapper/#int/words=man_dude_guy_bro&smoothing=3, met en évidence les variations utilisées pour appeler un homme, souvent un ami, aux États-Unis. Elle nous permet d'observer des répartitions bien distinctes à travers les états, pour des variantes pourtant connues de tous.

Les mots disponibles étant très variés, certains spécialistes trouveront sur ce site de nombreuses ressources pour leurs travaux. Il est également possible de récupérer les cartes lissées ou non et celles de chaque comté. Cela peut également être utile pour élaborer des corrélations avec les faits d'actualités rencontrés en 2013 et 2014.

Jack Grieve est également à l'origine d'une application qui cartographie l'expansion des néologismes aux États-Unis (https://qz.com/465820/how-brand-new-words-are-

spreading-across-america/). L'application, qui a pour titre « How brand new words are spreading across America » (Comment les nouveaux mots se diffusent à travers l'Amérique), permet à l'utilisateur de sélectionner un mot parmi une liste et d'observer dans quelle région ce mot est apparu et de quelle manière il s'est répandu sur Twitter entre Octobre 2013 et Novembre 2014. Jack Grieve a sélectionné des mots très rarement utilisés en 2013 (moins de une fois par millions de mots) et a conservé ceux devenus communs ou très répandus en 2014. Nous pouvons ainsi noter dans cette liste « amirite » (Am I right ? - J'ai raison ?), « bruuh » (évolution de « bro » - frère, mec), « on fleek » (au point, parfait). Ce dernier est intéressant car, inventé par une utilisatrice du site Vine, il a commencé à être utilisé en Juillet 2014. En Novembre 2014, cette même expression était devenue très répandue et connue dans tout le pays.

À l'époque d'Internet et des réseaux sociaux, l'expansion d'un néologisme devient plus facile qu'autrefois. Mais il est intéressant de noter que certains néologismes comme « unbothered » restent concentrés dans le Sud-Est tandis que d'autres s'étendent sur tout le territoire. Comme nous l'avons vu (fig. 2) , des mots comme « dude » ou « bro » sont préférentiellement utilisés par une région et presque rejetés par d'autres. Nous pouvons ainsi penser que certains néologismes ne s'étendront jamais sur tout le territoire tandis que d'autres s'effaceront peu à peu jusqu'à rester présent dans une zone géographique plus restreinte. Pouvons-nous retrouver ce phénomène à l'échelle de la France ? Pouvons-nous dans ce cas considérer ces mots comme des régionalismes ? En France, la question des régionalismes n'est pas assez souvent étudiée sur les réseaux sociaux pour pouvoir émettre une réponse aujourd'hui. Des analyses supplémentaires pourront certainement un jour rendre compte de l'évolution des néologismes et des régionalismes sur les réseaux sociaux en France.

1.5.3 Les régionalismes aux États-Unis

L'étude qui va être présentée ici, « Understanding U.S. Regional Linguistic Variation with Twitter Data Analysis » (Comprendre la variation linguistique régionale aux U.S. grâce à l'analyse de données sur Twitter), a été rédigée par Huang Y. et al. en 2016. Le but de cette étude est d'analyser les variations lexicales présentes sur Twitter. Ils ont tout d'abord constitué un corpus de tweets géolocalisés aux États-Unis. Ils ont ainsi rassemblé 924 millions de tweets écrits par 6.6 millions d'utilisateurs uniques entre le

7 octobre 2013 et le 6 octobre 2014. Ces auteurs ont ensuite utilisé une liste d'alternances lexicales formées de deux ou plusieurs manières d'évoquer le même concept, où chaque appellation est une variante, souvent indépendante du contexte [2] (ex. : bag/sack, clearly/obviously). Cette liste de 211 alternances a été prélevée des travaux de Grieve et al. (2013).

Les auteurs ont ensuite éliminé 152 alternances, insuffisamment présentes dans les tweets des utilisateurs. Ils ont alors voulu calculer la préférence de chaque comté en termes de variante. Ils ont tout d'abord déterminé A(w,v) comme A l'alternance ayant une variante w la plus fréquemment utilisée dans le corpus et v la ou les variantes moins utilisées. T(u,c,w) est alors un nombre total de tweets envoyés par un utilisateur u dans un comté c et qui contiennent la variante w. Dans ce cas, la « variante-preference » VP est calculée ainsi :

$$VP(u,c,A) = T(u,c,w)/(T(u,c,w)+T(u,c,v))$$

Cela donne un ratio compris entre 1 et 0. Notons que si la variante w est bien la plus utilisée pour cet utilisateur, ce ratio sera supérieur à 0,5 et inversement.

Selon ces auteurs, la mobilité de l'utilisateur a une plus grande importance que l'origine géographique de ce dernier :

« Since mobility and migration have strong influence on the formation of linguistic characteristics of a place, we use the location of each tweet instead of finding a home county for each Twitter user» (notre traduction : Puisque la mobilité et la migration ont une forte influence dans la formation des caractéristiques linguistiques d'une zone, nous utilisons la localisation de chaque tweet au lieu de chercher un comté d'origine pour chaque utilisateur sur Twitter) (Huang et al. 2016).

Ils n'ont ainsi relevé que la localisation de chaque tweet et ont donc dû considérer un nouvel utilisateur pour chaque nouveau comté d'où ce dernier a envoyé un tweet, afin

[2] « A lexical alternation consists of two or more different words with the same referential meaning, referred to as variants » (notre traduction)

de rester cohérents dans leurs calculs. Cela sous-entend qu'une personne parle davantage en fonction de son lieu d'habitation que son lieu de naissance. Cependant, une personne qui vient d'emménager prendra certainement plusieurs mois avant d'employer les expressions propres à cette région, et continuera d'utiliser les siennes pendant une certaine période. De plus, un tweet peut être écrit lors d'un déplacement ou de vacances, ce qui pourrait donc fausser les résultats. L'utilisateur ne renseigne pas toujours sa localisation sur son compte, mais celle-ci peut indiquer la ville où il a habité le plus longtemps. Une étude approfondie pourrait permettre de savoir s'il est plus pertinent de relever la position indiquée sur le compte de l'utilisateur ou la position de chaque tweet, compte tenu de la quantité de données et du lissage effectué postérieurement.

La préférence de variance moyenne (MVP) est ensuite calculée pour tous les utilisateurs d'un comté et pour chaque alternance. U(c,A) représente alors le nombre total d'utilisateurs ayant utilisé l'alternance A(w,v). La préférence moyenne d'un comté pour les variantes d'une alternance est calculée de la manière suivante :

$MVP(c,A) = \sum_u VP(u,c,A) / U(c,A)$.

Nous pouvons noter que ce calcul donne à chaque utilisateur un poids égal, indépendamment de leur nombre de tweets.

Suite à une étude probabiliste, ces chercheurs retiennent 38 alternances significatives. Ils ont alors pu créer des cartes représentant la répartition géographique des variantes selon les comtés. Cependant, leurs relevés sur Twitter n'ont pas permis d'obtenir une répartition équitable en ce qui concerne le nombre d'utilisateurs et de tweets par comté, il y a également une différence de taille entre chaque comté. Différentes opérations de lissage (fig. 3) ont ainsi été réalisées afin d'homogénéiser les résultats et leur représentation sur la carte.

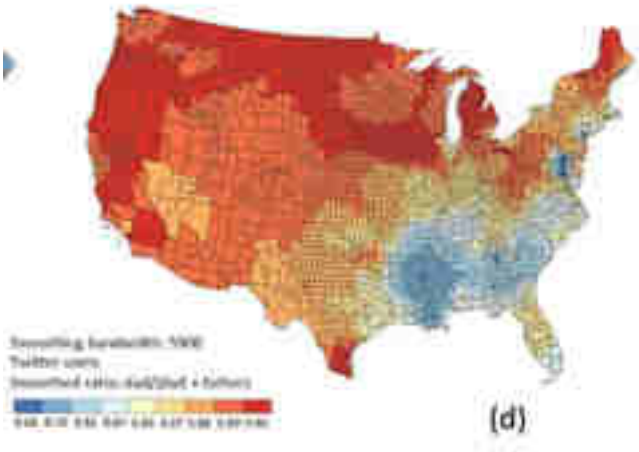

Figure 3 : Carte lissée de l'alternance « dad/father » (Ibid.)

Cette étude tente, pour finir, de définir des régions qui pourraient partager des ressemblances dialectales et de comparer celles-ci avec les travaux obtenus par des chercheurs comme Labov et Grieve.

1.5.4 Inconvénients de Twitter

Plusieurs problèmes sont relevés par les auteurs de l'étude que nous venons de présenter. Tout d'abord, le réseau social Twitter n'a pas toujours des données très fiables, notamment à cause de la démographie des utilisateurs et de leur langage pouvant contenir des abréviations ou des fautes. De plus, les tweets géolocalisés ne représentent que 2 à 3% de la totalité des tweets, il est alors nécessaire de rassembler un corpus assez conséquent pour palier le nombre de données inutilisables. La technique du text-mining a été testée par Xu, Wong et Yang (2013) afin de relever des informations géographiques directement dans le texte des tweets mais cela demande un nombre de données encore plus important. Pour finir, Twitter contient de nombreux spams envoyés par des organisations souhaitant faire de la publicité ou présenter l'actualité. En l'occurrence, un MVP mettant chaque utilisateur à la même échelle a permis de diminuer ce problème. Malgré les inconvénients de Twitter, les résultats de cette étude montrent que l'analyse en « Big Data » a de nombreux avantages par rapport à la méthodologie traditionnelle, puisqu'elle permet d'obtenir des résultats semblables pour un échantillon plus important.

1.5.5 Études similaires

Une étude similaire à celle de Huang Yuan et al. est celle de Gonzalo Donoso et David Sanchez (Dialectometric analysis of language variation in Twitter) (2017), celle-ci étudie la variation sur un corpus constitué de tweets situés principalement en Espagne. Elle se base sur 11 millions de tweets géolocalisés relevés en 2014 et 2016. Nous pouvons remarquer que toutes les études portant sur ce thème s'aident d'un corpus très conséquent et nous savons que nous ne pourrons pas rassembler la même quantité de données, ce qui rendra nos analyses probablement moins fiables. Toutefois, cette étude espagnole réalise plusieurs traitements qui pourront nous intéresser par la suite ; notamment la détection de la langue, faite à l'aide du Compact Language Detector (CLD) (McCandless 2012). Il existe d'autres méthodes mais celle-ci semble obtenir de bons résultats, cela permet de ne relever que les tweets écrits en français, car certains régionalismes peuvent être empruntés à l'allemand ou au flamand. Les auteurs retirent également les mots dièses et les mentions d'utilisateurs accompagnées d'une arobase. De plus, cette étude s'aide d'une liste de variations déjà réalisée pour la langue espagnole et effectue une analyse des distances géographiques entre les différentes variations.

L'étude « Mapping German Tweets to Geographic Regions » (Cartographier les tweets allemands en régions géographiques) de Tatjana Scheffler et al. (2014) est intéressante dans le sens où ces chercheurs effectuent la démarche inverse par rapport aux études mentionnées précédemment. En effet, ils tentent de retrouver la localisation des tweets en étudiant le dialecte et le lexique employés. Toutefois, cette méthode serait difficile à réaliser en France car les variations linguistiques n'y sont pas autant documentées qu'en Allemagne.

1.6 Prise en compte des entités nommées et du métadiscours

1.6.1 Cas du métadiscours

Un des points délicats de ce projet est d'analyser la fréquence des régionalismes, en ne gardant que les emplois réels de ces derniers.

Figure 4 : Exemple de tweet contenant un emploi réel de régionalisme

Dans l'exemple ci-dessus, le régionalisme « tancarville » (qui désigne un séchoir à linge), est utilisé dans ce que l'on appelle un emploi réel. Cependant un régionalisme peut être retrouvé dans des situations de communication ou des contextes différents de ceux attendus dans une utilisation authentique. Dans un premier temps, une personne peut employer un régionalisme pour décrire sa propre façon de parler ou celle d'une autre personne. Dans ces cas, on parlera de métadiscours, où la forme régionale sert par exemple à décrire ou demander dans quelle région celle-ci est employée ou encore à évoquer la disparition ou l'expansion du régionalisme. Ce métadiscours sera fréquemment présent dans les articles ou tweets de journalistes, de sociologues, de linguistes, mais également chez des personnes non spécialistes et simplement curieuses. Si nous prenons l'exemple de Twitter, un utilisateur pourra faire remarquer l'emploi d'une expression atypique ou affirmer sa manière de nommer un concept (fig. 5).

Figure 5 : Exemple de tweet représentant un métadiscours

L'idéal serait donc de relever toutes les mentions d'une région, d'une origine spatiale, de l'équivalent le plus répandu de la variante et bien d'autres éléments afin de distinguer ces emplois par rapport à un emploi réel, à travers un programme de reconnaissance automatique.

1.6.2 Cas des entités nommées

Dans un second temps, nous devrons identifier les entités nommées, où les régionalismes peuvent se trouver dans un nom d'enseigne, d'évènement, de lieu, de personne… (fig. 6)

Figure 6 : Exemple de tweet contenant une entité nommée

Nous ne considérons pas une entité nommée comme une utilisation réelle puisque le sens premier du régionalisme est remplacé, le locuteur n'a dans ce cas pas d'autres choix que d'utiliser cette appellation et cela ne nous apporte aucune information sur les particularités langagières de cette personne. Nous tenterons donc de détecter automatiquement les entités nommées dans notre corpus. De nombreux auteurs ont tenté d'effectuer une classification exhaustive de ces entités, mais c'est une tâche

difficile et une liste officielle ou admise par tous n'a pas encore été définie (Hatmi 2014). Notre but premier n'étant pas de classer les entités mais simplement de les détecter, nous nous inspirerons de ces catégories uniquement pour établir des règles rendant compte de chaque situation possible.

1.7 Conclusion

Cette première partie nous a permis d'explorer les différentes facettes de notre sujet d'étude. Nous pouvons ainsi aborder la réalisation concrète de notre projet en prenant en compte les connaissances dont nous avons fait l'acquisition. Cela nous permet également de prendre conscience des difficultés que nous pourrons rencontrer.

La deuxième partie de ce projet de recherche consistera à élaborer un corpus et effectuer les analyses nécessaires à l'étude de la vitalité des mots et expressions régionales. Nous élaborerons dans un premier temps une liste de régionalismes pertinents pour nos recherches. Le corpus se fera alors à partir d'une sélection de tweets contenant ces régionalismes. Dans un second temps, il nous faudra construire des outils de reconnaissance d'entités nommées et de métadiscours. Pour cela, nous pourrons nous inspirer du travail d'Evangelia Balouta et l'adapter à notre corpus. Notre dernière tâche sera alors d'analyser la répartition spatiale et la vitalité des véritables utilisations de régionalismes.

2. Données et méthode

2.1 Introduction à la classification automatique

L'apprentissage automatique peut se faire principalement de deux manières : supervisée et non-supervisée. La première est une méthode où l'on prédit des cibles connues, la deuxième est une méthode explicative, qui va faire émerger des relations entre les données et/ou regrouper des éléments selon des similarités.

Notre étude porte sur la classification de tweets en trois catégories connues (réel, métadiscours et entité nommée), il s'agit donc d'une méthode supervisée, plus précisément d'une classification automatique, qui est à différencier du clustering (méthode non supervisée) et de la régression (prédiction de valeurs, supervisée).

La classification automatique est sans doute l'une des techniques les plus étudiées en machine learning mais également la plus difficile à maîtriser. Elle consiste à prédire la classe d'un élément à l'aide d'un échantillon déjà catégorisé. Contrairement à l'apprentissage non supervisé, les différentes classes à prédire sont prédéfinies et l'algorithme doit associer les éléments non catégorisés à ces classes. On utilise alors différents features (aussi appelés descripteurs ou variables prédictives) pour aider l'algorithme dans sa classification, ce sont par exemple des catégories sémantiques, des vecteurs de mots ou encore des données statistiques comme le nombre de tokens par phrase.

Une fois l'étape d'apprentissage réalisée, notre modèle entraîné pourra prédire la classe d'autant de tweets que l'on a à disposition. Il est donc important de constituer un corpus d'entraînement solide qui ne faussera pas les résultats. De ce fait, notre première tâche a été de répertorier de nombreux régionalismes, pour ensuite sélectionner ceux qui permettront d'obtenir des exemples variés et pertinents.

2.2 Élaboration du corpus

2.2.1 Sources de régionalismes

La première source de candidats a bien sûr été le site « francaisdenosregions.com » et le livre de Mathieu Avanzi (2017b). Les régionalismes de ces sources ont été étudiés à l'aide d'enquêtes en ligne, nous savons ainsi qu'ils sont fréquemment employés dans certaines régions. De plus, il est facile de trouver leurs origines géographiques puisque la majorité d'entre eux sont répertoriés dans un des dictionnaires de régionalismes cités précédemment.

Les autres régionalismes proposés pour réaliser l'apprentissage du modèle sont issus de mon expérience, de mes proches ou encore des réseaux sociaux. En effet, plusieurs régionalismes me sont venus spontanément grâce à mes origines et mes déplacements. Je suis née dans la région qu'était autrefois la Picardie, puis j'ai vécu suffisamment de temps dans les villes de Tours, Lyon et Strasbourg pour entendre ou employer moi-même plusieurs expressions atypiques. Ma famille d'origine picarde et mes amis m'ont bien sûr aidé à en découvrir de nouvelles lors de la réalisation de ce mémoire (ex : « ziroux » qui désigne une personne sensible et facilement écœurée, « babache » qui signifie « idiot » en picard). Pour finir, j'ai relevé de nombreux régionalismes grâce à des communautés Facebook ou encore des articles évoquant les différences linguistiques du français. Ces ressources provenant d'Internet m'ont principalement apporté des connaissances sur des expressions du Sud de la France qui m'étaient jusqu'à maintenant inconnues et qui se sont pourtant avérées très présentes sur Twitter (ex : « tarpin » pour dire « beaucoup » ou « très »). Cependant, il était parfois difficile de trouver des sources confirmant leur statut de régionalisme.

2.2.2 Critères de sélection des régionalismes

La sélection des régionalismes qui serviront à l'apprentissage s'est faite selon plusieurs critères :

_ Homonymie : Un régionalisme a été écarté si sa forme pouvait avoir un autre sens répandu nationalement. Par exemple, une « loque », du « brun » ou encore une « poche » (qui désignent respectivement une serpillère, des excréments ou un sac plastique) possèdent un deuxième sens connu de tous qui serait ainsi beaucoup plus

représenté sur Twitter. Une tâche de désambiguïsation bien trop coûteuse aurait été nécessaire pour réaliser le corpus d'apprentissage avec ces régionalismes. Ils pourront cependant servir à tester le modèle d'apprentissage automatique si chaque tweet est analysé manuellement.

_ Graphie : Les régionalismes présentant trop de formes graphiquement différentes ont été retirés, c'est notamment le cas des lexèmes alsaciens « schluck » (gorgée) ou « schlappes » (chaussons) qui se disent plus qu'ils ne s'écrivent. Leur graphie varie régulièrement selon les individus et ils n'ont pas de forme bien définie. Sans doute pour cette raison, ils n'étaient pas non plus très présents sur Twitter. Les régionalismes ayant au maximum trois formes différentes ont cependant été pris en compte (ex : « clenche/clanche »), si ceux-ci étaient suffisamment fréquents.

_ Spécialités : Les noms de spécialités culinaires régionales n'ont pas été pris en compte pour la constitution du corpus. Le but de cette étude étant de représenter la vitalité des régionalismes, il n'est selon moi pas pertinent d'analyser les noms de plats ou de pâtisseries, puisque ceux-ci peuvent être employés par tout francophone. Chaque utilisation est alors une entité nommée, qui ne permet que de montrer la popularité de cette spécialité.

_ Expansion : S'il existe plusieurs manières de désigner un objet ou concept, seules les formes les moins fréquentes ou répandues ont été conservées. Par exemple, « étendoir » face à « étendage » ou « tancarville », bien que la première forme ne soit pas connue de tous, elle est la plus répandue et offrira sans doute plus d'emplois authentiques que de métadiscours. Il a donc été décidé qu'elles ne serviraient pas à l'apprentissage du modèle.

À la suite de cette première sélection, il était important d'étudier des exemples de tweets afin d'observer comment étaient utilisés ces mots et à quelle fréquence. En effet, un premier regard a permis de se rendre compte que certains régionalismes possédaient d'autres sens qu'un dictionnaire n'aurait pas donné, comme par exemple un nom de personnalité ou un détournement de sens (ex : le correcteur « tipp-ex » utilisé pour désigner une personne ayant la peau blanche). Le contexte de ces tweets était également important, selon la date de recherche d'exemples, des événements d'actualité comme les « gilets jaunes » ou la sortie d'une vidéo sur Youtube pouvaient

avoir un impact sur la qualité de l'apprentissage. Cela a également influencé la fréquence d'apparition des régionalismes, certains mots étaient ainsi très présents à certaines périodes mais devenaient presque introuvables un mois plus tard. D'autres mots candidats n'étaient tout simplement jamais mentionnés sur Twitter, soit en raison de leur manque de vitalité, soit parce qu'il s'agissait de mots employés presque exclusivement à l'oral (ex : les helvétismes « witz » pour dire « blague » et « poutzer » qui signifie « faire le ménage »). Une autre raison venait du fait que la plateforme ne se prête pas à certains contextes que l'on rencontre dans la vie quotidienne ; l'expression « être nareux » ou « être ziroux », c'est-à-dire être sensible au niveau de l'hygiène ou être facilement écœuré, a peu de chance d'être employé sur Twitter, si ce n'est pour pointer la valeur régionale de ces expressions. Enfin, certains régionalismes étaient très présents mais ne représentaient que des utilisations réelles, il y avait donc peu d'intérêt à les retenir pour l'apprentissage du modèle. Nous pourrons cependant étudier leur vitalité si le modèle est assez performant pour prédire la classe de nouveaux régionalismes.

2.2.3 Collecte de tweets

Ces premières recherches ont été réalisées à l'aide d'un script Python personnalisé et de la librairie Tweepy. Le programme « getTweets.py » (Annexe 1) se connecte à l'API Twitter à l'aide d'identifiants donnés lors de l'inscription puis recherche chaque mot de la liste de régionalismes. L'API Search standard permet de récupérer une quantité de tweets (ici 100) datant de moins d'une semaine et renvoie un dictionnaire contenant toutes les métadonnées du tweet (informations concernant le compte de l'utilisateur, la localisation du tweet…). L'API ne recherche pas que le mot clé tel quel, la recherche se fait également avec ou sans accents aigus et graves (les régionalismes contenant des trémas ou des accents circonflexes doivent être recherchés avec et sans accent). Les flexions peuvent également être trouvées et la recherche n'est pas sensible à la casse. Le point négatif de cette API est qu'elle cherche le mot clé dans l'intégralité du dictionnaire de métadonnées renvoyé, il peut alors être trouvé dans le nom d'utilisateur sans être dans le tweet en lui-même. Il a donc été nécessaire de vérifier si le tweet contenait le régionalisme. De plus, les retweets ont été supprimés (ce sont des tweets qu'un utilisateur a partagé sur sa page) car cela aurait engendré des doublons inutiles à l'apprentissage. Nous avons programmé l'API pour qu'elle recherche des tweets français, nous n'avons cependant pas réussi à filtrer en fonction du pays. Il faut pour

cela utiliser l'API Stream qui récupère les tweets en temps réel, mais où la recherche de mot clé est plus complexe.

Nous obtenons en sortie de ce programme un fichier par régionalisme, où chaque ligne correspond aux métadonnées d'un tweet, contenues dans un objet de type « json ». Les informations rendues par l'API ne sont pas modifiées afin de conserver le plus de données possibles.

Il est important de noter que toutes les API standards de Twitter ont un taux limite de récupération de tweets, il n'est donc pas rare que l'application se mettent en pause lorsqu'une requête est trop volumineuse. Pour finir, le texte de tweets est tronqué à 140 caractères par défaut, mais puisque Twitter a repoussé sa limite à 280 caractères en 2017, il est important de définir l'argument « tweet_mode= 'extended' » afin d'obtenir des tweets complets.

2.2.4 Annotation

Une fois les régionalismes sélectionnés, la tâche suivante a été de les annoter, c'est-à-dire désigner leur catégorie (R : réel, M : métadiscours, E : entité nommée), afin de faire en sorte que le modèle apprenne de manière supervisée. Une première annotation rapide a été effectuée à l'aide d'un script Python (Annexe 2). Celui-ci classe les tweets en fonction de mots clés, choisis subjectivement par rapport au contenu des tweets relevés, ces mots ne correspondront donc sans doute pas à d'autres données. Le programme cherche également si le régionalisme du tweet est entouré de guillemets, ou s'il commence par une arobase, un dièse ou par « http ».

Ces éléments clés ont été choisis à la suite de différentes observations. Tout d'abord, un tweet comportant du métadiscours aura plus de chance de contenir l'équivalent « standard » du régionalisme, c'est-à-dire la variante la plus employée, notamment en France. De plus, le verbe « dire » et la mention d'une zone géographique seront souvent présents et le régionalisme sera fréquemment entouré de guillemets. Dans le cas des entités nommées, il est plus difficile de prévoir un contexte qui correspondra à tout exemple, nous avons ainsi relevé des mots clés qui paraissaient être pertinents dans les tweets dont nous disposions. Les signes arobase et « http » sont des sources sûres d'entités nommées mais le dièse est un cas plus complexe. En effet, un mot dièse peut être intégré dans une phrase, sa catégorie dépendra alors du contexte,

mais il peut également être écrit seul en fin de tweet. Pour une première annotation, nous avons considéré que la présence d'un dièse était une marque de métadiscours puisque dans la majorité des cas, la personne met en valeur l'utilisation de ce mot pour sa valeur régionale.

Le script d'annotation a pu être amélioré à l'aide du logiciel de traitement de texte « Antconc ». Grâce à la sortie de la première annotation effectuée, nous avons pu séparer les données en trois fichiers, un pour chaque catégorie. Une fois importés dans le logiciel, ces fichiers nous ont permis d'observer plus simplement le contexte proche des régionalismes à l'aide du concordancier. Nous avons également pu relever d'autres indices grâce à l'étude des collocations et à l'onglet « Keyword List », notamment en comparant chaque fichier à celui des utilisations réelles. Le script d'annotation rapide a alors été relancé plusieurs fois avec différents mots clés jusqu'à obtenir une annotation satisfaisante.

Cette annotation restait tout de même approximative, il a donc été important de réaliser une vérification de chaque tweet manuellement, afin d'obtenir le meilleur échantillon d'apprentissage possible. Des ajustements ont ainsi été réalisés par rapport à la sortie de l'annotation rapide. De plus, quelques tweets non pertinents pour l'apprentissage ont été annoté « O » (pour Out) afin d'être supprimés à l'étape suivante. Il s'agissait de doublons ou de tweets inclassables sans connaissance de la conversation en cours.

Afin de faciliter la ré-annotation, des fichiers ne contenant que le texte des tweets ont été créés. Pour rappel, les fichiers au format csv étaient jusqu'à présent constitués de données au format « json » comportant toutes les métadonnées du tweet, c'est-à-dire les informations concernant l'utilisateur, les coordonnées géographiques du tweet ou encore le nombre de retweets. Il est également possible de récupérer les entités du texte, notamment le nombre de liens hypertextes, les mentions d'utilisateurs (commençant par une arobase) et les mots dièses. Lors de la création des fichiers, nous avons relevé ces informations afin de ne pas avoir recours une seconde fois aux fichiers contenant les dictionnaires (dossier « train_corpus_full ») lors de l'élaboration du modèle d'apprentissage. Les fichiers obtenus en sortie du script « create_training_file.py » comportent donc une annotation, le texte du tweet et le nombre de liens hypertextes, de mentions d'utilisateurs et de mots dièses, séparés par

des virgules. Le choix de garder ces entités repose sur le fait que certains tweets en contiennent un nombre très important. Les mentions d'utilisateurs sont le signe d'une conversation entre plusieurs personnes, ce qui peut signifier un langage plus oral et familier. Les mots dièses rendent souvent compte d'un langage bien particulier, associé à une volonté d'être présent sur les réseaux sociaux et de mettre en valeur certains mots, cela peut être dans un but marketing ou encore pour attendre une réponse d'autres utilisateurs. Enfin, les tweets automatiques évoquant l'actualité, la météo ou le trafic routier contiennent généralement un lien hypertexte, il est donc intéressant de relever la présence de ces derniers.

Après avoir vérifié manuellement chaque tweet, nous avons pu mettre en place le modèle de classification automatique.

2.2.5 Régionalismes retenus pour l'apprentissage

Ci-dessous se trouve la liste des régionalismes sélectionnés pour l'apprentissage du modèle ainsi que leur description.

Tancarville :

« Tancarville » est très intéressant à étudier dans le sens qu'il est régulièrement employé dans des cas de métadiscours ou d'entité nommée. En effet, cette appellation du séchoir à linge peut surprendre lorsqu'on la découvre, ce qui permet d'obtenir des tweets pertinents tels que ceux cités précédemment (p. 29). Mais le principal atout de ce régionalisme réside dans le fait qu'il s'agit également d'une commune en Seine-Maritime. De nombreuses mentions de ce mot font ainsi référence aux structures de la commune, notamment son pont, qui fut sans doute l'inspiration pour les créateurs de la marque de séchoir à linge « tancarville ». Cette marque est aujourd'hui devenue un mot à part entière pouvant être entendu sur une zone très étendue du Nord-Ouest. En raison de sa récence et de son statut de marque, ce régionalisme n'est pas présent dans le dictionnaire des régionalismes de France (DRF) mais il a fait l'objet d'une étude dans le cadre du projet « Français de nos régions » (Avanzi 2016a).

Catégorie	Nombre de tweets
Entité nommée	28
Métadiscours	8
Réel	9
Total	45

Tableau 2 : Répartition des tweets après l'annotation (tancarville)

Etendage/Etandage :

Tout comme « tancarville », le régionalisme « étendage » fait référence à un séchoir à linge. Son utilisation est plus restreinte, elle se situe principalement dans la région anciennement appelée Rhône-Alpes et en Suisse francophone (*Ibid.*). La prise en compte d'une deuxième forme « étandage » fut nécessaire compte tenu du nombre d'occurrences trouvées sur Twitter. Bien que l'on ait tendance à dire que le mot « étendoir » est considéré comme la norme connue de tous, des tweets nous prouvent que certaines personnes ne connaissaient pas cette appellation, qui était entièrement remplacée par la forme « étendage » dans leur vocabulaire. Ce phénomène se retrouve dans des tweets comme :

« […] J'habite à Montélimar, j'ai dit étendage une fois au boulot ils ont vu flou. Étendoir c'est quoi ce mot ? Un étendage point ».

Nous avons cependant rencontré un problème avec ce régionalisme par rapport aux tweets automatiques. En effet, un compte d'utilisateur était consacré à la génération automatique de sujets d'actualité et il s'est avéré que « étendage » était un mot récurrent sur ce compte. Nous avons décidé d'annoter ces tweets manuellement pour qu'ils soient retirés du corpus mais cela pourrait être amélioré en ajoutant une condition sur le nom d'utilisateur ou sur la source du tweet si ce phénomène se retrouve dans plusieurs régionalismes (la source correspond à la plateforme sur laquelle le tweet a été écrit, elle sera souvent atypique si le tweet est automatique).

Catégorie	Nombre de tweets
Entité nommée	0
Métadiscours	4
Réel	8
Total	12

Tableau 3 : Répartition des tweets après l'annotation (étendage et étandage)

Chicon :

Le régionalisme « chicon » est utilisé principalement en Belgique francophone pour désigner une endive française. Tandis que dans cette zone géographique, l'endive est une sorte de salade frisée. La région Hauts-de-France emploie également le régionalisme « chicon » mais le terme n'est pas autant répandu (Avanzi 2017c). Les tweets contenant ce mot sont plutôt variés mais nous obtenons un grand nombre d'entités nommées car de nombreux exemples mentionnent l'utilisateur « @endive_chicon », certains tweets trop répétitifs ont donc été retirés du corpus. Cette forme est présente dans le DRF.

Catégorie	Nombre de tweets
Entité nommée	23
Métadiscours	6
Réel	9
Total	38

Tableau 4 : Répartition des tweets après l'annotation (chicon)

Chocolatine :

« Chocolatine » était un régionalisme indispensable à cette étude. Cette variante de « pain au chocolat » génère énormément de débats depuis une polémique de Jean-François Copé, où celui-ci évoquait un pain au chocolat en 2012. Cela a grandement contribué à mettre en lumière les régionalismes, le nombre d'articles à ce sujet a augmenté et les populations se sont de plus en plus intéressées à leurs particularités linguistiques afin d'affirmer leurs origines. « Chocolatine » génère toujours autant de débats en raison de sa zone d'utilisation très étendue. En effet, cette forme est utilisée approximativement dans les régions de la Nouvelle-Aquitaine, de l'Occitanie et du

Québec (Avanzi 2016b). De plus, si les régionalismes ont dans la majorité des cas un équivalent que tout français peut considérer comme « standard », ici chaque camp semble rejeter l'appellation de l'autre, et c'est ce qui engendre ce « jeu » pour savoir qui sortira vainqueur. Ce débat est fréquemment retrouvé dans les tweets, ces données variées permettent d'améliorer la classification du métadiscours puisqu'elles sont applicables à tout autre régionalisme. D'autres variantes comme croissant au chocolat, petit pain au chocolat ou couque au chocolat existent mais notre programme ne permet pas de gérer le domaine polylexical pour le moment. Cette forme est présente dans le DRF.

Catégorie	**Nombre de tweets**
Entité nommée	3
Métadiscours	45
Réel	11
Total	59

Tableau 5 : Répartition des tweets après l'annotation (chocolatine)

Clenche/Clanche :

« Clenche » sert à désigner une poignée de porte dans une zone assez étendue du Nord. Ce régionalisme peut être entendu en Normandie, en Lorraine, en Belgique mais aussi à moindre mesure dans la région Hauts-de-France (Avanzi 2017d). Nous avons également pris en compte la forme « clanche » en raison de sa fréquence d'apparition et des quelques tweets mentionnant cette différence de graphie, que nous avons classé comme des cas de métadiscours. De plus, de nombreux tweets comportant ce régionalisme contenaient des énumérations de régionalismes de certaines régions, ils apportaient donc de nombreux indices pertinents pour la classification du métadiscours. « Clenche » désignait à l'origine une partie du loquet de la poignée pour finalement désigner la poignée dans son ensemble. La forme verbale de ce régionalisme existe également mais elle n'a pas été explicitement recherchée avec toutes ses flexions. De plus, le verbe « clancher » désigne au Québec l'action de se faire battre ou de réaliser une action rapidement. Nous n'avons pas retiré ces exemples puisqu'ils ne détérioraient pas l'apprentissage et que cette expression aurait

pu être retrouvée en Europe. Ces formes sont présentes dans le DRF et la Chasse aux belgicismes.

Catégorie	Nombre de tweets
Entité nommée	6
Métadiscours	53
Réel	61
Total	120

Tableau 6 : Répartition des tweets après l'annotation (clanche et clenche)

Drache :

L'équivalent du régionalisme « drache » est une averse, ce régionalisme est également utilisé comme un verbe pour déclarer qu'il pleut fortement mais nous n'avons pas recherché les flexions de cette forme verbale. Emprunté au flamant « draschen », ce régionalisme est principalement utilisé en Belgique et dans les villes françaises limitrophes à ce pays. Cependant, selon Mathieu Avanzi (*Ibid.*), « drache » serait de plus en plus utilisé en dehors de ces régions. Nous avons retrouvé un grand nombre de tweets comportant cette forme mais le nombre de tweet géolocalisé étant faible, nous ne pourrons sans doute pas observer son expansion à grande échelle. Les tweets que nous avons observés contenaient en majorité des utilisations authentiques de « drache », mais nous l'avons conservé en raison des quelques exemples de métadiscours et d'entités nommées visibles.
Ce régionalisme est présent dans la Chasse au belgicisme.

Catégorie	Nombre de tweets
Entité nommée	1
Métadiscours	8
Réel	65
Total	74

Tableau 7 : Répartition des tweets après l'annotation (drache)

Carabistouille :

« Carabistouille », qui signifie « mensonges » ou « balivernes », est un régionalisme qui fut très présent dans la presse et les réseaux sociaux en 2018. Cela est dû au

Président de la République Emmanuel Macron, qui a utilisé ce terme lors d'une interview télévisée. Cet usage a engendré beaucoup d'étonnement ou de moqueries de la part des français car « carabistouille » était principalement utilisé en Belgique francophone et en Hauts-de-France (Avanzi 2018a). Aujourd'hui encore, de nombreux tweets évoquant ce régionalisme mentionnent Emmanuel Macron. Ainsi nous considérons que dans la majorité de ces cas, il s'agit de métadiscours. Nous avons tout de même relevé plusieurs cas d'usage authentique, ce qui nous permettra d'analyser si l'utilisation de ce régionalisme s'est répandue dans le reste de la francophonie suite à cette mise en lumière médiatique.
Ce régionalisme est présent dans le Dictionnaire des belgicismes.

Catégorie	**Nombre de tweets**
Entité nommée	2
Métadiscours	19
Réel	23
Total	44

Tableau 8 : Répartition des tweets après l'annotation (carabistouille)

Ducasse et frairie :

Tout comme « frairie », « ducasse » désigne une fête de village ou une foire annuelle. Il existe de nombreuses appellations de cet évènement à travers la France, la Belgique et la Suisse. Pour nos recherches, nous avons sélectionné les variantes « fest-noz », « frairie », « feria », « messti », « ducasse » et « kilbe » mais seules « ducasse » (Hauts-de-France) et « frairie » (Poitou-Charentes) ont retenu notre attention. Pour plus d'informations à ce sujet, un article très complet est disponible sur le site « francaisdenosregions.com », celui-ci décrit toutes les variantes citées ci-dessus et leurs origines (Avanzi 2017e). Le terme « kermesse » n'a pas été sélectionné en raison du sens qu'il possède en France métropolitaine, c'est-à-dire une fête annuelle organisée par les établissements scolaires. L'appellation « feria » était quant-à-elle très fréquente, mais ne comportait pas assez de cas de métadiscours ou d'entités nommées. De manière générale, les régionalismes sélectionnés n'étaient pas fréquemment utilisés dans un contexte de métadiscours, il était cependant intéressant

de relever les cas d'entités nommées. Dans un premier temps, nous nous sommes interrogés sur l'annotation que devaient avoir des tweets comme ci-dessous :

« 🚲 🏀 La Ducasse de la rue de Lille c'est du 17 au 19 mai au Parc du Perroy @VilleBethune.➡ De nombreuses animations seront au rendez-vous pour le plus grand plaisir des grands et des petits. https://t.co/5UOMERyXTl »

« Ducasse d'Ath: Ce week-end aura lieu le « 42e week-end de la Fête des Mères » - Blog Sudinfo-Ath https://t.co/vVI3uC2jQN »

Nous avons finalement considéré ces exemples comme des utilisations authentiques de régionalismes plutôt que des entités nommées. Cependant, un cas particulier a été fréquemment relevé, il s'agit du célèbre chef cuisinier Alain Ducasse.

« Romain Meder Chef du restaurant Alain Ducasse au Plaza Athénée remporte le prix du meilleur plat @Taste_ofParis 👍👍 https://t.co/iosAEQuwZ8 »

De nombreux tweets comme celui-ci mentionnaient son restaurant ou sa participation à TopChef, il a donc été intéressant d'utiliser ces exemples pour améliorer la reconnaissance des entités nommées. Ce régionalisme est présent dans le DRF.

Catégorie	**Nombre de tweets**
Entité nommée	32
Métadiscours	1
Réel	17
Total	50

Tableau 9 : Répartition des tweets après l'annotation (ducasse)

Catégorie	**Nombre de tweets**
Entité nommée	0
Métadiscours	1
Réel	4
Total	5

Tableau 10 : Répartition des tweets après l'annotation (frairie)

Petitou, minot, tchot et petiot :

Nous nous sommes ensuite focalisés sur les appellations d'un enfant. Le DRF cite de nombreux régionalismes synonymes de « gamin » (tiot, chétit, drôle, gone, minot, petit, pitchoun), ou d'enfant en bas âge (petitou, gari, mousse, nin, niston, pitchoun). Pour notre corpus d'apprentissage, le régionalisme « petitou » - ou « petitoune » au féminin- a été retenu, ainsi que « minot(e) », tchot(e) et « petiot(e) ». La forme « tchot » est une graphie de plus en plus fréquente du régionalisme « tiot » (écrit également « tchiot »), originaire des Hauts-de-France. Selon le DRF, le régionalisme « petitou » était présent sur une large zone couvrant approximativement l'actuelle Occitanie. Cependant, nous avons relevé peu d'exemples de ce régionalisme sur Twitter, ce qui peut être le signe d'une perte de vitalité. La forme « petiot » n'est pas définie comme un régionalisme par le DRF, mais nous l'avons conservé afin d'observer son aire d'utilisation actuelle. Nous faisons l'hypothèse que cette forme serait plus employée dans le Nord de la France, en raison de sa proximité acoustique avec le régionalisme « tiot ». De plus, « Petiot » étant un nom de famille très répandu, nous avons pu relever de nombreux cas d'entités nommées, utiles à l'entraînement de notre classifieur. Enfin, le régionalisme « minot » était initialement présent en Provence-Alpes-Côte-d'Azur. Nous avons rencontré de nombreux exemples de cette forme sur Twitter, la raison principale semble être l'existence d'un club junior de football nommé « Les Minots de l'OM ». Nous ne savons cependant pas si c'est cette organisation qui a influencé la vitalité de ce régionalisme ou si son usage ne s'est jamais atténué au fil du temps.

Catégorie	**Nombre de tweets**
Entité nommée	12
Métadiscours	0
Réel	30
Total	42

Tableau 11 : Répartition des tweets après l'annotation (petiot)

Catégorie	Nombre de tweets
Entité nommée	7
Métadiscours	0
Réel	60
Total	67

Tableau 12 : Répartition des tweets après l'annotation (minot)

Catégorie	Nombre de tweets
Entité nommée	2
Métadiscours	0
Réel	4
Total	6

Tableau 13 : Répartition des tweets après l'annotation (tchot)

Catégorie	Nombre de tweets
Entité nommée	3
Métadiscours	0
Réel	6
Total	9

Tableau 14 : Répartition des tweets après l'annotation (petitou)

Natel et GSM :

« Natel » et « GSM » désignent tous les deux un téléphone portable. Le premier est utilisé en Suisse francophone tandis que le deuxième est majoritairement employé en Belgique francophone. Comme pour les autres régionalismes de ces pays, les études de Mathieu Avanzi ont mis en valeur qu'ils étaient également utilisés par certaines personnes des régions françaises limitrophes (Avanzi 2017f). « GSM » est le sigle d'un standard international de téléphonie sans fil : Global System for Mobile Communications. « Natel » est quant à lui une marque de téléphone mobile uniquement développée en Suisse. Les tweets contenant ces régionalismes révélaient principalement des emplois authentiques. Le nombre élevé de cas d'entités nommées provenait en majorité de noms d'utilisateurs mais quelques exemples de métadiscours pertinents ont tout de même été trouvés.

Catégorie	Nombre de tweets
Entité nommée	16
Métadiscours	1
Réel	29
Total	47

Tableau 15 : Répartition des tweets après l'annotation (gsm)

Catégorie	Nombre de tweets
Entité nommée	2
Métadiscours	4
Réel	27
Total	33

Tableau 16 : Répartition des tweets après l'annotation (natel)

Pour conclure cette section, ci-dessous se trouve un graphique (fig. 8) récapitulatif des quantités de régionalismes par classe.

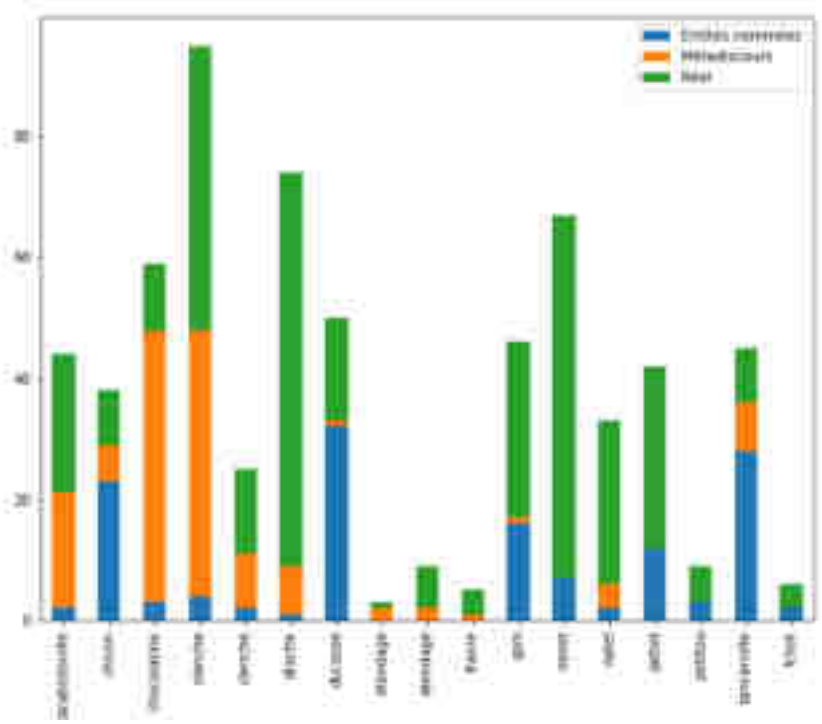

Figure 7 : Nombre de tweets par régionalisme et type d'annotation

Notre corpus contient finalement 137 tweets catégorisés en entités nommées, 150 tweets catégorisés en métadiscours et 363 tweets contenant de réelles utilisations de régionalismes, pour un total de 650 tweets.

2.3 Pré-traitements

2.3.1 Importation des données

Pour qu'un modèle puisse apprendre à classifier des données, il faut tout d'abord effectuer un pré-traitement. La première tâche consiste à charger les données et observer les proportions des différentes classes, c'est aussi à cette étape que l'on peut combler ou supprimer d'éventuelles données manquantes. La seconde partie représente le pré-traitement ; c'est ici que nous définissons les différents descripteurs qui serviront à la classification. Chaque type de descripteur est traité afin d'obtenir un ensemble de données facilement interprétable et pertinent pour l'apprentissage. Une fois les différents pré-traitements définis, des chaînes de traitement, communément appelées « pipeline », sont associés aux descripteurs afin de réaliser une vectorisation ou un encodage binaire de manière séquentielle. Ces chaînes de traitement permettent d'effectuer des traitements et de transformer les données pour les rendre plus compréhensibles pour le modèle. Enfin, la chaîne de traitement applique un classifieur sur les données. Chacune de ces étapes est appliquée sur l'échantillon d'entraînement puis sur celui de test. Pour finir, les prédictions du (ou des) modèle(s) sont analysées afin d'être éventuellement améliorées. Nous pouvons alors effectuer des opérations comme l'optimisation des hyperparamètres du modèle ou la sélection des descripteurs les plus pertinents pour obtenir de meilleurs résultats.

Dans notre cas d'étude de tweets, les données au format CSV sont tout d'abord importées et stockées dans un « dataframe », ces tableaux permettent de réaliser de nombreux traitements de manière intuitive et visuellement accessible. Les données sont alors réparties dans des colonnes où chaque ligne est associée à un index, ce qui rend la navigation très simple. Comme expliqué précédemment, les données d'entrée comportent déjà 5 colonnes : type, tweet, hashtag, user_mention et url. Une sixième colonne « regionalism » est créée à partir du nom du fichier afin de récupérer la forme du régionalisme utilisée pour la recherche initiale dans l'API Twitter.

La quantité de tweets dans chaque classe est représentée en graphique (fig. 8) afin de repérer rapidement les inégalités. Cela permet notamment de redimensionner la proportion des données si nécessaire.

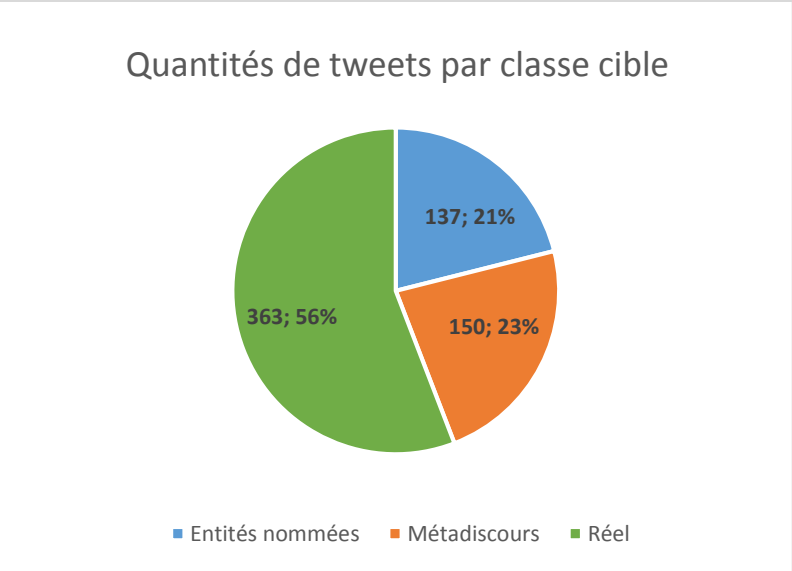

Figure 8 : Répartition des classes dans le corpus d'apprentissage

Ici, nous observons que notre corpus est déséquilibré, mais cette répartition a été conservée puisque, en prenant en compte les régionalismes non sélectionnés, la majorité des tweets représente un cas authentique d'utilisation de régionalisme. De cette manière, la modèle apprendra que la classe « réelle » a plus de chance d'apparaître.

Les graphiques ont été générés à l'aide des librairies « pandas » et « matplotlib ». Il est également intéressant d'observer la quantité de liens hypertextes, de mots dièses et de mentions d'utilisateurs en fonction du type de régionalisme. La méthode « describe » permet d'obtenir toutes les informations statistiques principales telles que la moyenne, l'écart-type ou le maximum de ces trois entités.

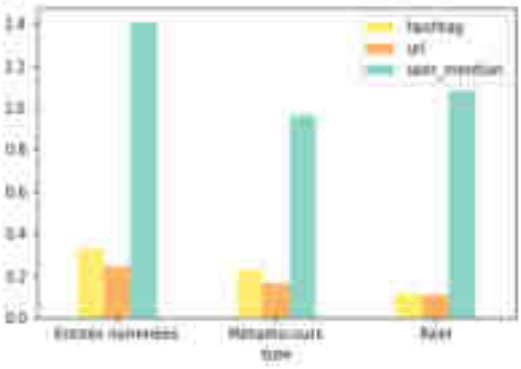

Figure 9 : Moyenne d'entités par classe cible

Nous remarquons alors que les liens hypertextes et les mentions d'utilisateurs sont légèrement plus nombreux dans les cas d'entités nommées (fig. 9). Par ailleurs, la classe réelle semble contenir moins de liens hypertextes et de mots dièses, ces tendances pourront donc sans doute aider le classifieur dans sa tâche.

2.3.2 Choix des librairies

Pour le prétraitement des tweets, nous commençons par l'analyse du régionalisme seul. Pour cela nous utiliserons la librairie « spaCy », qui possède un outil de reconnaissance d'entités nommées entraîné sur des corpus de différentes langues. La librairie possède trois modèles pour la langue française (« Spacy Documentation ») : « fr_core_news_md » est le modèle le plus volumineux, il a été entraîné grâce aux corpus Sequoia et WikiNER, qui comportent tous deux des textes d'actualités médiatiques. Ce corpus permet de réaliser tous les traitements disponibles sur spaCy (tokenisation, dépendances, lemmatisation, entités nommées…). L'exactitude de ce modèle est très élevée sur un corpus bien structuré mais il est très probable que ses performances soient moindres sur un corpus provenant de réseaux sociaux tel que le nôtre.

La librairie spaCy possède également le modèle « fr_core_news_sm », celui-ci est identique au premier à l'exception qu'il est moins volumineux et prend donc moins de temps à importer. Il obtient cependant de moins bons résultats que le premier modèle, c'est pour cela que nous avons décidé d'utiliser le premier modèle. Il existe un également un modèle de langue française très peu volumineux, qui est en général uniquement utilisé pour la tokenisation. Nous avons également essayé le modèle de

reconnaissance d'entités nommées multilingue, mais celui-ci ne rencontrait pas d'aussi bons résultats que le modèle français.

En ce qui concerne le traitement des tokens, nous avons réalisé des essais avec les librairies « NLTK » et « spaCy ». Leurs résultats étaient similaires mais notre choix s'est porté sur spaCy pour sa facilité d'utilisation et ses performances en reconnaissance d'entités nommées. Nous avons toutefois utilisé les « stopwords » ou « mots-vides » de NLTK pour nos essais car ceux-ci sont moins vastes que ceux disponibles dans spaCy. Enfin, il est important de noter que NLTK possède une méthode « TweetTokenizer » (« TweetTokenizer » 2019) mais celle-ci n'est pas adaptée au français, nous ne l'avons donc pas utilisée.

2.3.3 Annotation du régionalisme

La première étape de l'analyse est de localiser le régionalisme dans le tweet à partir de la colonne « regionalism ». Suite à la tokenisation du tweet, l'indice du token correspondant au régionalisme est récupéré afin de vérifier plusieurs conditions. Tout d'abord, le token commence-t-il par une arobase ? Cette question est importante car tout régionalisme commençant par une arobase est considéré comme une entité nommée puisqu'il s'agit d'un nom d'utilisateur (ex : « @endive_chicon »). Puis, nous vérifions si le régionalisme est contenu dans un lien hypertexte grâce à la méthode « like_url » de spaCy. Cela est également considéré comme un cas d'entité nommée mais nous annotons, dans un premier temps, chaque condition par une lettre distincte afin d'observer les quantités obtenues. Les liens sont ainsi annotés « U » pour « url » et les arobases sont annotées « A ».

Les régionalismes restants sont annotés selon l'analyse d'entité nommée de spaCy. Si un régionalisme est reconnu comme étant une entité nommée, cette libraire lui assigne un type (personne, lieu…) et un code pour connaître la position du token dans l'entité (« B » si le token commence l'entité, « I » s'il est dans une entité, « O » s'il n'est pas dans l'entité).

« @Daarjeeling @ManonAubryFr Et bien comme d'hab: on vend le pont d'Avignon, et pour 900 000 il faudra rajouter celui de ***Tancarville*** *non? »*

Ici, Tancarville est une entité nommée uni-lexicale, nous sommes donc dans un cas où le régionalisme commence l'entité.

*« En route vers le zéro déchet avec la famille Zéro **Carabistouille** ! […] »*

Dans cet exemple, l'entité nommée est poly-lexicale et le régionalisme se situe à l'intérieur de celle-ci.

Après avoir réalisé quelques tests, nous avons estimé que le type de l'entité nommée n'était pas assez fiable, nous n'avons donc conservé que la position du token dans l'entité. L'annotation continue ensuite avec l'analyse des tokens situés avant et après le régionalisme afin de repérer la présence de guillemets grâce à la méthode « is_quote », cela permet d'obtenir un indice de métadiscours. Nous relevons également la présence d'un dièse devant le régionalisme. La recherche ne se fait que sur le dièse car lors de la tokenisation, spaCy considère qu'un mot dièse correspond à deux tokens. Ces deux indices sont respectivement annotés « Q » (quote) et « H » (hashtag). Exemple de régionalisme annoté « Q » :

*« @DevedeuxJu Ah ! « **carabistouille** » longtemps que je n'avais pas entendu ça »*

Exemple de régionalisme annoté « H » :

« Tout comme les irremplaçables #flot, #clanche et #cornet, #clarteux mériterait d'entrer dans le dictionnaire https://t.co/nK50DCaGzT »

Dans le cas où le régionalisme n'est pas repéré dans le tweet, nous marquons le tweet d'une annotation « NF ». Dans notre corpus d'apprentissage, le nombre de « NF » est nul puisque la présence de chaque régionalisme a été vérifiée dès la collecte de tweets via l'API. Cependant, pour prédire la classe de nouveaux tweets, nous avons utilisé des tweets que nous avions collectés selon différentes méthodes, sans réaliser systématiquement cette première vérification. L'annotation « NF » nous permet donc de relever ces tweets potentiellement non filtrés et de les supprimer. De plus, lors de l'apprentissage du modèle, cette annotation nous a permis de déceler et corriger certaines failles dans notre traitement.

Afin de vérifier la pertinence des indices relevés, nous analysons ensuite leur fiabilité par rapport à l'annotation manuelle réalisée en amont. De cette manière, nous observons que les régionalismes reconnus comme étant au début d'une entité nommée sont moins fiables que ceux étant à l'intérieur d'une entité. En effet, seulement 41% des annotations « B » correspondent réellement à une entité nommée

contre 69.5% pour les annotations « I ». Dans la majorité des cas, ces erreurs sont causées par une ou plusieurs majuscules.

*« **Tancarville** c'est un joli mot »* (cas de métadiscours reconnu comme étant le début d'une entité nommée)

*« J'aurais donc passé deux heures grand max en jeans aujourd'hui, je me suis pris une MEGA **DRACHE**, je suis rentrée et de nouveau en pyjama […] »* (cas réel reconnu comme faisant partie d'une entité nommée)

Cependant, la librairie spaCy a l'avantage d'oublier peu d'entités nommées puisque l'annotation « O » ne correspond que dans 3% des cas à une entité nommée. Il s'agit dans la majorité des cas d'oublis de majuscule ou de nom de famille prononcé sans prénom, ce qui rend la reconnaissance plus difficile.

Comme cela était attendu, les annotations « A » et « Q » ont une exactitude de 100%. En effet, l'arobase fait toujours référence à un utilisateur (et donc une entité nommée) et la présence de guillemets de chaque côté du régionalisme correspond systématiquement à un signe de métadiscours. Enfin, l'annotation H apparaît comme un bon indice de métadiscours puisqu'il atteint les 83% d'exactitude. Cet indice doit cependant être pris en compte avec précaution puisqu'un mot dièse peut également correspondre à un emploi réel ou une entité nommée :

*« Le #Pontde**Tancarville** souffle cette #année ses 60 bougies [...] »* (entité nommée contenue dans un mot dièse)

*« Finalement, #Macron aura pris la même #**drache** que #Hollande lors des #cérémonies officielles du #8Mai1945 »* (emploi réel contenu dans un mot dièse)

Ces annotations sont ensuite associées à des colonnes, où chaque valeur est remplacée par 0 ou 1 (Faux ou Vrai). Nous obtenons ainsi les colonnes « maybe_in_EN » pour l'annotation « I », « maybe_EN » pour « B » et « EN » pour les annotations « A » et « U » (bien qu'aucun régionalisme de notre corpus n'ait été trouvé dans un lien hypertexte). Les colonnes « reg_quote » pour l'annotation « Q » et « reg_hashtag » pour l'annotation « H » sont également créées.

2.3.4 Traitements textuels

Le texte des tweets est ensuite nettoyé, plusieurs caractères sont retirés afin d'améliorer la tokenisation et l'apprentissage. Les liens hypertextes, les mentions d'utilisateur et les dièses sont supprimés du texte original. En effet, les liens et les mentions sont extrêmement nombreux et n'apportent aucune information pertinente pour l'apprentissage des vecteurs de mots. De plus, la fréquence de ces entités a déjà été relevée, il n'est donc plus nécessaire de les conserver. Pour les mots dièses, seul le dièse est retiré car, comme nous l'avons vu précédemment, un mot dièse est segmenté en deux tokens par spaCy et le terme restant peut être intégré dans une phrase et donc apporter un sens ou une information importante. Bien sûr, le fait qu'un régionalisme soit précédé par un dièse est important mais cette information est déjà visible dans les colonnes précédemment définies. Les tweets sont également normalisés, c'est-à-dire que les caractères répétés plus d'une fois sont retirés, tout comme les chiffres et les espaces multiples. De plus, les formes abrégées "jdis" et "koi" sont respectivement corrigées en "je dis" et "quoi", puisque ces formes sont importantes pour la reconnaissance des métadiscours. Des espaces sont également ajoutés autour des parenthèses et des points pour améliorer la tokenisation. Enfin, le régionalisme est supprimé du texte original afin de faire en sorte que l'algorithme classifie en fonction du contexte plutôt que du régionalisme cible. Nous devrions ainsi pouvoir prédire la classe de tweets comportant de nouveaux régionalismes, absents des données d'apprentissage.

Ainsi, *« @Brevesdepresse et tout cela 15 jours avant les élections européennes...\nj'ai hâte de voir si #Macron va arriver à ses fins, avec toutes ses #carabistouille, le 26 Mai...\nVive le #RIC... »* devient :

« et tout cela 0 jours avant les élections européennes. j'ai hâte de voir si Macron va arriver à ses fins, avec toutes ses, le 0 Mai. Vive le RIC. ».

2.3.5 Tokenisation

L'étape suivante consiste à vectoriser le tweet, c'est-à-dire transformer ses tokens en vecteurs afin que le modèle d'apprentissage puisse les prendre en compte et les associer entre eux. Il existe plusieurs types de vectorisation que nous développerons plus loin. Dans tous les cas, il est nécessaire de définir comment le tweet sera normalisé et quels éléments vides de sens (stopwords) seront retirés, puisque chaque

token conservé formera un descripteur pour la classification. En ce qui concerne la normalisation, plusieurs choix s'offrent à nous. Nous pouvons simplement segmenter le tweet en tokens ou nous pouvons effectuer une normalisation de type lemmatisation ou désuffixation. La première méthode consiste à rassembler différentes formes fléchies en une seule forme canonique. En d'autres termes cette méthode remplace le token par son entrée lexicale dans le dictionnaire. Cela permet d'obtenir moins de descripteurs mais cette méthode peut être sensible aux abréviations comme dans le langage SMS. La désuffixation (ou racinisation) permet quant-à-elle de réduire un token à sa racine. De la même manière, nous obtenons une seule forme pour plusieurs formes fléchies, ce qui diminue le nombre de descripteurs à prendre en compte. Cette méthode serait plus adaptée lorsqu'un corpus est mal orthographié mais ce fait n'est pas garanti pour un corpus français. Nous comparerons donc les résultats de classification avec une désuffixation et avec une lemmatisation. Le choix des « mots-vides » à prendre en compte est également important puisqu'ils définissent quels tokens seront retirés du corpus, en raison de leur manque de pertinence sémantique. Notre analyse portera donc également sur la prise en compte de mots-vides, dans le but d'observer leur impact sur le traitement de corpus.

2.3.6 Vectorisation

Pour qu'un modèle puisse apprendre à catégoriser un texte brut, nous avons besoin de transformer ce texte en vecteurs. Il existe pour cela deux méthodes principales sur scikit-learn : CountVectorizer et Tf-idfVectorizer. Ces deux fonctions possèdent plusieurs attributs ; le premier définit comment le texte est segmenté, c'est à cette étape qu'il est possible d'ajouter une normalisation. Il est ensuite possible d'ajouter une liste de mots-vides qui ne feront pas partie des descripteurs. La tokenisation peut également se faire en suivant un pattern de n-grammes, c'est-à-dire des combinaisons de « n » éléments. Enfin, les attributs « min_df » et « max_df » définissent les fréquences minimales et maximales qu'un descripteur doit avoir dans un texte ou document, il s'agit d'une proportion d'apparition si le chiffre donné est décimal, autrement le nombre de texte dans lequel le descripteur doit apparaître est pris en compte. Les deux méthodes de vectorisation sont différentes, le CountVectorizer ne fait que vectoriser le nombre d'occurrence de chaque descripteur tandis que le Tf-idf prend en compte la fréquence mais aussi la rareté des tokens ou des lemmes. Cette méthode associe un poids en fonction de l'importance d'un descripteur dans un texte

par rapport aux autres textes. Notre but sera ainsi de comparer les résultats obtenus avec les méthodes CountVectorizer et Tf-idfVectorizer, tout en modulant leurs paramètres, afin de choisir la plus adéquate.

Deux autres descripteurs sont ensuite créés afin d'observer la structure graphique des tweets : la longueur des tweets et le nombre de majuscules par rapport à la longueur des tweets. Nous nous sommes portés sur ces choix grâce à nos observations faites lors de l'annotation manuelle. En effet, de nombreux cas d'entités nommées ont été relevés dans un contexte bien structuré en matière de ponctuation et de majuscules, il s'agissait régulièrement de tweets publicitaires ou d'informations concernant le trafic routier. Certains tweets sont également rédigés entièrement en majuscule et ces données peuvent éventuellement permettre au modèle d'apprentissage de relier ce phénomène à une classe. Ces deux descripteurs sont vectorisés grâce à la méthode « DictVectorizer » de scikit-learn qui génère un dictionnaire comportant le nom du descripteur et sa valeur dans le but de les rendre interprétables par un classifieur automatique. La méthode « DictVectorizer » est également utilisée sur les colonnes précédemment décrites qui permettront de reconnaître la nature du régionalisme.

En raison du nombre insuffisant de descripteurs consacrés à la détection de métadiscours, il a été nécessaire de développer un pré-traitement ayant pour but de créer un score en fonction de la présence de lemmes-clés fréquents dans les contextes de métadiscours[3]. Ce score se fait alors à partir d'une liste de mots clés, indépendants du régionalisme en question. Ils ne sont pas exclusivement présents dans les cas de métadiscours mais de cette manière, les tokens propres à notre corpus d'apprentissage devraient avoir moins d'incidence sur la classification. Le modèle serait alors capable de mieux généraliser son apprentissage vers de nouveaux tweets.

2.4 Apprentissage

Ces procédés sont ensuite ajoutés à une chaîne de traitement pour la classification. Cette méthode permet de regrouper les traitements des descripteurs pour ensuite

[3] Cette démarche pourrait cependant être améliorée, elle a principalement été mise en place pour pallier le manque de précision du modèle.

appliquer différentes chaînes de traitement à un jeu de données de manière modulable et séquentielle. En effet, même si plusieurs descripteurs sont définis, il est facile de retirer un groupe afin d'essayer plusieurs classifications. De plus, chaque étape est associée à un nom, ce qui nous permet de modifier rapidement les paramètres d'une étape si nécessaire. Une fois les pré-traitements effectués, la chaîne de traitement se termine par un classifieur qui renvoie un score de prédiction.

Un corpus d'apprentissage a besoin d'être segmenté, la première partie sert à entraîner le modèle, la deuxième permet de tester le modèle et d'obtenir des prédictions sur les données de cette partie. La classe attendue de chaque tweet n'est donc pas donnée au modèle, cela nous permet ensuite de comparer les classes prédites aux classes attendues. Grâce à notre chaîne de traitement, il est très facile d'effectuer les pré-traitements sur une partie, d'entraîner le modèle à associer des descripteurs à une classe, de prétraiter la seconde partie puis de prédire les classes de ces exemples.

Il fallait cependant savoir quel modèle d'apprentissage serait le plus performant pour notre corpus. Pour cela, nous avons utilisé la validation croisée. Il s'agit d'une technique qui segmente un corpus en plusieurs plis afin d'appliquer un modèle sur différents ensembles entraînement/test. Ainsi, nous obtenons autant de scores de prédiction qu'il y a de plis, ce qui nous permet de récupérer une moyenne ou un écart-type pour un seul modèle d'apprentissage. Il est important de réaliser ces segmentations sur l'intégralité du corpus afin d'éviter le surapprentissage, c'est-à-dire le moment où le modèle se sera trop spécialisé et où la généralisation vers de nouvelles données deviendra difficile. Si nous essayons différents classifieurs exclusivement sur la partie entraînement du corpus, les scores que nous obtiendrons ne seront peut-être pas représentatifs du corpus intégral. Une segmentation en plis sur tous les tweets permet ainsi d'éviter ce problème.

La prédiction de classes cibles engendre plusieurs résultats utiles pour le jugement de l'efficacité d'un modèle. La précision rend compte de la proportion d'exemples bien classés dans une certaine cible. Cependant il est important d'observer le rappel d'une classe cible, celui-ci représente la proportion d'exemples bien classés par rapport au nombre d'exemples attendus dans cette classe. Pour comparer rapidement des modèles ou des paramétrages, nous pouvons utiliser la f-mesure qui prend en compte

ces deux valeurs. La librairie scikit-learn propose également une f-mesure pondérée, c'est-à-dire qu'elle rend compte du déséquilibre des données, notamment lorsqu'une classe est moins représentée qu'une autre en termes d'exemples.

2.4.1 Liste des modèles

Nous présenterons dans cette section les modèles ayant obtenus les meilleurs résultats avec notre corpus de tweets. Dans la majorité des cas de classification, il est impossible de prévoir quel modèle aura la meilleure exactitude, nous avons donc essayé un grand nombre de classifieurs jusqu'à atteindre un résultat satisfaisant.

- Decision Tree (Arbre de décision) :

Lors de l'apprentissage, le but de ce modèle est de constituer une arborescence pour classer les données d'entraînement. Il crée des noeuds en testant des règles de classification grâce aux descripteurs, ce qui génère des feuilles contenant une classe associée aux données confirmant cette règle et cette classe. Ce modèle se spécialise peu à peu jusqu'à ce que l'intégralité des données d'entraînement soient triées. Certaines feuilles non pertinentes peuvent ensuite être supprimées lors du post-élagage afin d'éviter le surapprentissage. Ce modèle simple a étonnement obtenu de bons résultats avec notre corpus, nous avons cependant préféré nous concentrer sur sa version améliorée : le modèle RandomForest qui est décrit plus loin dans cette section.

- Neural network (Réseaux de neurones artificiels) :

Ces modèles se basent sur le fonctionnement des neurones humains. Chez l'Homme, l'information circule dans les neurones sous forme de courant électrique et chimique, lorsque plusieurs neurones se rencontrent en un point, une sommation est faite entre les synapses (connexions entre les neurones) inhibitrices et excitatrices. Chaque neurone a alors un poids qui est additionné à l'autre pour obtenir une nouvelle intensité de courant. De la même manière, les neurones artificiels reçoivent en entrée plusieurs descripteurs associés à un poids, ils sont additionnés et traités selon un seuil à une ou plusieurs reprises (ce sont les couches internes). Lors de l'apprentissage, l'algorithme met à jour les seuils et les poids en fonction de chaque classe à obtenir en sortie. C'est la fonction d'activation qui va ensuite décider quelle sortie doit être activée grâce aux

poids. Les neurones artificiels sont bien souvent non supervisés mais il existe certains modèles supervisés :

- Perceptron et MultiLayer Perceptron (Perceptron multi-couche)

Un perceptron est un modèle de neurone artificiel simple, il ne possède qu'une sortie et une couche interne, il est similaire à un modèle linéaire. Le MultiLayer Perceptron possède quant-à-lui plusieurs couches internes pour modifier les poids. De plus, sa rétropropagation permet de faire des allers-retours entre la sortie et l'entrée afin d'ajuster les poids en fonction de l'erreur obtenue en sortie (l'erreur est la comparaison entre la classe attendue et la classe prédite). Les différentes sorties de ce modèle permettent de l'utiliser pour de la classification à plusieurs labels et classes.

Les réseaux de neurones peuvent être très performants mais nos résultats ont grandement varié en fonction des pré-traitements réalisés, passant de 60% à 80% pour certains d'entre eux. De plus, la taille de notre corpus a sans doute eu un impact sur la qualité des prédictions de ce modèle.

- Régression logistique (Logistic regression) :

La régression linéaire et la régression logistique sont similaires à la différence que la première permet de prédire des valeurs tandis que la deuxième est utilisée pour les tâches de classification. La régression linéaire consiste à séparer un ensemble de points grâce à une droite ou un plan ayant un équation de type « $y = ax + b + \varepsilon(x)$ »; où y est la variable à prédire, a et b sont les coefficients que le modèle va déterminer et x est la variable prédictive, epsilon représente le bruit des mesures (Rakotomalala). Ce modèle est en particulier efficace pour des données simples, il est alors plus judicieux d'essayer le « support vector machine » (SVM) pour des situations plus complexes.

La régression logistique est un cas particulier de la régression linéaire. Elle utilise des vecteurs afin d'obtenir une classification de type vrai/faux. La librairie scikit-learn possède également un SVM adapté à la classification.

Nous avons obtenu de très bons résultats avec ce modèle mais le paramétrage de celui-ci est assez restreint, il ne nous aurait donc pas permis d'obtenir une forte amélioration dans nos résultats.

- SVM (Support Vector Machines) :

Les machines à vecteur de support sont des modèles très utilisés pour la régression – c'est-à-dire la prédiction de valeur – mais peuvent également l'être pour la classification. Ce modèle permet de résoudre des problèmes plus complexes que les régressions linéaires, notamment lorsque la séparation des données ne peut pas être réalisée par une simple droite. La séparation des données peut alors se faire grâce à des équations de type polynômes ou gaussiennes qui prennent en compte des marges de chaque côté de la droite.

Cette méthode reste efficace avec un très grand nombre de descripteurs et peut même obtenir des résultats comparables aux réseaux de neurones. Il existe différents types de SVM, sur scikit-learn, le modèle le plus adapté à notre tâche est le C-Support Vector Classification puisqu'il permet de catégoriser selon plus de deux classes. Le SVM est cependant très sensible aux proportions de données par classe, ce qui explique probablement les résultats décevants que nous avons obtenus (74% d'exactitude).

- Gradient boosting :

Cet algorithme fait partie des techniques ensemblistes, qui font coopérer plusieurs modèles d'apprentissage simples, principalement à l'aide d'arbres de décision. Cela permet d'améliorer l'exactitude, de réduire la variance dû à de mauvaises données d'apprentissage et d'éviter le surapprentissage (*Ibid.*). Le boosting a la particularité d'être séquentiel et de pondérer les éléments précédemment mal classés à chaque nouvel arbre, mais cela peut représenter un problème si le nombre de données aberrantes est élevé. Cet algorithme nous a permis d'obtenir de très bons résultats.

- Random forest (forêts aléatoires) :

Les forêts d'arbres décisionnels apprennent sur plusieurs arbres de décision avec différents groupes de données indépendants, le nombre d'arbres est à définir par l'utilisateur. La performance des arbres de décision habituels peut être altérée en fonction de l'ordre de prise en compte des descripteurs, ce modèle permet alors de résoudre ce problème et peut également réduire la complexité de l'arborescence.

Ce modèle est une amélioration de la méthode ensembliste bagging, il obtient de manière générale des résultats similaires au boosting. Il n'est pas difficile à paramétrer

et évite le problème de surapprentissage. Cette méthode est cependant sensible au nombre de descripteurs pertinents. Pourtant, c'est avec elle que nous avons obtenu la plus haute exactitude parmi tous les classifieurs essayés.

Modèle	F-mesure
Baseline (classer dans la cible la plus fréquente)	55.58
Arbre de décision	78.15
Régression logistique	82.46
Perceptron multi-couches	74.76
Machine à vecteur de support	74.00
Gradient Boosting	81.84
Forêts aléatoires	84.46

Tableau 17 : Taux d'exactitude des modèles sur trois plis

2.5 Conclusions pré-traitements

Nous avons ensuite utilisé les deux modèles les plus performants - c'est-à-dire Random Forest et Gradient Boosting - pour essayer plusieurs combinaisons de pré-traitements. Nous nous sommes finalement arrêtés sur ces conclusions :

2.5.1 Traitement du tweet

Après avoir réalisé plusieurs essais, nous avons conclu que le CountVectorizer permettait d'obtenir de meilleurs résultats de classification que la méthode Tf-idf. Cela est probablement lié au fait que les tweets sont pour la plupart très courts et que leur thème varie en fonction du régionalisme. Le poids des mots liés à un régionalisme peut donc potentiellement entrer en conflit avec ce qui permet réellement de déceler la classe du tweet.

De plus, nos essais nous ont permis de conclure que la lemmatisation serait le choix le plus adapté à notre corpus. Cela s'explique probablement par le fait que les tweets sont différents du langage SMS. Ils contiennent bien sûr des fautes d'orthographe mais ils ne sont que très rarement écrits de manière abrégée et le nombre de fautes est atténué par l'essor des outils de correction automatique comme le T9, ce qui rend la

lemmatisation satisfaisante. Cependant, ce ne sont que des observations déduites de notre corpus et cela peut ne pas correspondre à tout le réseau social.

En ce qui concerne la prise en compte de mots-vides, nous avons tout d'abord remarqué que la ponctuation était nécessaire à l'apprentissage, sans doute parce qu'elle reste rare dans les tweets. Sa présence peut ainsi apporter des indices sur le profil de l'utilisateur, et par extension le type d'usage du régionalisme. Tout comme la présence de majuscule, une ponctuation très fréquente peut être le signe d'un tweet généré automatiquement (où les emplois authentiques de régionalismes se feront plus rare). Après plusieurs essais, le meilleur choix fut de n'utiliser aucun stopword pour notre pré-traitement, cela est lié au choix de segmenter les tweets en n-grammes de mots allant de 2 à 4 et de remplacer le régionalisme par le marqueur « regio » plutôt que de le supprimer. Tous ces traitements ont pour but d'éviter le surapprentissage lié à la taille de notre corpus. En effet, nous avions dans un premier temps essayé de classifier de nouveaux régionalismes et de nouveaux tweets avec des mots-vides (sans ponctuation) et des lemmes simples, sans le régionalisme. Cependant, nous avons remarqué deux choses : de nombreux cas de métadiscours étaient oubliés et certains régionalismes étaient catégorisés comme une entité nommée simplement parce que leur tweet contenait un lemme très fréquent dans notre corpus. Par exemple, le lemme « pont » était uniquement présent dans nos exemples d'entités nommées, il avait ainsi une grande importance dans la classification. La lemmatisation en n-grammes permet alors d'obtenir des descripteurs tels que « on dit regio » ou « ou regio ». De plus, puisque le marqueur « regio » est très fréquent, les n-grammes ont tendance à se former autour du régionalisme, ce qui évite de prendre en compte le contexte des tweets longs.

2.5.2 Entités

Pour rappel, les entités regroupent les liens hypertextes, les arobases et les mots-dièse contenus dans un tweet. Le fichier brut affichait à l'origine le nombre d'occurrences de ces entités, sans limitation. C'est dans le but d'améliorer l'analyse du classifieur que nous avons essayé principalement deux configurations. La première méthode consistait à ne garder qu'une valeur binaire, c'est-à-dire « 1 » en cas de présence d'une entité ou « 0 » dans le cas contraire. La deuxième mettait en place une limite de valeur pour ne pas obtenir de valeurs trop hautes et singulières. Notre

choix s'est finalement porté sur une plage de valeur allant de zéro à trois entités, remplaçant ainsi les données supérieures à cette limite par la valeur trois.

2.5.3 Entités nommées

Concernant l'annotation d'entité nommée de chaque régionalisme, nous expliquions plus haut que l'annotation « B », signifiant que le mot cible commence une entité nommée, était moins fiable en raison de son pourcentage d'exactitude faible. Nous avons donc essayé de classifier les tweets en prenant en compte l'annotation des entités nommées de plusieurs manières. Une première approche a été de fusionner les annotations « B » et « I » (qui signifie que le régionalisme est à l'intérieur d'une entité nommée), mais cela ne nous a pas apporté d'assez bons résultats. Nous avons également essayé de ne pas prendre en compte l'annotation « B » mais cela provoquait une baisse du taux de rappel des entités nommées. Notre choix s'est donc arrêté sur la prise en compte de toutes les annotations.

2.5.4 Métadiscours

Notre liste de mots-clés a évolué au fil des essais. En effet, la reconnaissance du métadiscours a toujours souffert d'un rappel trop bas, notre but a donc été de l'élever à une valeur plus satisfaisante. Pour enrichir notre liste de mots-clés, nous nous sommes servis d'un CountVectorizer, uniquement appliqué aux exemples de type « métadiscours ». Grâce à cela, nous avons pu relever les lemmes les plus fréquents dans cette catégorie. Cette liste a eu beaucoup d'influence sur la qualité des résultats, elle doit donc être modifiée avec précaution. Pour rappel, la liste de mots-clés permet d'obtenir un score en fonction de leurs occurrences et de la longueur de la liste. Il a donc été pertinent de comparer la moyenne et les quartiles de ces scores en fonction du type de régionalisme lors des essais.

2.6 Paramétrage des modèles

Les modèles que nous avons précédemment essayés n'étaient que très peu paramétrés afin de réaliser une première sélection. Ceux ayant obtenus les meilleurs résultats ont ensuite été optimisés afin d'éviter le surapprentissage et d'améliorer le rappel et la précision de chaque classe. Nous nous sommes ainsi concentrés sur les modèles Gradient Boosting et Random Forest. L'optimisation s'effectue à l'aide de la méthode « GridSearchCV » de scikit-learn. Cette technique performante teste

plusieurs combinaisons d'hyper-paramètres pour enregistrer la configuration la plus adaptée à nos données. Les hyper-paramètres à essayer sont définis en entrée, l'utilisation de la chaîne de traitement permet également de tester plusieurs paramètres de vectorisation des données. Cependant, il est important de noter que la méthode « GridSearchCV » utilise la validation croisée et peut donc prendre beaucoup de temps à s'exécuter si le nombre d'hyper-paramètres est élevé.

Les algorithmes de Boosting et RandomForest se basent tous deux sur l'optimisation d'arbres de décision, ils possèdent ainsi plusieurs paramètres en commun, en particulier pour la formation des arbres de décision (fig. 10).

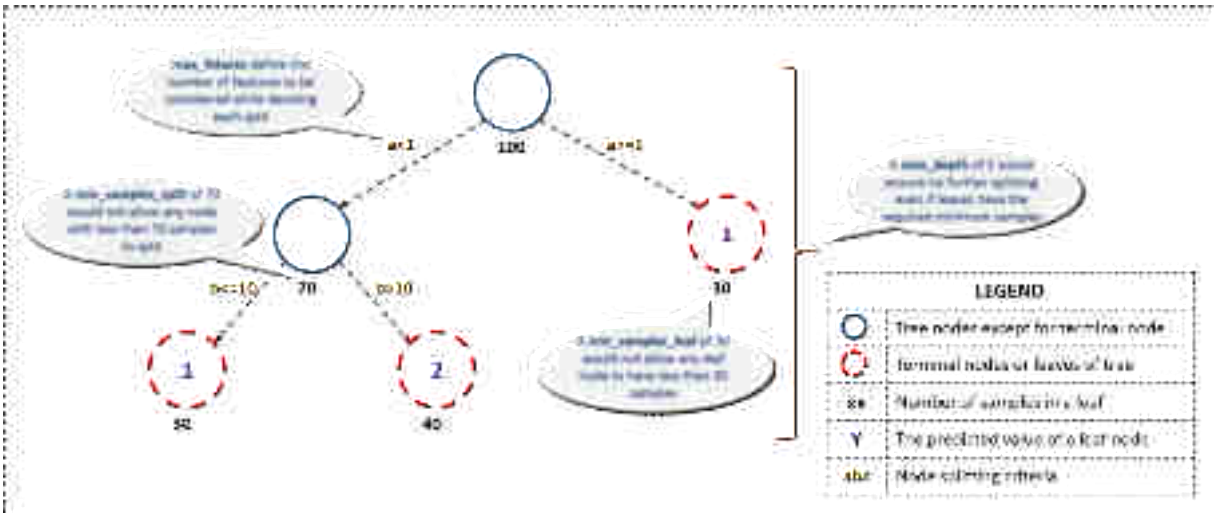

Figure 10 : : Explication du paramétrage des arbres (« Complete Guide to Parameter Tuning in Gradient Boosting (GBM) in Python » 2016)

_max_features : nombre de descripteur maximal à prendre en compte pour l'élaboration d'un arbre. Pour rappel, les descripteurs définissent les règles de classification qui permettent de segmenter les données.

_min_samples_split : nombre d'exemple minimal à prendre en compte pour tester une règle de classification

_min_samples_leaf : nombre d'exemple minimal à obtenir après un test de classification.

_max_depth : nombre de segmentation maximal autorisé

_n_estimators : nombre d'arbre à intégrer dans l'algorithme

Différentes combinaisons de paramètres ont ainsi été essayées sur les modèles « Random Forest » et « GradientBoosting ». Il était important d'essayer d'augmenter certaines valeurs, notamment « n_samples_split » et « n_samples_leaf », respectivement définies 1 et 2 par défaut. Cela engendre un fort taux de surapprentissage puisque cela signifie qu'une règle de classification peut être définie sur uniquement 2 exemples. Le paramétrage de « max_depth » résout également ce problème si l'on diminue sa valeur.

2.6.1 Résultats du paramétrage

Une fois les modèles paramétrés et les nouvelles prédictions réalisées, nous avons pu comparer les performances des modèles. De manière générale, « Random Forest » a obtenu de meilleurs résultats que le « Gradient Boosting ». Ci-dessous se trouvent les résultats de prédiction obtenus après le paramétrage des forêts aléatoires. Pour comparaison, l'optimisation du « Gradient Boosting » a permis d'obtenir une f-mesure de 87%, accompagné d'un rappel de 73% pour le métadiscours et 78% pour les entités nommées, ce qui n'est pas suffisant pour nos besoins.

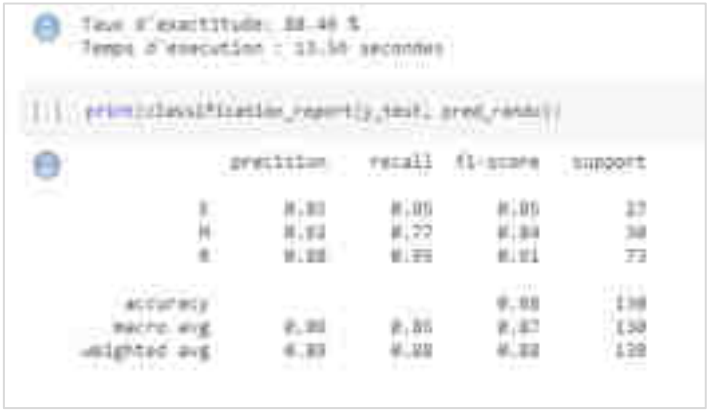

Figure 11 : Score d'exactitude et rapport de classification (Random Forest)

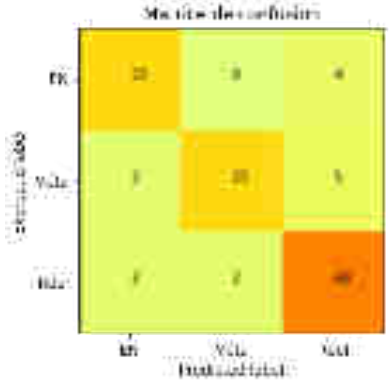

Figure 12 : Matrice de confusion (Random Forest)

Nous pouvons voir que les exemples réels sont très peu oubliés (0.95 de rappel), ce qui était notre but. Neuf tweets sont malheureusement mal classés sur les 78 tweets classés comme étant réels, mais ce modèle a l'avantage d'oublier peu de cas de métadiscours (seulement 5 sur 28). Si nous regardons ces instances mal classées, nous remarquons que des mots dièses comme #TeamChocolatine ou #Languefrançaise auraient pu être pris en compte par l'algorithme, une analyse plus poussée des mots dièses pertinents pourrait ainsi améliorer les résultats.

En ce qui concerne les descripteurs sélectionnés pour l'apprentissage du modèle Random Forest paramétré (Annexe 3), le score de métadiscours (calculé à partir d'une liste de lemmes pertinents) et le descripteur « EN », reconnaissant les mentions d'utilisateurs, sont les plus utiles à la classification. Le pourcentage de majuscule et la longueur du tweet semblent également avoir une importance, tout comme le descripteur « maybe_in_EN » qui reconnaît les régionalismes contenus dans des entités nommées poly-lexicales. Par ailleurs, le nombre de mots-dièses et de liens hypertextes dans le tweet semblent moins pertinents que le nombre de mention d'utilisateurs. Enfin, nous retrouvons quelques n-grammes pertinents tels que « dire regio », « alain regio », « pain au chocolat », « pont de regio » ou encore « on dit ». Pour rappel, le marqueur « regio » sert à remplacer le régionalisme dans le tweet.

Il est important de préciser que certains de nos pré-traitements nous ont permis d'obtenir une f-mesure plus élevée que 88%. Cependant, comme expliqué précédemment, ces apprentissages ne nous permettaient pas de prédire de manière satisfaisante la classe de nouveaux régionalismes provenant de tweets en plus grande quantité. Nous avons ainsi choisi de conserver ce modèle pour l'annotation de nouveaux tweets.

3. Comparaison avec étude d'Evangelia Balouta

La principale différence entre la démarche d'Evangelia Balouta (2017) et celle présentée ici repose dans le fait que la première prend en compte des descripteurs sélectionnés manuellement tandis que nous avons opté pour une méthode qui relève plus de descripteurs mais de manière moins précise et contrôlée.

Dans notre démarche, plus de lemmes sont sélectionnés grâce au CountVectorizer mais nous retrouvons tout de même certains descripteurs pris en compte par

Evangelia Balouta (*Ibid.*, p. 73), notamment la présence de parenthèses, de tiret, de point, d'un déterminant possessif, ou encore d'un article défini ou indéfini… De plus, nos autres descripteurs sélectionnés manuellement rejoignent également ses résultats. Nous retrouvons ainsi les lemmes « mot », « dire », « appeler », « prononcer », « apprendre » dans nos listes respectives de lemmes permettant de reconnaître les cas de métadiscours. Cependant, certains descripteurs sélectionnés par Evangelia Balouta ne se prêtent pas au registre familier ou courant que l'on peut trouver dans un corpus de tweets (ex : « considérer », « employer » …), nous n'avons donc pas jugé nécessaire de les ajouter à nos descripteurs. Par ailleurs, plusieurs descripteurs spécifiques au corpus d'Evangelia Balouta ont été pris en compte lors de ses travaux. Bien que le CountVectorizer ait relevé certains lemmes spécifiques à notre corpus d'entraînement, nous avons tenté de limiter au maximum ce biais en prenant en compte des bi-grammes et tri-grammes afin d'obtenir une meilleure généralisation lors de l'étape de prédiction. Enfin, la sélection de ses descripteurs se faisait sur les voisins proches du régionalisme cible. En raison du contexte restreint des tweets (280 caractères), nous avons estimé qu'il n'était pas nécessaire de réaliser cette opération.

En ce qui concerne l'apprentissage de la classification automatique, le logiciel Weka utilisé par Evangelia Balouta est certes très complet, mais notre utilisation de scikit-learn nous a permis d'avoir plus de contrôle sur le paramétrage des modèles et la sélection des descripteurs. De plus, notre projet entièrement réalisé sous Python nous donne la possibilité de classifier des tweets de manière plus fluide et optimale et d'intégrer ce projet à une chaîne de traitement.

4. Visualisation des tweets

4.1 Prédiction et génération de fichiers

Le script « prediction.py » permettant de prédire la classe de nouveaux tweets reprend les principales étapes de pré-traitement du script d'apprentissage. À l'exception que le modèle obtenu lors de l'apprentissage est importé et n'a donc pas besoin d'être de nouveau entraîné. De plus, nous ne réalisons pas d'annotation préalable pour les

classes cibles. Chaque régionalisme est traité indépendamment afin d'obtenir une visualisation de carte pour chacun d'entre eux. Les informations pertinentes de chaque tweet sont ensuite écrites dans deux fichiers au format CSV, l'un servira au stockage de tous les tweets et éventuellement au parcours manuel des informations, l'autre permettra de visualiser les tweets géolocalisés, il contient les coordonnées ainsi que le tweet et la classe prédite par le modèle. L'utilisateur est également présent dans le fichier afin d'observer si plusieurs occurrences d'un régionalisme correspondent à une même personne. De plus, nous avons inscrit le lieu d'habitation renseigné par l'utilisateur dans sa description, afin de comparer les coordonnées du tweet et l'origine géographique de l'utilisateur, dans le cas où cette dernière est exploitable.

Le nombre de tweets pouvant être pris en compte dans la prédiction est infini mais pour réaliser les cartes présentées plus loin, nous avons collecté 1390 tweets pour 34 régionalismes (ou formes de régionalisme) et 1157 utilisateurs distincts. Sur ce corpus, 921 tweets comportaient un usage réel, 272 contenaient un régionalisme annoté comme une entité nommée et 197 étaient des cas de métadiscours, selon les résultats de notre prédiction.

4.2 Qlik Sense Cloud

Une fois le fichier au format CSV créé, il n'est pas difficile de projeter les données sur une carte afin d'observer l'expansion de certains régionalismes. Nous avons choisi l'outil Qlik Sense (« Data Visualization Software in the Cloud | Qlik Cloud ») pour sa rapidité de prise en main et son efficacité. C'est un outil de visualisation de données et d'aide à la décision. Fondée en 1993, la société Qlik a développé son produit Qlik View et plus récemment Qlik Sense, qui est un outil plus accessible à un public non expert. Il existe une version « Cloud » où certaines fonctionnalités sont gratuites, dont le partage de données entre plusieurs utilisateurs et appareils. La version « Bureau » est également gratuite mais est davantage réservée à un usage personnel. En plus de sa facilité d'utilisation, Qlik Sens propose de nombreuses fonctions géographiques, maniables à l'aide d'une interface graphique, ce qui nous a paru tout à fait approprié à notre besoin. Cependant, il est bien sûr possible de s'orienter vers d'autres outils de visualisation comme Leaflet et d3.js en JavaScript, Plotly sur Python ou d'autres logiciels similaires à Qlik comme Tableau, Power BI ou encore SAP.

La visualisation des données est une étape importante, elle permet une compréhension intuitive des résultats ainsi qu'un partage des connaissances. Les nombreux graphiques, tableaux et cartes disponibles sur Qlik facilitent la réalisation de cette étape. Cet outil permet d'afficher toute sorte de données mais les extensions de fichier Qlik (.qvd) ne pouvant être décodées sans le logiciel, il ne sera pas possible de joindre les résultats proprement dits. Nous joindrons donc dans ce projet des captures d'écrans des analyses réalisées.

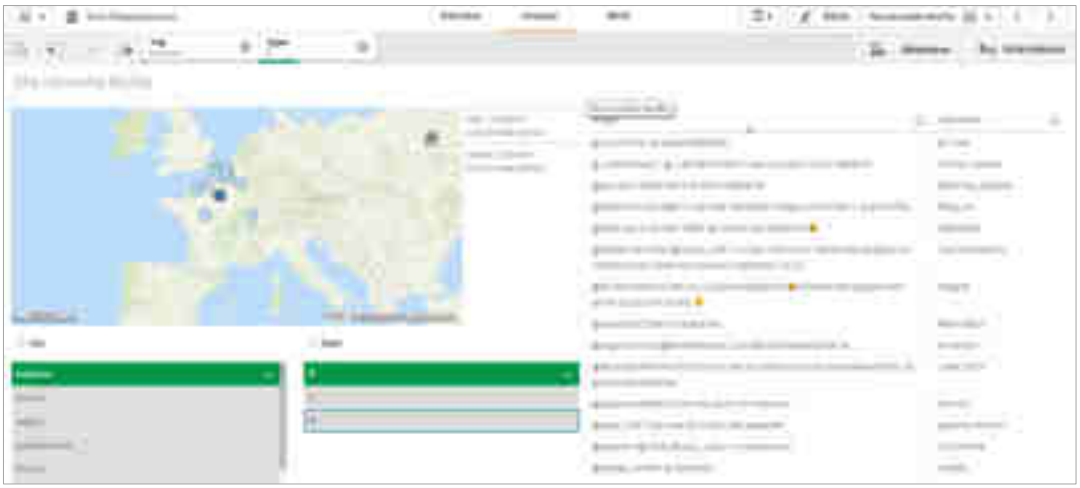

Figure 13 : Exemple de visualisation de données sur Qlik Cloud

L'espace de travail de Qlik Cloud peut être modulé selon nos envies. Ici, nous pouvons sélectionner un régionalisme, sa catégorie prédite par le classifieur et avoir une vue sur les tweets et leur localisation. Sur les exemples que nous montrerons ci-dessous, une carte a été générée à partir de la sélection d'un régionalisme et de ses utilisations dites « réelles ». Sur les cartes que nous présenterons ci-dessous, la taille des points a été calculée à partir du nombre d'utilisateurs différents pour chaque lieu. Cela nous permet de ne prendre en compte qu'un tweet par utilisateur. La légende à droite de ces cartes donne accès aux quantités minimales et maximales d'utilisateurs pour un lieu.

4.3 Étude de la vitalité et de l'expansion

Dans cette section, nous analyserons la vitalité d'une sélection de régionalismes ayant obtenus des résultats satisfaisants à la suite de plusieurs collectes de tweets et d'une annotation automatique. Il est important de noter que, pour obtenir plus d'occurrences géolocalisées, nous avons pris en compte la localisation du tweet (point jaune) et le lieu d'habitation de l'utilisateur renseigné dans sa description (point rouge).

4.3.1 Régionalismes du Nord (Belgique et Nord de la France métropolitaine)

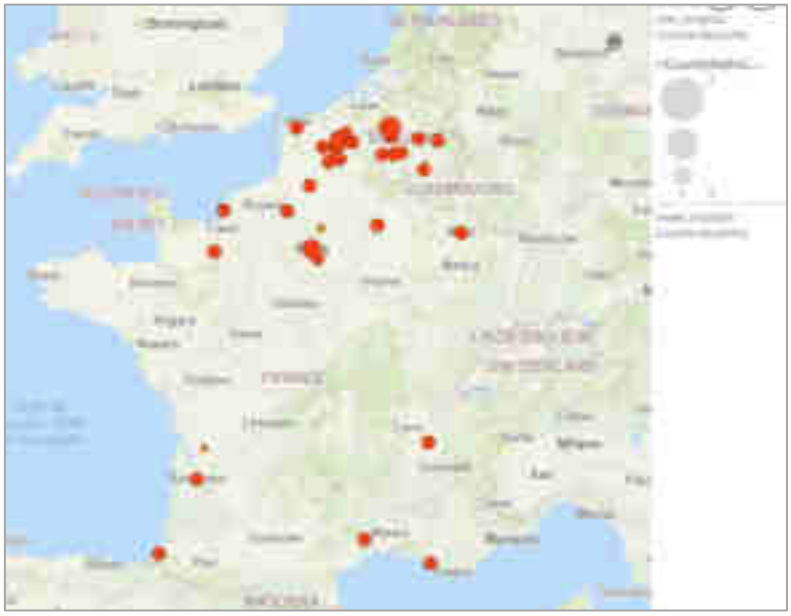

Figure 14 : Répartition géographique du régionalisme « drache »

Cette carte met en valeur l'utilisation du régionalisme « drache », nous notons que la répartition des tweets est principalement située autour de Lille et en Belgique, ce qui corrobore les résultats du site « françaisdenosregions.com » (Avanzi 2017a). Nous remarquons cependant plus d'occurrences en Île de France et dans le Sud. Nous noterons toutefois que deux des points visibles dans le Sud représentent des cas de métadiscours mal catégorisés par la classification automatique.

Figure 15 : Répartition géographique du régionalisme « babache »

Ici, nous analysons un régionalisme qui n'a pas encore fait l'objet d'une étude ; « babache ». Ce terme désigne une personne niaise, simple d'esprit ou idiote en picard. Effectivement, nous observons que son utilisation est restreinte au nord des Hauts-de-France, sans traverser la frontière. Pourtant, ce régionalisme semble être très fréquent, nous pouvons ainsi faire l'hypothèse que sa zone d'utilisation pourrait s'étendre dans les années à venir.

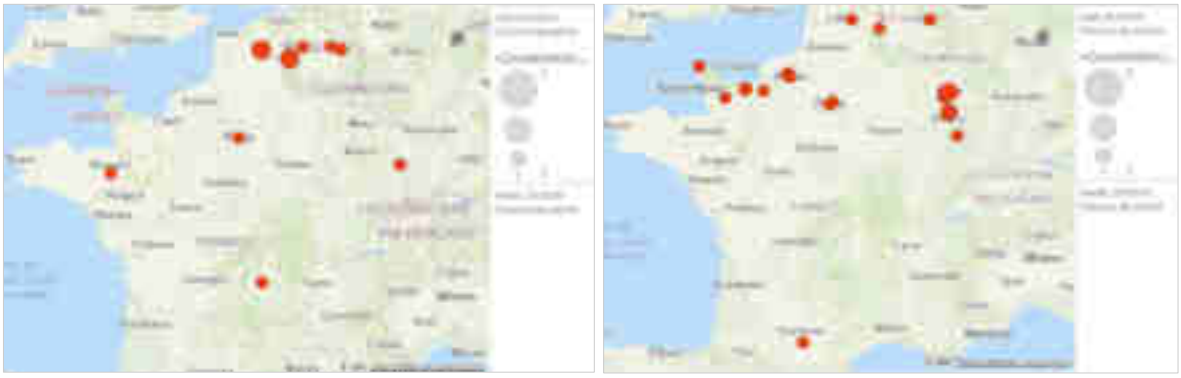

Figure 16 : Répartition géographique du régionalisme « chicon » (à droite), « clenche » et « clanche » (à gauche)

En ce qui concerne les régionalismes « chicon » et « clenche », nous retrouvons les mêmes répartitions que celles décrites dans les articles de Mathieu Avanzi. À l'exception de quelques points, mettant probablement en valeur le déménagement de certains utilisateurs.

4.3.2 Régionalismes du Sud

Figure 17 : Répartition géographique du régionalisme « dégun »

Cette carte nous permet d'observer l'expansion géographique du régionalisme « dégun », qui signifie « personne », dans le sens « aucun individu ». Cette forme

originaire de l'occitan s'est popularisée grâce au slogan marseillais « on craint dégun ». Nous pouvons penser que ce slogan, plutôt que de contribuer à l'expansion de « dégun », a restreint l'emploi de ce régionalisme à une certaine communauté voulant affirmer son identité. Les utilisations visibles en dehors de la côte méditerranéenne appartiennent probablement à des supporters de l'Olympique de Marseille ou à des personnes ayant vécus dans cette zone géographique.

Figure 18 : Répartition géographique du régionalisme « tarpin »

Sur cette carte nous pouvons observer les zones d'utilisation de « tarpin », qui signifie « beaucoup » ou « très ». Ce régionalisme est intéressant car il est extrêmement présent dans notre corpus mais son utilisation reste restreinte à la côte méditerranéenne. C'est sans doute un exemple de régionalisme qui ne s'étendra jamais au reste de la France, mettant en valeur l'identité d'une population.

4.3.3 Comparaison de régionalismes

- Appellations de kebab :

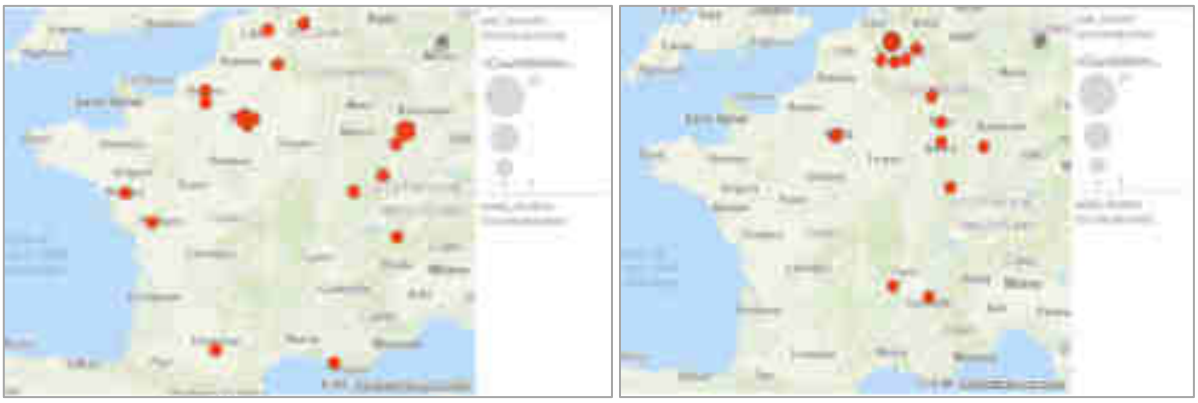

Figure 19 : Répartition géographique des régionalismes « döner » et « dürüm »

Pour les appellations de « kebab », nous avons sélectionné les formes « döner » et « dürüm », toutes deux avec et sans accents. Les cartes ci-dessous tendent à prouver que la forme « döner » est répandue sur une large zone mais en réalité, les occurrences visibles en dehors de l'Alsace, du Sud et de Paris sont des exemples de tweets mal orthographiés, où le verbe « donner » est écrit « doner ». Cela nous permet tout de même d'observer une plus grande expansion que celle citée dans l'article de « français de nos régions » dédié aux appellations du kebab (Avanzi 2018b). De la même manière, nous observons une certaine vitalité du régionalisme « dürüm » (ou durum), qui s'étend dans le Nord-Est de la France.

- Fête du village :

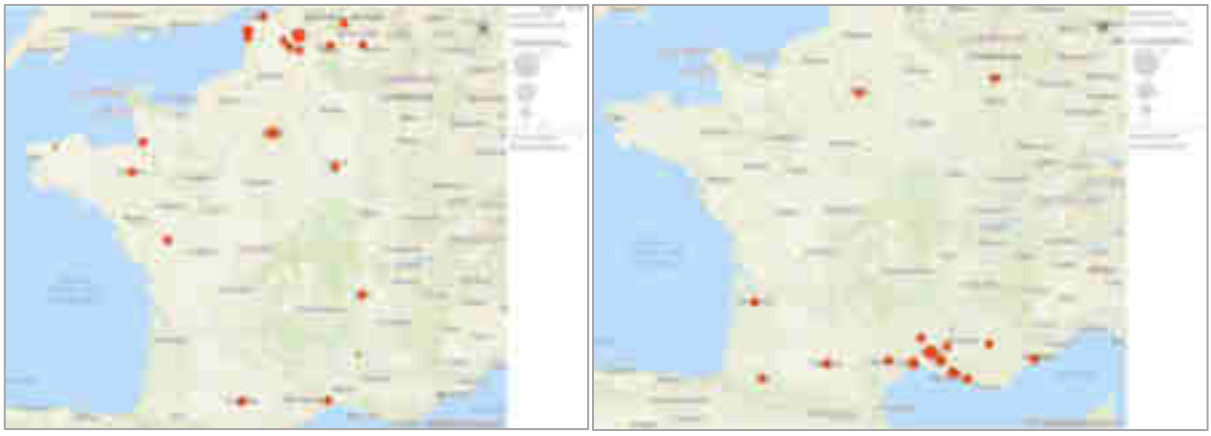

Figure 20 : Répartition géographique des régionalismes « ducasse » et « feria »

En ce qui concerne l'appellation de la fête du village, nous observons dans un premier temps de nombreuses occurrences de « ducasse » en dehors de sa zone d'expansion

supposée (les Hauts-de-France). Une rapide vérification des données nous a cependant montré que les points situés en dehors de cette zone représentaient, dans la majorité des cas, une nomination du chef Alain Ducasse. Par ailleurs, l'étude du régionalisme « feria » nous offre une répartition géographique différente des résultats de Mathieu Avanzi (Avanzi 2017g), puisque nous retrouvons beaucoup d'occurrences autour de Nîmes et peu dans le Sud-Ouest. Une collecte ultérieure de nouvelles données pourrait toutefois nous indiquer si ces tweets ont été influencés par la date d'import, qui pourrait coïncider avec la date d'une certaine fête de village.

- Téléphone portable/mobile :

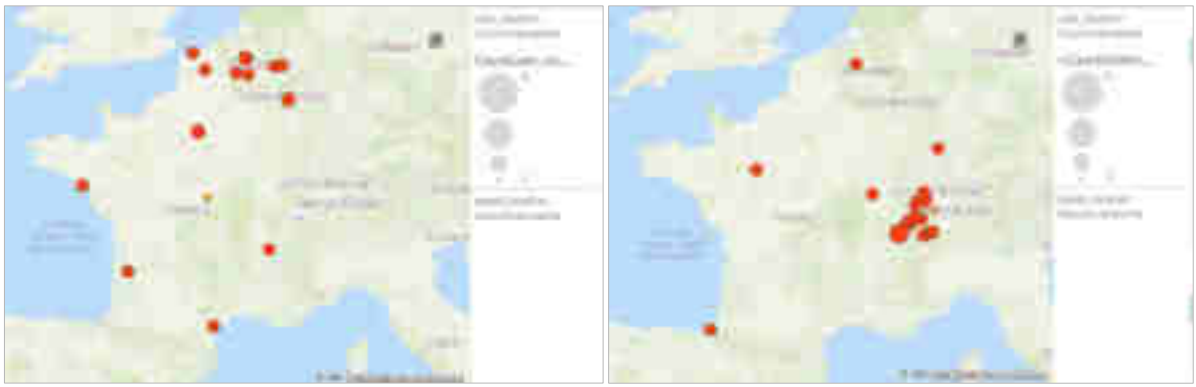

Figure 21 : Répartition géographique des régionalismes « gsm » et « natel »

Dans le cas du téléphone portable, nous observons que l'appellation « natel » reste très restreinte à la Suisse, comme nous l'avions prédit. Cependant, la forme « GSM » est présente dans bien plus de lieux. Ces utilisations réelles peuvent être dues à des déménagements de personnes ayant habité en Belgique. Une deuxième hypothèse serait que ces occurrences proviennent de personnes d'origine africaine ou canadienne, où le sigle « GSM » est parfois employé.

Conclusion

Ce projet avait pour but d'étudier la vitalité de régionalismes francophones sur Twitter. Pour cela, nous avons mis en place une étape de classification automatique permettant de ne garder que les utilisations authentiques de ces particularités linguistiques. De cette manière, les régionalismes intégrés à une entité nommée ou à un métadiscours ont été retirés dans la mesure du possible.

À partir d'un corpus de 650 tweets, nous avons réussi à obtenir une exactitude de 88.46%. Ce pourcentage pourrait être amélioré grâce à un corpus d'apprentissage plus large, où les catégories « métadiscours » et « entité nommées » sont mieux représentées. Concernant l'exactitude de chaque classe cible, nous avons choisi d'obtenir un meilleur rappel des régionalismes « réels » au détriment du rappel du métadiscours. En d'autres termes, nous avons fait en sorte que les cas authentiques soient moins souvent mal catégorisés, même si cela signifie que certains cas de métadiscours ne seront pas reconnus. Ce choix s'explique par le fait que la majorité des régionalismes ont peu de visibilité en dehors de leur zone d'utilisation, ils ne sont donc pas souvent employés dans des cas de métadiscours. Les entités nommées sont quand-à-elle bien reconnues de manière générale mais nous rencontrons des déficits lorsque cette entité n'est pas poly-lexicale, en raison de la présence non régulière de majuscules sur Twitter.

Certaines améliorations restent envisageables pour ce projet. En effet, nous avons réalisé nos pré-traitements dans l'optique que les régionalismes à prédire ne seraient que mono-lexicaux. Il serait alors pertinent de prendre en compte les régionalismes composés de plusieurs éléments, mais cela demanderait quelques manipulations dans le processus de tokenisation de spaCy. Par ailleurs, les formes verbales n'ont pas été prises en compte, il serait alors intéressant d'intégrer la conjugaison de certains régionalismes à notre recherche de tweets.

Pour conclure, l'étude de la vitalité des particularités linguistiques nécessite de rassembler un grand nombre de données. Une collecte de tweets en temps réel et l'utilisation éventuelle d'une API premium pourraient nous permettre de réaliser de meilleures analyses. Le processus pourrait alors contenir la construction d'une carte et la visualisation des données géolocalisées en temps réel. De plus, il serait

intéressant d'observer l'évolution de la vitalité des régionalismes, il faudrait pour cela rassembler des tweets sur une période plus étendue. De toute manière, il sera nécessaire d'effectuer une recherche à grande échelle pour pouvoir compenser la quantité de tweets non géolocalisés.

Bibliographie

AMBROGI Pascal-Raphaël et DIOUF Abdou, 2015, *Dictionnaire du bon usage au service du sens et de la nuance*, Paris, Honoré Champion (coll. « Champion classiques. Série "Références et dictionnaires" »), 525 p.

AVANZI Mathieu, 2018a, *Les carabistouilles de Macron – Français de nos régions*, https://francaisdenosregions.com/2018/04/12/les-carabistouilles-de-macron/ , 2018.

AVANZI Mathieu, 2018b, *Variations sur les dénominations du kebab*, https://francaisdenosregions.com/2018/09/11/variations-sur-le-kebab/ , 2018, consulté le 28 mai 2019.

AVANZI Mathieu, 2017a, *Les régionalismes du Nord-Pas-de-Calais et de Wallonie*, https://francaisdenosregions.com/2017/03/01/les-regionalismes-du-nord-pas-de-calais-et-de-wallonie/ , 1 mars 2017, consulté le 26 mai 2019.

AVANZI Mathieu, 2017b, *Atlas du français de nos régions*, s.l., Armand Colin, 160 p.

AVANZI Mathieu, 2017c, *Les régionalismes du Nord-Pas-de-Calais et de Wallonie*, https://francaisdenosregions.com/2017/03/01/les-regionalismes-du-nord-pas-de-calais-et-de-wallonie/ , 2017.

AVANZI Mathieu, 2017d, *La « chasse aux belgicismes » cartographiée*, https://francaisdenosregions.com/2017/12/08/la-chasse-aux-belgicismes-cartographiee/ , 2017.

AVANZI Mathieu, 2017e, *L'évolution des dénominations de la 'fête de village' (1906-2016) – Français de nos régions*, https://francaisdenosregions.com/2017/09/19/les-denominations-de-la-fete-de-village/ , 2017.

AVANZI Mathieu, 2017f, *Ce que les Suisses, les Belges et les Québécois ne disent pas comme les Français : le cas du téléphone – Français de nos régions*, https://francaisdenosregions.com/2017/04/25/smartphone-natel-gsm_/ , 2017.

AVANZI Mathieu, 2017g, *L'évolution des dénominations de la 'fête de village' (1906-2016) – Français de nos régions*, https://francaisdenosregions.com/2017/09/19/les-denominations-de-la-fete-de-village/ , 2017, consulté le 26 mai 2019.

AVANZI Mathieu, 2016a, *Des régionalismes du parler lyonnais (et alentours)*, https://francaisdenosregions.com/2016/12/06/des-regionalismes-du-parler-lyonnais-et-alentours/ , 2016.

AVANZI Mathieu, 2016b, *Pain au chocolat, petit pain ou chocolatine ?*, https://francaisdenosregions.com/2016/09/01/pain-au-chocolat-ou-chocolatine/ , 2016.

AVANZI Mathieu, BARBET Cécile, GLIKMAN Julie et PEUVERGNE Julie, 2016, « Présentation d'une enquête pour l'étude des régionalismes du français », s.l., EDP Sciences, vol.27.

BALOUTA Evangelia, 2017, « Étude de la vitalité des mots régionaux du français », 2017, p. 106.

BLANCHET Philippe et BULOT Thierry, 2013a, *Une introduction à la sociolinguistique : pour l'étude des dynamiques de la langue française dans le monde / Thierry Bulot et Philippe Blanchet.*, s.l., Editions des archives contemporaines, 166 p.

BLANCHET Philippe et BULOT Thierry, 2013b, *Une introduction à la sociolinguistique : pour l'étude des dynamiques de la langue française dans le monde / Thierry Bulot et Philippe Blanchet.*, s.l., Editions des archives contemporaines, 166 p.

CARLI Augusto, 2007, « Pour une mise à jour du concept de vitalité linguistique », s.l.

CHEVILLET François, 1991, *Les variétés de l'anglais*, Paris, Nathan (coll. « Collection créée par Henri Mitterand#Série "Langues étrangères" »), 231 p.

COOKE Martin, BARKER Jon, LECUMBERRI ML Garcia et WASILEWSKI K., 2013, « Crowdsourcing in speech perception » dans *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*, s.l., p. 137–172.
DONOSO Gonzalo et SANCHEZ David, 2017, « Dialectometric analysis of language variation in Twitter », s.l., Association for Computational Linguistics.
DUBOIS Jean, 1965, *Grammaire structurale du français, nom et pronom*, Ed. revue et corr., Paris, Larousse (coll. « Grammaire structurale du français »).
FRANCARD Michel (ed.), 2015, *Dictionnaire des belgicismes*, 2e édition revue et augmentée., Louvain-La-Neuve, De Boeck Duculot, 406 p.
GLESSGEN Martin et THIBAULT André, 2005a, *La lexicographie différentielle du français et le Dictionnaire des régionalismes de France: actes du colloque en l'honneur de Pierre Rézeau*, Strasbourg, Presses Univ. de Strasbourg, 314 p.
HATMI Mohamed, 2014, « Reconnaissance des entités nommées dans des documents multimodaux », 2014, p. 154.
HUANG Yuan, GUO Diansheng, KASAKOFF Alice et GRIEVE Jack, 2016, « Understanding U.S. regional linguistic variation with Twitter data analysis », *Computers, Environment and Urban Systems*, 2016, vol. 59, p. 244‑255.
LABOV William, 2006, *The social stratification of English in New York City*, 2nd ed., Cambridge [UK] ; New York, Cambridge University Press, 485 p.
MCCANDLESS Mike, 2012, *Automatically exported from code.google.com/p/chromium-compact-language-detector*, s.l.
RAKOTOMALALA Ricco, *Enseignements de Ricco RAKOTOMALALA à la Faculté de Sciences Economiques*, http://eric.univ-lyon2.fr/~ricco/cours/index.html.
RÉZEAU Pierre (ed.), 2001, *Dictionnaire des régionalismes de France: géographie et histoire d'un patrimoine linguistique*, Bruxelles, De Boeck/Duculot, 1140 p.
SCHEFFLER Tatjana, GONTRUM Johannes, WEGEL Mathias et WENDLER Steve, 2014, « Mapping German Tweets to Geographic Regions », 2014, p. 8.
SÉGUY Jean, 1973, « Les Atlas linguistiques de la France par régions », *Langue française*, 1973, vol. 18, nº 1, p. 65‑90.
THIBAULT André et KNECHT Pierre, 2004, *Dictionnaire suisse romand: particularités lexicales du français contemporain*, Nouv. éd. rev. et augm., Genève, Zoé, 885 p.
WARNANT Léon, 1973, « Dialectes du français et français régionaux », *Langue française*, 1973, vol. 18, nº 1, p. 100‑125.
Nltk.tokenize package — NLTK 3.4.1 documentation, https://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.casual.TweetTokenizer , 2019.
French · spaCy Models Documentation, https://spacy.io/models/fr.
Data Visualization Software in the Cloud | Qlik Cloud, https://www.qlik.com/us/products/qlik-sense/qlik-sense-cloud, consulté le 27 mai 2019.

Annexe 1 : Collecte de tweets

Annexe 2: Première annotation des tweets

Annexe 3: Descripteurs sélectionnés par le modèle d'appprentissage et leur poids

meta_score	0.083260
tN	0.071910
le regio	0.042526
maybe_in_tN	0.035574
pourtant_maj	0.030804
length	0.030020
dire regio	0.023722
reg_quote	0.022620
au chocolat	0.018127
alain regio	0.017521
pain au chocolat	0.014719
reg_hashtag	0.013653
regio de	0.013423
= regio	0.013242
un regio	0.013183
pain au	0.012356
poignée de porte	0.011401
maybe_tN	0.010788
user_mention2	0.010789
poignée de	0.010207
pont de	0.009534
mon regio	0.009215
de porte	0.008058
regio '	0.007873
ou regio	0.007024
pont de regio	0.006820
dit regio	0.005598
regio le	0.005291
regio =	0.005279
' regio '	0.005141
# .	0.004997
hashtag2	0.004940
url2	0.004886
= regio =	0.004631
alain regio ,	0.004500
on dit	0.004402
' regio	0.004300
le regio de	0.004259
pour le	0.004244
on dit regio	0.004170
regio ou	0.003623
j' avoir	0.003563
regio ?	0.003409
il regio	0.003283
on dire	0.003171
un poignée	0.003126
parler de	0.003054