Caractérisation des registres de langues par extraction de motifs séquentiels émergents

Jade Mekki

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Jade MEKKI

Caractérisation de registres de langue par extraction de motifs séquentiels émergents

Thèse prévue à Nanterre, le 08/09/2022
Unité de recherche : UMR 6074 IRISA

Rapporteurs avant soutenance :

Farah BENAMARA — Maître de conférences, HDR, Université Paul Sabatier
Thierry CHARNOIS — Professeur des Universités, Université Paris 13 Nord

Composition du Jury :

Rapporteurs :	Farah BENAMARA	Maître de conférences, HDR, Université Paul Sabatier
	Thierry CHARNOIS	Professeur des Universités, Université Paris 13 Nord
Examinateurs :	Jean-Yves ANTOINE	Professeur des Universités, Université François Rabelais
	Olivier BAUDE	Professeur des Universités, Université Paris Nanterre
	Dominique LEGALLOIS	Professeur des Universités, Université Sorbonne Nouvelle
Dir. de thèse :	Damien LOLIVE	Maître de conférences, HDR, Université de Rennes 1
Co-dir. de thèse :	Delphine BATTISTELLI	Professeur des Universités, Université Paris Nanterre
	Gwénolé LECORVÉ	Chercheur, HDR, Orange (précédemment Université de Rennes 1)
Encadrant de thèse :	Nicolas BÉCHET	Maître de conférences, Université de Bretagne-Sud

Je marchais le coeur battant, la gorge sèche, et si parfait autour de moi était le silence de pierre, si compact le gel insipide et sonore de cette nuit bleue, si intriguants mes pas qui semblaient poser imperceptiblement au dessus du sol de la rue, je croyais marcher au milieu de l'agencement bizarre et des flaques de lumière égarantes d'un théâtre vide - mais un écho dur éclairait longuement mon chemin et rebondissait contre les façades, un pas à la fin comblait l'attente de cette nuit vide, et je savais pour quoi désormais le décor était planté.

Le Rivage des Syrtes, Julien Gracq

Remerciements

En premier lieu, mes remerciements vont à Farah Benamara, Thierry Charnois, Olivier Baude, Dominique Legallois et Jean-Yves Antoine qui ont examiné ma thèse et permis sa soutenance. Je tiens à remercier particulièrement Jean-Yves Antoine ainsi qu'Iris Eshkol pour avoir accepté de suivre mes travaux en participant à mon comité de suivi de thèse.

Plus largement, j'adresse mes remerciements à tous les membres de l'équipe *Expression* du laboratoire IRISA ainsi qu'à ceux du laboratoire MoDyCo. Leurs accueils m'ont donné un cadre professionnel stable et bienveillant tout au long de ma thèse.

Au delà du cadre professionnel, je tiens à remercier mes ami.e.s : Alice, Aline, Anna, Arthur, Benjamin, Camilla, Claire, Clara, Gabriele, Julia, Laura, Lucie, Lucille, Marine et Mingqiang. Vous avez tous participé à cette entreprise en annotant mes données, en faisant les évaluations perceptuelles, en relisant/corrigeant articles et manuscrit, en m'entraînant aux divers oraux et en discutant (plus ou moins tard dans la nuit) de ce que sont les registres de langue.

Evidemment je remercie également ma famille de m'avoir toujours soutenue, notamment Solange, mon frère et mes parents. Solange, je vous remercie de m'avoir donné un espace de travail serein durant les mois de confinement ainsi que d'avoir relu et corrigé ce manuscrit de thèse en un temps record. Jules, papa et maman, vous avez toujours été mes modèles de force, de volonté et de résilience. Vous êtes mon socle et mon point de départ. Merci.

Mickael, je ne trouverai pas de mots à la hauteur de ma reconnaissance. Ces années de thèse nous les avons vécues ensemble et rien ne pourrait expliquer ce que cela signifie pour moi. Ce travail est le nôtre et de cela j'en suis profondément fière. Merci d'avoir tenu à mes côtés.

Enfin, j'adresse mes derniers remerciements – mais non les moindres – à mes directeurs de thèse : Delphine, Gwénolé et Nicolas. Ces remerciement vont bien au delà des formalités exigées par l'exercice car j'aimerais sincèrement vous remercier pour votre confiance, votre patience, votre exigence, votre écoute et votre bienveillance. J'ai toujours admiré votre travail et j'ai été heureuse de faire cette thèse sous votre direction. Merci, notre rencontre aura marqué ma vie professionnelle et personnelle.

TABLE DES MATIÈRES

Chapitre 1

Introduction

1.1 Contexte d'étude

Le locuteur d'une langue sent que pour un même message il existe plusieurs manières de le dire. Ce phénomène linguistique est celui des registres de langue. Ils renvoient à un trait saillant du langage que tout locuteur saisit intuitivement. Par exemple, la différence entre ces trois phrases, qui pourtant veulent dire la même chose, est nettement perceptible : « Je kiffe bouffer des churros. », « J'adore manger des churros. », « Consommer des churros m'enthousiasme. ». La différence perçue résulte d'un phénomène de variation linguistique : les trois phrases font des écarts par rapport à un repère donné. La vision la plus simple et la plus répandue se base sur la norme grammaticale [1] pour juger les productions linguistiques. Toute production la suivant est « correcte » et toute production ne la suivant pas est « fautive ». Toutefois, aucun consensus sur la norme linguistique n'émerge de la littérature scientifique. De ce fait, les registres de langue renvoyant à un phénomène de variation linguistique ne trouve pas de définition stable dans la littérature.

Plusieurs travaux sociolinguistiques ont étudié les registres de langue pour comprendre à quels cadres humains (sociologique, culturel, etc.) ils correspondent. D'autres travaux en linguistique de corpus ont cherché comment ils sont identifiables. Que cela soit en sociolinguistique ou bien en linguistique de corpus, ces approches ont trois limites principales :

1. Les données utilisées sont peu nombreuses (comme dans Favart 2010) ; or une étude conduite à partir d'un corpus de petite taille réduit sa capacité à généraliser ses résultats.

2. Les niveaux d'analyse de la langue ne sont jamais analysés conjointement (comme dans Červenková 2014) ; or les relations entre niveaux sont aussi intéressantes. Par exemple, un mot qui se termine en « -asses » ne sera sûrement pas du même

1. Historiquement, c'est l'Académie française fondée en 1635 par Richelieu qui a pour but de fixer le bon usage à suivre.

registre s'il est un nom commun (« C'est de la vinasse. »), ou bien un verbe (« Qu'il le fasse bien. »).

3. Les registres de langue sont étudiés un à un (comme dans Ilmola 2012) ; or nous pensons que la perception des registres se fait par distinction entre eux. Par exemple, un énoncé courant pourra être perçu comme soutenu face à un registre familier ; et inversement pourra être perçu comme familier face à un registre soutenu.

Pour répondre à ces limites, nous proposons une approche explorant automatiquement les registres de langue. Elle se fonde sur un large corpus de textes, considère divers niveaux d'analyse de la langue en même temps et caractérise les registres de manière comparative.

1.2 Problématique et objectifs de la thèse

Le premier et principal objectif de cette thèse est de caractériser automatiquement des registres de langue en faisant l'hypothèse qu'ils sont identifiables par des traits particuliers qui les distingueraient les uns des autres. L'objectif secondaire est de proposer une chaîne de traitement suffisamment générique pour caractériser d'autres phénomènes linguistiques à partir de jeux de données contrastées comme des avis positifs vs. des avis négatifs par exemple.

Avant de caractériser notre objet d'étude, les registres de langue, nous avons dû le circonscrire car il ne trouve pas de définition stable dans la littérature. En effet, les registres de langue renvoient à la perception d'un phénomène de variation linguistique. Cette variation opère des écarts par rapport à une norme linguistique. Dès lors, la définition des registres de langue est liée à celle de la norme linguistique. Or, la définition de la norme linguistique soulève des questions sociolinguistiques et linguistiques : *La norme doit-elle être descriptive et basée sur l'usage ? Ou bien prescriptive et imposée par un organisme institutionnel ? La norme linguistique est-elle la même à l'oral et à l'écrit ?* Selon les réponses données à ces questions, les définitions de la norme et des registres diffèrent.

Pour caractériser un registre A (appelé registre cible), il nous faut ainsi établir une liste de descripteurs linguistiques qui le distingue d'un registre B (appelé registre source). Ce que nous appelons « descripteurs linguistiques » est un patron linguistique décrivant une séquence de une ou plusieurs unités linguistiques, avec des traits linguistiques de divers niveaux d'analyse de la langue. Par exemple, le descripteur $\langle$(*il*, pronom personnel) + (racine, verbe) + (adverbe)$\rangle$ décrit un syntagme de 3 unités linguistiques. Les syntagmes « il mange bien », « il parle joliment » ou « il chante mal » en sont des exemples. Des

descripteurs caractérisant un registre cible A sont des descripteurs plus présents dans les textes A que dans les textes B. Par exemple, les descripteurs m_1=⟨(adverbe, terminaison en -ant)⟩ et m_2=⟨(vous, sujet)⟩ sont des descripteurs caractéristiques du registre cible soutenu, car plus présents dans le texte (1) du registre cible, que dans le texte (2) du registre source. Les exemples des descripteurs sont soulignés dans les textes.

(1) Vous avez indiscutablement une bonne connaissance du fonctionnement du FMI et des enjeux qui sont à l'œuvre. Cependant, pour ma compréhension personnelle et celle, sans doute, de bien d'autres lecteurs de ce Blogg, je voudrais que vous nous éclairiez de vos lumières concernant certaines interrogations.

(2) Franchement, je me demandais ce que je foutais là et comment j'avais eu aussi peu d'amour-propre pour venir m'asseoir en face de cet abruti...

Dans le cadre de cette thèse, nous avons souhaité constituer un corpus volumineux pour représenter au mieux les usages réels des registres par les locuteurs. Un outil automatique est nécessaire pour découvrir à partir de ce corpus les descripteurs linguistiques caractéristiques des registres de langue. Les techniques de fouille de motifs séquentiels émergents permettent d'extraire des régularités caractéristiques d'un ensemble de données par rapport à un autre. La fouille de motifs séquentiels émergents présente trois principaux avantages : le respect de la notion d'ordre entre les mots, la représentation de chaque unité linguistique par différents traits linguistiques (partie grammaticale du discours, fonction syntaxique, lemme, etc.), l'émergence des descripteurs de manière comparative. Cependant, divers verrous limitent son utilisation. Tout d'abord, la qualité des motifs séquentiels émergents extraits est difficilement évaluable sans base de référence à laquelle les comparer. Ensuite, les algorithmes d'extraction de motifs séquentiels émergents nécessitent des ressources en calcul très importantes empêchant l'aboutissement de certaines extractions. Enfin, ils extraient un nombre de motifs séquentiels émergents très élevé et redondants entre eux entravant l'interprétation des résultats. Dès lors, pour caractériser les registres de langue par extraction de motifs séquentiels émergents, nous avons proposé dans cette thèse un ensemble de stratégies permettant à l'extraction de passer l'échelle et assurant l'interprétabilité des motifs séquentiels émergents. La chaîne de traitement proposée essaie de poser le moins d'*a priori* possible afin de pouvoir l'appliquer à d'autres phénomènes linguistiques, tels que les dialectes variant selon les zones géographiques ou les manières de parler propres à différentes époques historiques. Notre travail apporte quatre contributions principales :

1. Un large corpus de tweets constitué pour représenter les registres de langue : il compte 228 505 tweets en français annotés selon la tripartition de registres de langue que nous avons décidé de retenir (familier vs. courant vs. soutenu) [2]. L'annotation manuelle a été encadrée par un guide d'annotation pour les registres [3] dans lequel nous avons exposé un nouvel ensemble de descripteurs linguistiques pour l'analyse de **C**ommunications **M**édiées par **O**rdinateurs (CMO).
2. Une méthodologie pour vérifier la qualité des motifs séquentiels émergents extraits : elle a utilisé des données artificielles dans lesquelles nous connaissions *a priori* les motifs que nous voulions retrouver dans les motifs séquentiels émergents.
3. Un protocole de fouille de motifs séquentiels émergents à partir de données réelles est présenté. Il permet à la fouille de motifs séquentiels émergents de passer l'échelle malgré un corpus volumineux et sans poser d'*a priori* linguistique sur les motifs séquentiels émergents à obtenir.
4. Une approche organisée en deux temps pour réduire l'ensemble complet de motifs séquentiels émergents : le partitionnement des motifs séquentiels émergents et la sélection de motifs séquentiels émergents représentants. La qualité des motifs séquentiels émergents représentants a été confirmée par deux évaluations. La première a perceptuellement validé une sous-partie des motifs séquentiels émergents représentants. La seconde a automatiquement validé l'intégralité des motifs séquentiels émergents représentants.

La figure 1.1 donne une vision d'ensemble de notre chaîne de traitement. Cette dernière part de la notion de *variation linguistique* de laquelle nous avons défini les *registres de langue* et aboutit à des ensembles de motifs séquentiels émergents caractéristiques des registres de langue. Nous précisons dans la section suivante l'organisation du manuscrit qui détaille chacune des étapes de la figure 1.1.

1.3 Organisation du manuscrit

Pour situer notre étude, le chapitre 2 dresse un panorama des différentes définitions existantes dans la littérature francophone et anglophone. Il y est comparé notre conception des registres de langue avec celles données dans des travaux sociolinguistiques ou

2. Le corpus est accessible via ce lien : `http://tremolo.irisa.fr/tremolo-tweets-corpus/`.
3. Le guide d'annotation est disponible via ce lien : `https://hal.archives-ouvertes.fr/hal-03218217`.

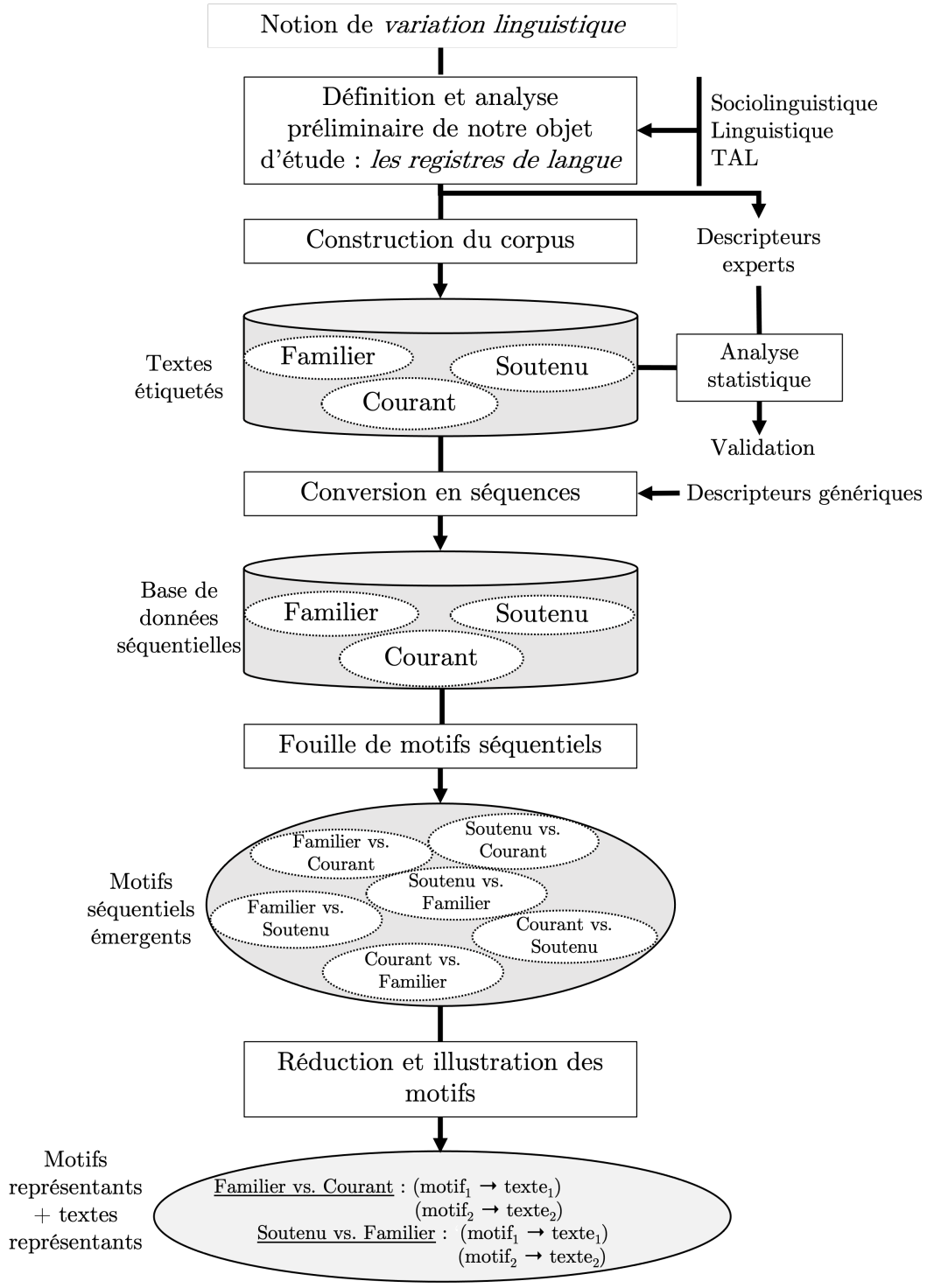

FIGURE 1.1 – Vision globale de notre chaîne de traitement pour la caractérisation des registres de langue.

linguistiques. Cette comparaison est l'occasion de souligner la diversité des variables pouvant être prises en compte pour l'analyse des registres de langue : le niveau social d'un locuteur, son réseau social, son statut hiérarchique dans une communauté, etc. Face à la profusion des approches, ce premier chapitre situe nos travaux. Il pose notre définition de la norme linguistique de laquelle découle notre définition des registres de langue en français.

Après avoir circonscrit notre objet d'étude, un corpus de textes est constitué pour en extraire les motifs linguistiques caractéristiques des registres de langue. Le chapitre 3 détaille la constitution linguistiquement motivée d'un corpus de tweets, appelé TREMoLo-Tweet, et son annotation en registres. Pour annoter l'ensemble des tweets, un modèle d'apprentissage automatique semi-supervisé est utilisé. Il généralise l'annotation manuelle d'un sous-ensemble de tweets encadrée par un guide d'annotation. Ce guide d'annotation est issu d'une analyse linguistique fine d'un sous corpus de tweets et de la prise en compte de divers travaux linguistiques ou TAL qui se sont déjà penchés sur l'analyse de ce type de textes (les CMO) entre autres. Il propose ainsi notamment des descripteurs linguistiques pour l'étude des CMO habituellement écartés comme les hashtags ou bien les pictogrammes. La qualité de l'annotation automatique, qui étend l'annotation manuelle à l'ensemble du corpus, est garantie par les résultats expérimentaux obtenus.

Le chapitre 4 présente les différents types de motifs proposés dans la littérature. Nous y montrons en quoi les motifs séquentiels émergents sont les plus pertinents pour caractériser les registres de langue. Notre chaîne de traitement vise à obtenir un ensemble de motifs séquentiels émergents de taille raisonnable afin de pouvoir les interpréter. Elle commence par la validation des techniques de fouille de motifs séquentiels émergents comme pertinentes pour notre tâche. Le chapitre 5 expose comment des données artificielles ont été utilisées pour vérifier la robustesse des motifs séquentiels émergents, avant de les extraire à partir de données réelles. Plusieurs contraintes croisées ont guidé le protocole de fouille de motifs à partir du corpus TREMoLo-Tweet. La première a été de réussir le passage à l'échelle de l'extraction de motifs séquentiels émergents tout en considérant l'intégralité du corpus dont la taille risquait de faire exploser le coût algorithmique. La seconde contrainte a été de ne pas poser d'*a priori* linguistique sur les motifs à découvrir. La dernière contrainte a été d'obtenir un ensemble de motifs séquentiels émergents le moins volumineux possible. Pour respecter ces contraintes nous avons joué sur diverses variables : le type de motifs (utilisation de motifs clos pour réduire le nombre de motifs sans perte d'information) ; le nombre et le type de traits linguistiques (utilisation de peu

de traits mais avec une dimension informative forte comme les sous-mots : *_Fille, ttes_*) ; les paramètres d'extraction (utilisation de la contrainte de gap pour écarter tous les motifs non contiguës). Notre protocole d'extraction est principalement validé par le passage à l'échelle de l'algorithme de fouille de motifs séquentiels émergents malgré la taille du corpus utilisé.

L'ensemble de motifs séquentiels émergents obtenu contient un nombre très élevé de motifs. Pour avoir un sous-ensemble de motifs séquentiels émergents représentants plus interprétable, nous avons utilisé des techniques de partitionnement de données. Dans un premier temps, les motifs sont regroupés selon leurs similarités. Le chapitre 6 introduit l'algorithme RGMSE (pour **R**e**G**roupement de **M**otifs **S**équentiels **É**mergents) qui regroupe les motifs sans fixer *a priori* le nombre de groupes à obtenir. Les partitions obtenues sont validées à l'aide de mesures vérifiant si les groupes de motifs séquentiels émergents sont bien différents entre eux, et si les motifs séquentiels émergents d'un même groupe sont bien similaires entre eux. Dans un second temps, pour chaque groupe de motifs séquentiels émergents, un motif représentant est sélectionné. La capacité de ces motifs représentants à caractériser les registres de langue a été validée par deux évaluations. La première est une évaluation perceptuelle qui a examiné une sous-partie des motifs séquentiels émergents représentants. La seconde est une évaluation automatique qui a testé la totalité des motifs représentants en utilisant des classifieurs binaires entraînés à distinguer deux registres. Ces classifieurs se sont entraînés à partir des motifs représentants utilisés comme descripteurs d'apprentissage. Les résultats obtenus confirment la capacité des motifs séquentiels émergents représentants à bien caractériser les registres de langue.

Le chapitre 7 de ce manuscrit montre que notre chaîne de traitement peut être utilisée pour caractériser d'autres phénomènes linguistiques à partir de données contrastées, en l'occurrence des genres de textes adressés aux enfants. Nous y exposons nos dernières expériences validant la robustesse de notre méthodologie et montrant la possibilité de l'appliquer pour d'autres cas d'usage.

Enfin, nous dressons un bilan de la thèse dans lequel nous revenons sur nos principales contributions. Puis, nous concluons ce manuscrit en proposant diverses pistes de futurs travaux utilisant notre approche.

Chapitre 2

La variation linguistique

Sommaire

Selon la situation de communication et ses objectifs, un locuteur choisit des manières d'écrire ou de parler différentes. Il adapte son énoncé en choisissant certaines formes linguistiques plutôt que d'autres. Par exemple, à la fin d'une lettre de motivation un candidat n'écrira pas « A+ » ou « tchao », ces expressions sont réservées à un cadre privé. Il restera plus formel en employant des formulations de politesse comme « cordialement » ou « respectueusement ». Les registres de langue renvoient au fait de percevoir « différentes façons de dire la même chose », c'est-à-dire de discerner des « variations linguistiques » (Labov 1988). Les textes (3), (4) et (5) [1] constituent des exemples de textes appartenant à trois registres différents. Ils montrent que les registres de langue sont un phénomène linguistique saillant : même avec des extraits de textes très courts et sans contexte, des registres sont nettement perceptibles ; on pourrait ici par exemple déjà dire que (3) tend à être identifié comme plutôt familier, (4) comme plutôt neutre, (5) comme plutôt soutenu, sans pour le moment chercher à donner une définition précise de cette opposition entre familier, neutre et soutenu.

(3) Bon, comme tout le monde doit le savoir, j'habite depuis bientôt 3 ans, au dessus d'un fou à lier, généralement dénommé le "Connard du premier". Bientôt la quille, nous déménageons pour le 29/12.

1. Les extraits sont issus de notre corpus TREMoLo-web introduit section 3.3.

(4) Le cimetière des éléphants Tempête sur la SPA dont les dirigeants sont mis en examen selon un scénario qui rappelle point par point celui de l'ARC. Le scandale est de même nature et de même portée.

(5) Peut-être, maintenant que mon cœur, incapable de vouloir et de supporter de son plein gré la souffrance, ne trouvait qu'une seule solution possible, le retour à tout prix d'Albertine, peut-être la solution opposée (le renoncement volontaire, la résignation progressive) m'eût-elle paru une solution de roman, invraisemblable dans la vie, si je n'avais moi-même autrefois opté pour celle-là quand il s'était agi de Gilberte.

Les registres de langue semblent être un phénomène intuitivement reconnaissable et aisé à saisir. Ils sont, d'ailleurs, enseignés dès l'école primaire à partir du CE1. Cependant, il n'existe aucun consensus sur leur définition dans la littérature scientifique. Cela représente une difficulté majeure pour nos travaux : comment extraire des motifs linguistiques caractéristiques des registres de langue sans définition stable de ces derniers ? Pour répondre à cette question il faut comprendre pourquoi, malgré la facilité avec laquelle les registres sont perçus, la communauté scientifique n'arrive pas à tomber d'accord sur leur définition. Trois éléments principaux peuvent y répondre : le rôle de la norme linguistique dans la perception des registres de langue ; le croisement de variables extra-linguistiques avec leur analyse ; la représentation de l'espace linguistique comme un continuum allant de la modalité orale vers la modalité écrite. Ce chapitre détaille chacun de ces éléments, afin de positionner nos travaux et de délimiter notre objet d'étude.

La première raison à l'absence de consensus sur la définition des registres est l'absence de consensus sur la définition de la norme linguistique. Comme une norme est nécessaire pour évaluer les variations linguistiques, sa définition est liée à celle des registres. Une évaluation sommaire situera dans le registre familier un énoncé ne suivant pas la norme, et dans les registres courant et soutenu un énoncé la suivant. Mais comment s'établit cette norme linguistique ? Est-ce l'usage qui fait la norme, ou bien la norme qui doit faire l'usage ? Par exemple, l'adverbe « cordialement » est passé du registre familier au registre soutenu. Au XVII^{e} siècle, « cordialement » avait une signification plus intime qu'aujourd'hui [2]. Aujourd'hui « cordialement » est majoritairement utilisé dans des échanges professionnels. Alors, est-ce parce que l'usage a glissé vers un registre plus soutenu que sa définition a été modifiée ? Ou bien, est-ce que parce que les dictionnaires en ont donné une

2. Il était utilisé pour dire "Du fond du cœur, dans tout l'élan de la vie morale la plus intime." selon la définition extraire du *TLFi*.

nouvelle définition [3] que les usages ont changé ? Dans la première section de ce chapitre, nous nous positionnons par rapport à la littérature en introduisant notre définition de la norme linguistique associant règles d'usage et règles grammaticales, après avoir dressé un panorama des travaux sur le sujet.

La deuxième difficulté pour définir les registres se trouve dans la diversité des variables prises en compte lorsqu'ils sont définis. Selon le domaine de l'étude (linguistique, sociolinguistique), et son aire linguistique (anglophone, francophone), les registres sont associés à diverses variables telles que la classe sociale du locuteur, son niveau scolaire, sa communauté, son aire géographique, etc. Cela résulte en une multitude de définitions illustrées dans la littérature par la co-existence des termes *niveau de langue*, *registre* et *style*. Nous présentons dans la section 2.2 des travaux en sociolinguistique, puis en linguistique, qui détaillent l'utilisation et le sens de ces terminologies. Nous nous positionnons par rapport à cette dernière en choisissant le terme *registre de langue*. Ce choix repose sur le souhait d'éviter toute notion normative et hiérarchisante qui pourrait se retrouver dans le terme *niveau de langue*, et la connotation francophone du terme *style* trop spécifique à un style individuel.

Enfin, le troisième élément de réponse demeure dans la représentation de l'espace linguistique dans lequel sont localisés les registres. Pour obtenir le nombre de registres distincts, l'espace est segmenté en diverses partitions. Dans la section 2.3, nous montrons dans un premier temps que cet espace linguistique est généralement représenté par un continuum allant de la modalité orale vers la modalité écrite. De nombreux travaux associent l'oral avec le registre familier, et l'écrit avec les registres courant et soutenu. Cette association indexe les registres sur la maîtrise de l'écriture. Or, nous ne pensons pas que l'oral devrait être systématiquement associé au registre familier, et la maîtrise de l'écrit aux registres courant et soutenu : il est possible d'utiliser les registres familier et soutenu à l'oral, tout comme à l'écrit. En outre, nous montrons que la caractérisation des modalités orale et écrite devient inconsistante face aux nouveaux moyens de communication. Le continuum que nous proposons est un espace linguistique permettant de les prendre en compte. Dans un second temps, nous détaillons plusieurs exemples de partitions variant selon les travaux. De toutes ces partitions, la plus répandue est la tripartition en registres familier, courant et soutenu. Comme elle est également la plus enseignée en milieu scolaire, nous avons choisi de caractériser les registres de langue en comparant ces trois registres

3. La 9^{e} édition du dictionnaire de l'Académie Française définit « cordialement » comme « De manière cordiale, formule de politesse par laquelle on conclut une lettre. ».

entre eux.

2.1 Notion de norme linguistique

L'objectif de cette section est de comparer diverses approches étudiant la norme linguistique pour introduire notre propre définition. Notre définition réunit les normes d'usage et grammaticale systématiquement séparées dans les autres définitions proposées. Les réunir évite d'évaluer les variations linguistiques en se basant uniquement sur la notion de fautes grammaticales. La comparaison des travaux sur la norme linguistique met en lumière la difficulté pour la définir, et les enjeux socio-politiques liés à sa définition. Dans la section 2.1.1 suivante, nous montrons que les diverses définitions de la norme reposent sur deux choix : le choix des valeurs qu'une langue devrait incarner, le choix de l'entité légitime pour édicter la norme. Puis, nous exposons en section 2.1.2 comment certaines approches abordent la norme à travers la définition d'un français standard comme variété linguistique de référence.

2.1.1 Norme(s) linguistique(s)

Choisir une norme linguistique revient à choisir un ensemble de règles à suivre. Le contenu de ces règles diffèrent selon l'entité en charge de leur établissement et selon l'idée de ce que doit être la langue française. Par exemple, des plaidoyers comme *Tirons la langue : plaidoyer contre le sexisme dans la langue française* de (Borde 2018) ou *Pour une langue sans sexisme* de (Labrosse 2021) voient la langue française comme une langue sexiste avec des règles grammaticales telles que le genre masculin qui l'emporte sur le genre féminin au pluriel. Ils remettent en question la norme linguistique actuelle résultante d'une vision patriarcale de la société puisqu'établie par des grammairiens masculins. Cet exemple montre que le choix d'une norme linguistique équivaut à se positionner sur des questions socio-politiques. Illustrant différentes prises de position, plusieurs types de normes existent dans la littérature. Cette section détaille des typologies de normes, avant de présenter notre propre définition de la norme linguistique.

Types de norme linguistique De nombreux travaux étudient la norme linguistique, pour n'en citer qu'une infime partie : (Rey 1972 ; Prikhodkine 2011 ; Pöll et Schafroth 2010 ; Moreau 1999 ; Maurais 2008 ; Marcellesi 1976 ; Manessy 1994 ; Lafontaine 1986 ; etc.) Parmi eux, nous retenons trois études : celle de (Gadet 2007), celle de

(Paveau et Rosier 2008) et celle de (Ledegen 2021). Nous les avons choisies car ces trois études proposent des typologies partageant des types communs de normes. Dans la première typologie, (Gadet 2007) oppose deux types de norme linguistique : la norme objective et la norme subjective. La norme objective renvoie à une contrainte issue d'un organisme officiel qui édicte les règles à suivre pour un bon usage de la langue. La norme subjective renvoie à une contrainte collective issue des jugements de valeurs sur les pratiques linguistiques d'une communauté de locuteurs. Elle a pour effet de renforcer la cohésion sociale de cette communauté. La principale limite de cette proposition est la mise en opposition des deux normes. Isolées, elles ne peuvent pas conduire à une langue moderne et commune. La norme objective risque d'être toujours en retard par rapport aux usages réels de la langue, car longue à être établie. Prenons, par exemple, l'édition actuelle du dictionnaire de l'Académie française publiée en 2011. 11 années ont été nécessaires pour mettre à jour la 8^{e} édition de 1935. La norme subjective, quant à elle, risque d'établir des règles linguistiques trop spécifiques à une communauté entravant la compréhension entre locuteurs de communautés différentes. Il faudrait associer ces deux normes pour assurer un ensemble de règles communes à l'ensemble des locuteurs et adaptables aux nouveaux usages.

Dans (Ledegen 2021) l'auteur reprend ces types de normes en les intégrant dans une typologie plus nuancée qui en distingue cinq au total. Elle montre graduellement comment la norme peut être un outil de description ou de prescription :

1. La norme objective, qui désigne les habitudes linguistiques partagées au sein d'une communauté.
2. La norme descriptive, qui explicite les normes objectives définies ci-dessus.
3. La norme prescriptive, qui donne un ensemble de normes objectives comme le modèle à suivre.
4. La norme subjective, qui concerne les attitudes et représentations linguistiques, et attachent aux formes des valeurs esthétiques, affectives, ou morales : élégant vs. vulgaire, chaleureux vs. prétentieux...
5. La norme fantasmée, qui est « un ensemble abstrait et inaccessible de prescriptions et d'interdits que personne ne saurait incarner et pour lequel tout le monde est en défaut » (Baggioni et Moreau 1997 dans Ledegen 2021).

La limite de cette typologie est qu'elle n'envisage pas la norme descriptive comme imposant des règles à suivre. Or, les locuteurs d'une même communauté suivent ces règles

implicites pour pouvoir être compris. La norme descriptive a donc une action prescriptive au même titre que la norme prescriptive. Dans la dernière typologie, (Paveau et Rosier 2008) décrit des types de *postures*[4]. Trois types de postures sont distingués :

1. La posture normative, qui est fondée sur le respect du bon usage tel qu'il est défini et conservé dans les grammaires et les dictionnaires, et tel qu'il s'exprime au sein de conseils et prescriptions pour parler une belle et bonne langue.
2. La posture puriste, qui se caractérise par une forte prégnance de valeurs esthétiques, politiques, pseudo-linguistiques et métaphoriques.
3. La posture scientifique, qui est revendiquée par la linguistique depuis Saussure, pour laquelle la norme légitime est celle, interne, des règles du système de la langue.

Trois types de normes se retrouvent dans toutes ces typologies : la norme prescriptive qui édicte des règles grammaticales de bon usage à suivre ; la norme subjective qui édicte des règles empreintes de valeurs idéologiques et esthétiques sur ce que devrait être la langue ; la norme objective qui dégage des règles à partir de l'usage de la langue. Partant de ce constat, nous avons travaillé sur une définition de la norme afin d'évaluer les variations linguistiques. Nous ne voulions pas baser la perception de la variation sur la notion de faute grammaticale, car cette dernière est insuffisante pour l'analyse des registres. Si le repère pour évaluer les variations linguistiques était la norme grammaticale, alors comment différencier les registres courant et soutenu la suivant tous les deux ? Une forme correcte grammaticalement mais inadaptée à la situation de communication peut-elle être perçue comme du registre soutenu ? Imaginons l'utilisation de la formulation « Je vous prie d'agréer, Messieurs, l'expression de mes sentiments respectueux » pour saluer une foule après un discours politique. La formulation ne sera pas perçue comme complètement soutenue bien qu'elle respecte la norme grammaticale. En effet, son utilisation ne suit pas la norme d'usage qui la réserve aux échanges écrits. Le non respect de l'usage crée une impression d'erreur limitant la reconnaissance du registre soutenu. Cet exemple met en avant le besoin d'une norme d'usage complémentaire à la norme grammaticale pour l'analyse des registres de langue.

Notre définition de la norme Dans le cadre de notre travail, les enjeux politiques, voire idéologiques, sous-jacents à l'établissement d'un modèle linguistique ne sont pas pris en compte, car hors de nos objectifs. C'est pourquoi nous n'avons pas considéré la

4. *Posture* est définie par le TFLi comme suit : "Attitude morale de quelqu'un".

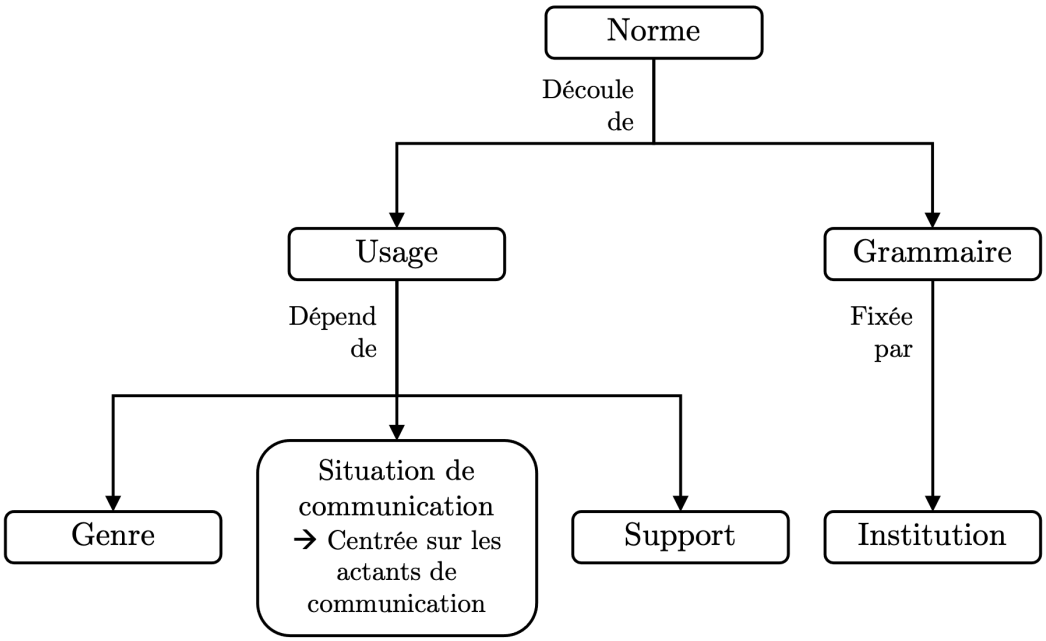

FIGURE 2.1 – Caractérisation des actions normatives

norme subjective. En revanche, nous pensons que la norme prescriptive est nécessaire pour assurer la compréhension entre locuteurs. Nous pensons également que la norme objective est importante pour permettre aux règles linguistiques de s'adapter aux nouveaux usages. Notre définition de la norme découle donc à la fois de la norme objective en se basant sur l'usage, et de la norme prescriptive en se basant sur la grammaire. La figure 2.1 explicite que l'usage dépend selon nous de trois variables : le genre (discours politiques, SMS, etc.), la situation de communication (milieu familial, milieu professionnel, etc.) et le support (oral, écrit sur papier, numérique, etc.). Différents usages émergent des règles linguistiques qui ne sont pas forcément formalisées et rédigées dans un document officiel tel qu'un dictionnaire, une convention ou bien une grammaire. À l'inverse, la norme grammaticale contient des règles entérinées et décrites par une institution à caractère officiel comme l'académie française. Suivant cette définition de la norme linguistique, nous décidons de considérer un texte comme pouvant appartenir à trois registres distincts définis comme suit :

- familier : lorsque la norme grammaticale et la norme d'usage ne sont pas suivies ;
- courant : lorsqu'il se conforme partiellement à la norme grammaticale et d'usage ;
- soutenu : lorsqu'il se conforme complètement à la norme grammaticale et d'usage.

Par exemple, les textes (6), (7) et (8) sont des tweets [5], respectivement, des registres familier, courant et soutenu.

(6) Bosh il lui a fait un sal boulot à kaaris il lui a montré que maintenant y'a plus de petit et grand

Le tweet (6) est perçu comme familier. Il ne suit pas la norme d'usage dès lors qu'aucun élément propre aux tweets, tels que les hashtags ou bien la mention d'un autre utilisateur, n'est utilisé. Il ne suit pas non plus la norme grammaticale avec des fautes d'orthographe comme l'utilisation de « y'a » pour « il y a ».

(7) Un grand chelem sans Nadal Federer et Djoko, c'est comme une ligue des champions sans Messi et Ronaldo #USOpen2020

Si le tweet (7) est perçu comme courant, c'est parce qu'il respecte la norme d'usage avec l'utilisation d'un hashtag, sans respecter totalement la norme grammaticale avec la forme contracter de « cela est ».

(8) [#LeSaviezVous] Le ministère a confirmé que les résultats de la C.L.A.S. élective qui s'est tenue le 4 mars 2020 sont valides. La responsable #UNSA Police Yvelines est la nouvelle vice-présidente de la CLAS78 ! url_path

Enfin, le tweet (8) est perçu comme soutenu, car il respecte les deux normes. La norme d'usage est suivie avec l'utilisation des hashtags pour référencer son contenu, ou bien l'insertion d'URL pour renvoyer à un contenu informatif complémentaire. La norme grammaticale est parfaitement respectée avec l'absence de fautes.

2.1.2 Variété linguistique de référence

Cette section montre comment certains travaux ont abordé la question de la norme linguistique à travers celle de la variété linguistique de référence. Ce que nous appelons la « variété linguistique de référence » renvoie au français choisi comme français modèle à suivre :

> Selon une définition de R.-L. Wagner, une attitude normative « implique que l'on ait discerné des niveaux entre plusieurs manières de s'exprimer, hiérarchisé ces niveaux et conféré à l'un d'eux la dignité de modèle » (Wagner et Quemada 1969 dans Authier et Meunier 1972).

Tout comme dans la section précédente, le choix d'un modèle est accompagné de questions socio-politiques. Par exemple, les travaux de (Valdman 1982) remettent en

5. Ils sont issus de notre corpus TREMoLo-Tweet présenté section 3.4.

question le choix d'un « français standard » renvoyant à une classe sociale bourgeoise et parisienne :

> « Il devient de plus en plus malaisé de répondre à la question : quel français enseigner ? En effet, la reconnaissance de la primauté du français standard (FS), défini comme "le parler soutenu de la bourgeoisie cultivée de la région parisienne", ne fait plus l'unanimité, tant à l'intérieur qu'hors de l'hexagone. » (Valdman 1982)

Illustrant la difficulté de tomber sur un consensus sur ce que devrait être le modèle à suivre, de nombreuses appellations différentes co-existent. Parmi elles, citons le « français standard » (Valdman 1982 ; Beaufort, Roekhaut et Fairon 2008 ; Charaudeau et Maingueneau 2002 ; Guerin 2008 ; Durand et Lyche 1999 ; Rebourcet 2008) ; ou bien le « français courant » (Bonnard 1981 ; Hellermann 1969 ; Glasco 2011 ; Rigat et Piccardo 2008), ou encore le « registre neutre » (Mercier, Verreault et Lavoie 2002 ; BERTOCCHINI et Costanzo 2010 ; Juan et Zhihong 2018). Ce principe de neutralité est repris par (Gadet 2007, p. 20) qui décrit le « français standard » comme une variété se voulant neutre : « il prétend à la neutralité devant les genres discursifs ». Toutes ces terminologies renvoient à une sorte de français dit standard/courant/neutre/de base à partir duquel sont évaluées les variations linguistiques :

> « Le français "de base", est un seuil entre ce qui est formel et ce qui ne l'est pas. Il est associé à l'usage correct : une langue épurée de tout énoncé erroné. En somme, il correspond à une entité linguistique qui peut être aussi bien écrite qu'orale. Le bon français, c'est le français correct. » (Rebourcet 2008)

Une stratégie mise en place pour définir ce français standard est de l'appréhender de manière négative, par élimination de ce qu'il n'est pas : « Le français standard n'est pas le français régional mais n'est pas non plus le français oral, ni même le français populaire. » (Rebourcet 2008). Mis en perspective avec les registres de langue :

> « Le français standard devient le français incarnant la limite entre ce qui est oral et informel, et ce qui est plus formel, plus soutenu et littéraire. [...] Il [le français standard] englobe, en quelques sortes, des caractéristiques qui ne permettent pas pragmatiquement de le définir mais qui le définissent plutôt en comparaison à d'autres réalités linguistiques. » (Ibid.).

Cette définition met en exergue l'intérêt de notre approche qui caractérise les registres de langue de manière comparative. Elle permettrait, par exemple, d'extraire les motifs linguistiques caractéristiques du registre courant renvoyant au français standard par rapport aux registres familier et soutenu. Nous voyons ici, comment notre approche pourrait être

utilisée par des travaux en sociolinguistique ou linguistique. La section suivante dresse justement un panorama des travaux étudiant la variation linguistique.

2.2 Entre sociolinguistique et linguistique

L'étude de la variation linguistique relève de la sociolinguistique et linguistique. De nombreux travaux sociolinguistiques cherchent à comprendre *pourquoi* les formes linguistiques varient et quelles variables socio-économiques ou culturelles sont éventuellement liées à ces variations. Tandis que les études linguistiques cherchent à décrire *comment* les formes varient et à exposer les moyens linguistiques mobilisés pour les produire. Selon le domaine de l'étude et son aire linguistique (francophone ou bien anglophone), la variation linguistique est appelée « niveau de langue », « registre » ou bien « style ».

- Le terme « niveau de langue » n'est employé que dans la littérature francophone et renvoie à des variétés linguistiques associées à une classe sociale (Stourdzé et Collet-Hassan 1969).
- Le terme « registre » :
  - dans la littérature francophone, renvoie aux variétés linguistiques mais sépare variations et classes sociales ;
  - dans la littérature anglophone, renvoie à des genres de textes et s'est imposé notamment grâce aux travaux de (D. Biber et Finegan 1990 ; D. Biber 1991 ; D. Biber 1994 ; etc).
- Le terme « style » :
  - dans la littérature francophone, renvoie aux ressources linguistiques mobilisées par un locuteur particulier et se place dans le domaine de la « stylistique » introduite par (Bally 1909) ;
  - dans la littérature anglophone, renvoie à des variations linguistiques comme définies dans (Labov 1988) ; dès lors, leur analyse est sociolinguistique.

À l'instar de ces définitions, nous parcourons dans cette section des études sociolinguistiques puis linguistiques, en différenciant les travaux anglophones et francophones.

2.2.1 En sociolinguistique

Cette section présente dans un premier lieu les travaux anglophones, puis les travaux francophones issus de la sociolinguistique.

2.2.1.1 Travaux sociolinguistiques anglophones

Les travaux fondateurs du sociolinguiste américain (Labov 1966) mettent en avant le lien entre la prononciation du /r/ et la classe sociale du locuteur de la ville de New York. Pour ce faire, il choisit 4 grands magasins qui visent 4 clientèles de classes sociales différentes. Dans tous ces magasins, le rayon de chaussures est au 4ème étage. Ainsi, lorsqu'il demande aux clients où se trouve l'étage pour les chaussures il note la prononciation du /r/ lorsqu'ils lui répondent « *the fourth floor* ». Une autre étude conduite en 1972 présentée dans (Labov 1972) sur le parler vernaculaire africano-américain met en lumière le fonctionnement linguistique de cette variété linguistique. Nous pouvons notamment citer le lien que cette étude met au jour entre les structures linguistiques choisies par le locuteur et la place que ce dernier occupe dans sa communauté. Citons également les travaux présentés dans (Wolfram 1969) qui fut l'un des premiers à étudier la variété de l'anglais américain à Détroit Michigan en 1969. Il montre une corrélation importante entre des variables phonologiques et grammaticales avec le statut social des locuteurs. Lorsqu'en 1974 l'étude présentée dans (P. Trudgill et al. 1974) applique la méthodologie proposée par (Labov 1966) en Angleterre, il montre à travers plusieurs variables phonologiques que leur prononciation est aussi liée au statut social du locuteur. Ce lien établi entre variables extra-linguistiques et productions linguistiques se retrouve dans les travaux de (Milroy 1986). Toutefois, il cherche à établir un lien entre des variations linguistiques et des communautés de locuteurs.

Ces variations linguistiques sont désignées avec le terme de « style » qui a été introduit dans la littérature scientifique anglophone par Labov. Il appelle style le degré de « surveillance » du locuteur porté à son discours, selon la situation dans laquelle il est produit. La notion de surveillance a été présentée dans les travaux de (Labov 1972), elle peut être décrite comme la conscience linguistique du locuteur sur son propre discours au moment où il le produit. Le terme de style dans une acceptation labovienne est largement repris dans la littérature scientifique anglo-saxonne. Parmi les travaux utilisant ce terme, nous pouvons mentionner les travaux présentés dans (Hymes 2013) qui proposent une liste de facteurs pouvant infléchir le style des locuteurs. Citons également les travaux présentés dans (Irvine 1985) qui étudient les styles en Amérique du Nord selon les différentes classes sociales. Les recherches introduites par (Khalid et P. Srinivasan 2020) analysent le style de différentes communautés en ligne [6] à travers l'étude de 262 traits linguistiques. Citons également les travaux présentés dans (Eckert 2003) qui ne se

6. 9 communautés issues de 3 plateformes

concentrent pas sur des communautés de locuteurs mais plus spécifiquement sur le style d'un locuteur. Ils observent le rôle du style comme outil de construction de la personnalité des locuteurs. Proche de cet angle d'analyse, (Auer 2008) regroupe différents articles sur le lien entre styles et constructions de l'identité. Les travaux de (Kiesling 2009) quant à eux cherchent à comprendre la construction de l'identité du locuteur lors de prise de position, il veut voir comment le style permet aux locuteurs de créer des prises de position et comment ces styles sont associés aux positions. Cette dimension psychologique des styles peut être illustrée par l'étude présentée dans (Taylor et Thomas 2008). Elle examine lors de prises d'otage, la relation entre la correspondance des styles linguistiques (entre négociateurs et preneurs d'otages) et le résultat de la négociation.

2.2.1.2 Travaux sociolinguistiques francophones

Dans la littérature francophone, la notion de variation linguistique se retrouve dans celle de « niveau de langue ». La notion de niveau de langue est une notion francophone qui vient de deux champs de recherches différents : la stylistique et la didactique des langues. D'une part, c'est autour de questions posées par la traduction que l'École de la Bibliothèque de stylistique comparée menée par (Malblanc 1944) fait émerger la notion de niveau de la langue dans les années cinquante. D'autre part, c'est en didactique que la notion est utilisée dans les années soixante par le mouvement de renouvellement pédagogique de la langue maternelle. Cette origine historique peut en partie expliquer l'importance des travaux traitant des niveaux de langue en traduction. Ces travaux montrent la complexité sociolinguistique des niveaux de langue en illustrant la difficulté à traduire ces variations linguistiques qui découlent de deux aires socioculturelles différentes. Pour n'en évoquer que quelques uns, nous pouvons citer les travaux présentés dans (Gallardo 2005) qui tentent d'utiliser un critère de fréquence afin de créer des tables d'expressions figées italien/français. Au même titre que dans le cadre du travail mené par F. Gadet, son travail différencie la variation géographique dite diatopique, et sociale dite diastratique. Le terme choisi dans ses travaux pour désigner la variation diastratique est celui de niveau de langue. Nous pouvons également mentionner le travail sur la traduction phraséologique présenté dans (Xatara 2002). Il met particulièrement en avant la difficulté de traduire les expressions idiomatiques [7] en respectant les niveaux de langue d'une langue source à une

7. Elles sont définies par (Caillies 2009) comme suit : « Elles constituent des locutions, connues comme telles et pouvant être répertoriées dans des dictionnaires, dont la signification est supposée ne pas résulter de la composition des significations des mots qui les constituent. »

langue cible. Les travaux présentés dans (Chuquet et Paillard 1987) soulignent quant à eux la difficulté de certaines transpositions français/anglais que posent les niveaux de langue. Ils exposent la nécessité de « retourner en amont du procédé de traduction afin d'examiner les principaux facteurs qui déclenchent son emploi ». Ils renvoient à des facteurs liés au contexte d'énonciation et au locuteur afin de comprendre et donc de traduire au mieux cette variation linguistique. Lors d'une étude sur le niveau familier dans les romans québécois présentée dans (Demaizière 1989), les niveaux de langue sont montrés comme intrinsèquement liés au contexte d'énonciation. D'après cette étude, les niveaux de langue donnent des informations sur le locuteur telles que sa classe sociale, ses orientations politiques, la zone géographique où il évolue ainsi que son histoire. Ces nombreuses informations en font un objet d'étude riche dont elle en souligne l'importance.

Les niveaux de langue comme indicateurs de la variété sociale se retrouvent aussi dans l'étude présentée dans (Gadet 1996). Cette étude décrit les niveaux de langue comme une variation diastratique, c'est-à-dire qu'elle renvoie à la diversité sociale des locuteurs. Par exemple, la partition de l'espace linguistique tirée de l'étude présentée dans (C. Demanuelli et J. Demanuelli 1991) va du registre populaire au registre soutenu, et révèle une certaine corrélation entre variation linguistique et classe sociale avec l'usage du terme « populaire ». Le terme niveau de langue est utilisé pour désigner :

> « Un ensemble de marqueurs soit au même niveau d'analyse soit à différents niveaux, convergeant dans une même direction, [qui] contribuera à imprimer à un texte un certain niveau de langue (populaire / courant / familier / soutenu). » (C. Demanuelli et J. Demanuelli 1991)

Ce lien entre variations linguistiques et classes sociales est étendu au « milieu du locuteur » dans l'étude présentée dans (Chuquet 1990), où les niveaux de langue sont définis comme :

> « différents types d'usage linguistique qui varient selon le milieu et la situation où se trouve le locuteur (milieu socioculturel, classe d'âge, milieu professionnel, etc.) » (Chuquet 1990)

L'étude de ce « milieu du locuteur » à travers sa perception des niveaux de langue se retrouve (Cajolet-Laganière et Martel 2011). Cette étude met en regard un travail linguistique qui liste des marqueurs linguistiques de différents niveaux de langue avec une étude sociolinguistique qui fait état de la perception des locuteurs de ces marqueurs. Leurs travaux révèlent la force de l'association faite par (Cajolet-Laganière et Martel 2011) entre niveaux de langue et variables sociales. Cette double étude est également intéressante pour la mise en exergue de l'hypothèse faite par les chercheurs sur

la capacité qu'auraient tous les locuteurs de percevoir les niveaux de langue. Cette faculté, qui serait innée, est exprimée dans l'étude présentée dans (Authier et Meunier 1972) où les auteurs parlent de « sentiment de niveaux de langue » : « une échelle relative, correspondant à une intuition nette qu'il y a des niveaux ou registres divers s'opposant dans la langue ». L'idée que les niveaux de langue soient une « capacité langagière » des locuteurs est reprise dans les travaux de (Fuchs 2021a) qui différencient niveau de langue et registre de langue. Ainsi, pour cette étude, le niveau de langue renvoie à différents moyens de s'exprimer pour les locuteurs ; tandis que le registre de langue renvoie au discours produit : « On appelle "registres de langue" les usages que font les locuteurs des différents "niveaux de langue disponibles", en fonction des situations de communication. » (Fuchs 2021a).

2.2.2 En linguistique

Les travaux sociolinguistiques précédemment présentés étudient la variation linguistique en considérant des critères tels que la classe sociale du locuteur, ou bien sa communauté culturelle. À l'inverse, sous l'angle de la linguistique, les travaux tendent à décrire de quelle manière les formes varient. Cette section en présente un panorama, à nouveau dans la littérature anglophone, puis francophone.

2.2.2.1 Travaux linguistiques anglophones

En linguistique, la variation linguistique est appelée *register* dans la littérature scientifique anglophone. Citons par exemple, les travaux présentés dans (Ferguson 1982) qui définissent les registres de langue comme une variation « dans laquelle la structure linguistique varie en fonction des occasions d'utilisation ». L'étude de (Ure 1982) précise les « occasions d'utilisation » et parle d'activités humaines dont les registres de langue seraient les produits : « chaque communauté linguistique a son propre système de registres[...] correspondant à l'éventail des activités que ses membres exercent normalement ».

En linguistique de corpus, le terme registres de langue s'est notamment imposé à travers les travaux de Douglas Biber (D. Biber 1991 ; D. Biber 1994 ; D. Biber 1995 ; S. Biber D. e. C. 2001 ; D. Biber 2012 ; D. Biber et Egbert 2018 ; D. Biber et Conrad 2019). Dans ses travaux, Biber définit un registre comme « une variété linguistique associée à une situation particulière d'utilisation (en comprenant des buts particuliers de communication) » (D. Biber et Conrad 2019). L'importance du contexte déjà évoquée

dans les définitions de (Ferguson 1982 ; Ure 1982) est retrouvée dans celle donnée par Biber. En effet, pour (D. Biber et Conrad 2019), le registre est une perspective d'analyse prise parmi trois possibles : le registre, le genre et le style. Elles se différencient entre elles autour de quatre axes principaux : (1) les textes pris en compte, (2) les types de traits linguistiques choisis, (3) leurs distributions et (4) leurs interprétations (table 2.1).

Axes principaux	**Registre**	**Genre**	**Style**
(1) Textes	extraits de textes	textes complets	extraits de textes
(2) Traits linguistiques	tous traits lexico-grammaticaux	expressions spécialisées, organisation rhétorique, mise en forme	tous traits lexico-grammaticaux
(3) Distribution des traits linguistiques	fréquents dans les textes de la variété	peu fréquents et apparaissent à un endroit particulier du texte	fréquents dans les textes de la variété
(4) Interprétation	les traits servent des fonctions communicatives importantes dans le registre	les traits sont conventionnellement associés au genre	les traits ne sont pas directement fonctionnels ; ils sont préférés sur le plan esthétique

Table 2.1 – Définition des caractéristiques des registres, des genres et des styles extraite de (D. Biber et Conrad 2019)

La table 2.1 montre que l'identification d'un registre repose sur des « descripteurs linguistiques qui ont toujours des rôles fonctionnels » (ibid.), c'est-à-dire qu'ils sont choisis selon le contexte et l'objectif de la communication. En cela, le registre s'oppose au style dans la mesure où, selon Biber, les descripteurs linguistiques du style ne sont pas fonctionnels car ils reflètent « plutôt des préférences esthétiques, associées à des auteurs particuliers ou des périodes historiques » (ibid.).

Afin d'étudier les registres, Biber observe quantitativement la variation de certains traits linguistiques sélectionnés *a priori* sur un très large corpus [8] selon différents axes :

8. Le corpus utilisé est composé des corpus suivants :
   - The TOEFL 2000 Spoken and Written Academic Language Corpus - T2K-SWAL
   - The Longman Spoken and Written English Corpus - LSWE
   - A Representative Corpus of Historical English Registers - ARCHER

oral/écrit, formel/informel, etc. Son objectif est d'identifier les cooccurrences de traits linguistiques selon ces axes. Par exemple, dans (D. Biber et Conrad 2019), une des analyses se concentre sur le comportement de traits linguistiques dans la presse écrite et les écrits académiques. Il est observé une présence plus importante du phénomène de nominalisation dans les écrits académiques et une présence plus importante des adjectifs attributs dans les écrits journalistiques.

La méthodologie a été appliquée sur d'autres types de corpus. Citons par exemple, les travaux menés sur des textes issus du Web (D. Biber et Egbert 2018). Dans le cadre de ces travaux, les auteurs présentent un large corpus [9] qui compte 50 millions de mots. Ils mettent en avant le fait qu'il est catégorisé selon différentes classes de textes telles que des interviews, des sous-titres de films, des textes dits formels, des billets de blogs, etc. Ils appellent ces classes des registres de textes. Cet usage du terme registre montre la différence de sens avec ce que nous appelons registres de langue. Ainsi, le terme registre dans la littérature anglo-saxonne semble désigner des genres de discours dans le sens de (Maingueneau 2012) qui les définit comme suit :

> « Des étiquettes comme "magazine", "vaudeville", "entretien d'embauche", "talk-show", etc., désignent ce qu'on entend habituellement par genres de discours, c'est-à-dire des dispositifs de communication qui ne peuvent apparaître que si certaines conditions socio-historiques sont réunies. » (Maingueneau 2012)

Dans notre démarche, nous caractérisons un phénomène linguistique qui se rapproche plus de la notion de style que celui de registre, selon les caractéristiques que Biber propose table 2.1. En effet, les traits linguistiques saillants dans les registres de langue ne sont pas directement fonctionnels, mais relèvent bien des choix esthétiques des locuteurs qui préfèrent (consciemment et inconsciemment) tels ou tels traits linguistiques.

2.2.2.2 Travaux linguistiques francophones

Dans la littérature francophone, les études linguistiques de (Gadet 1997) montrent que les variations linguistiques touchent divers niveaux d'analyse de la langue tels que le niveau phonologique (par exemple l'élision du *u* : "t'es mort" au lieu de "tu es mort"), le niveau morphologique (des terminaisons de mot non standards : "politicard"), le niveau lexical (des emprunts aux langues étrangères : "je suis dead") et le niveau syntaxique (la non inversion sujet/verbe dans une phrase interrogative : "Tu vas bien ?").

9. Corpus of Online Registers of English - CORE

Afin d'appréhender ce vaste phénomène que sont les variations linguistiques, nous avons considéré les travaux de (Gadet 1996) comme point de départ. De fait, les découpages de l'espace linguistique que Gadet opère sont largement repris dans la littérature, pour n'en citer que quelques uns : (Zribi-Hertz 2011 ; Pierozak 2003 ; Quillard 2000 ; Tyne 2012 ; Dewaele 2001 ; Pettit 2005). Plus précisément, elle partitionne l'espace linguistique en variétés de langue selon deux typologies (Gadet 1996) : selon les usagers, selon l'usage.

1. Selon les usagers :
   (a) Diachronique : la diversité dans le temps,
   ex : roman du XV^{e}siècle vs. roman du XX^{e}siècle
   (b) Diatopique : la diversité dans l'espace géographique,
   ex : *pain au chocolat* vs. *chocolatine*
   (c) Diastratique : la diversité dans la société,
   ex : *cheum* vs. *laid*
2. Selon l'usage :
   (a) Diaphasique : diversité selon le contexte d'énonciation,
   ex : dicussion entre amis vs. échange professionnel
   (b) Diamésique : diversité selon le support de l'énonciation.
   ex : conversation orale vs. échange d'emails

Notre travail porte sur un phénomène de l'ordre de la variation diastratique et diaphasique, mais il ne prend en compte aucune variable sociolinguistique comme le statut social du locuteur : il se concentre uniquement sur l'objet qu'est le texte produit. En cela, nous nous éloignons de la majorité des travaux qui se penchent sur la variation diastratique.

Dans la littérature scientifique francophone, l'étude de la variation linguistique se focalise surtout sur les formes linguistiques mobilisées et délaisse les éléments non linguistiques qui concernent le locuteur. Prenons par exemple les travaux présentés dans (Frei 1971) qui proposent une « Grammaire des fautes ». Après avoir constitué un corpus de textes écrits qui couvrent toute la France et illustrent « la langue courante », il en tire une typologie des fautes les plus fréquentes. La faute est vue par Frei comme symptomatique d'un « déficit du français ». Le locuteur opère une sorte de régularisation spontanée des irrégularités arbitraires de la langue normée. Frei présente les fautes comme venant palier des « besoins linguistiques » du locuteur. Il propose une typologie de cinq besoins :

1. l'assimilation, c'est-à-dire le besoin d'associer des signes à une signification ;

2. la différenciation, c'est-à-dire le besoin de clarté des propos ;
3. l'économie, c'est-à-dire le besoin d'une compréhension facile et rapide ;
4. l'invariabilité, c'est-à-dire le besoin de minimiser l'effort de mémoire ;
5. l'expressivité, c'est-à-dire le besoin de rendre son propos frappant.

Dès lors, selon Frei, le locuteur fait varier son énoncé selon ses besoins linguistiques aux dépens du respect de la norme grammaticale. Dans les études francophones, nous pouvons également citer les travaux présentés dans (Blanche-Benveniste et al. 1990) sur la variation syntaxique de la langue parlée. Son travail relève non seulement les formes régulières dans les variations produites, mais surtout propose de nouveaux outils afin de pouvoir décrire la variation de la langue parlée. En effet, les outils proposés jusqu'alors étaient conçus pour de l'écrit et devenaient inopérants pour de l'oral. Ainsi, (Blanche-Benveniste et al. 1990) propose une description de la variation syntaxique à l'oral, ainsi qu'un nouveau cadre théorique avec des outils opérationnels pour la langue parlée. Citons aussi l'étude de la variation considérée comme un trait de style, c'est-à-dire propre à un locuteur particulier. Le traité présenté dans (Bally 1909) prône une stylistique linguistique dont l'objet d'étude n'est pas le discours produit par le locuteur mais « ses ressources stylistiques » qu'il a à sa disposition. « La stylistique n'a pour objet ni le phénomène général du style à travers les langues ("tâche chimérique"), ni le style d'un écrivain particulier : elle doit étudier la langue parlée et ses ressources stylistiques. » (Calvet 2021).

Enfin, citons notre étude préliminaire menée en 2018 qui a validé une liste de descripteurs linguistiques pour la caractérisation d'un texte selon son registre de langue (familier, courant, soutenu). À partir d'un état de l'art mené sur la notion de registre dans la littérature linguistique et sociolinguistique, (Mekki, Battistelli, Lecorvé et al. 2018) établit une liste de 72 descripteurs [10], dont 30 ont été validés sur un corpus de textes français de registres distincts. Le corpus est composé d'écrits français [11] : *Albertine disparue* de Proust pour le registre soutenu, des archives de *L'Humanité* et *Le Monde* pour le registre courant et *Kiffe Kiffe Demain* de Guene, *L'Assommoir* de Zola et *Voyage au bout de la nuit* de Céline pour le registre familier. Considérant ces trois corpus, chacun spécifique à un registre, un descripteur est valide pour un registre donné si, parmi les différents corpus, la valeur du descripteur est significativement supérieure pour le corpus dédié à ce registre, à celle des autres. La table 2.2 présente les résultats de cette étude

10. La liste complète est donnée par l'annexe II.
11. Ces données ont été tokenisées, puis étiquetées en parties du discours avec Treetagger.

ID	Descripteur	Source	F	C	S
Niveau lexical (5)					
1	Element ponctuant	(Gadet 2003)	+		
2	Onomatopés	(Ilmola 2012)	+		
3	"là" ponctuants	(Gadet 1997)	+		
4	Termes à redoublement ("tonton", "dodo")	(Gadet 1997)	+		
5	Planificateurs du discours	(Branca-Rosoff 1999)			+
Niveau morphosyntaxique (13)					
6	Contraction de "cela" en "ça"	(Gadet 1997)	+		−
7	Négation sans "ne"	(Bilger et Cappeau 2004)	+		−
8	Sujet "on" transposé en "nous"	(Bilger et Cappeau 2004)	−		+
9	Terminaison en "-asse"	(Ilmola 2012)	+		
10	Terminaison en "-ouze"	(Ilmola 2012)	+		
11	Terminaison en "-o"	(Ilmola 2012)	+		
12	Verbe "être" au singulier devant un syntagme nominal singulier	(Bilger et Cappeau 2004)	+		
13	"ça" + verbe		+		
14	Dérivation d'un adverbe d'un nom ou adjectif ("vachement")	(Ilmola 2012)	+		
15	Verbe du premier groupe	(Gadet 1997)	+		
16	Emploi du passé simple		−		+
17	Emploi du passé composé			+	
18	Emploi du présent de l'indicatif			+	
Niveau syntaxique (10)					
19	Emploi fautif de relation en "que"	(Gadet 2003)	+		
20	Interrogative sans inversion sujet/verbe	(Gadet 2003)	+		
21	Interrogation en "est-ce que"	(Ilmola 2012)		+	
22	Maintien de "des" devant un adjectif au lieu de "de"	(Kalmbach 2012)	+		
23	Rajout de "à lui/elle" après un pronom personnel "son/sa"	(Gadet 2003)	+		
24	Emploi de pronoms relatifs	(Gadet 2003)			+
25	Adverbe + parataxe ("vraiment bien")		+		
26	Inversion "en" et COI à l'impératif ("donne m'en")	(Kalmbach 2012)	+		
27	"c'est... qui" ("c'est lui qui a fait")		+		
28	Effacement du pronom "il" impersonnel	(Favart 2010)	+		
Niveau phonétique (2)					
29	Elision de "e"	(Favart 2010)	+		
30	Elision du "i" du pronom "qui" devant une voyelle	(Ilmola 2012)	+		

Table 2.2 – Table des descripteurs validés positivement (+) et négativement (−) pour les registres familier (F), courant (C) et soutenu (S) extraite de (Mekki, Battistelli, Lecorvé et al. 2018)

préliminaire : il est précisé, pour chaque descripteur [12], s'il a été validé comme positivement (+) ou négativement (−) discriminant. Un descripteur positivement discriminant est un descripteur dont la forte fréquence est caractéristique d'un registre, tandis que pour un descripteur négativement discriminant c'est sa faible fréquence qui est caractéristique. Ces résultats confirment la possibilité de caractériser les registres de langue avec des descripteurs linguistiques à divers niveaux d'analyse de la langue, et valident une liste de descripteurs issus de la littérature scientifique. Les principales limites de ce travail préliminaire sont l'absence de formalisation du critère de significativité permettant de valider un descripteur comme caractéristique d'un registre et l'identification manuelle de ces descripteurs. Partant de ces premières validations, nous proposons une méthodologie pour la caractérisation des registres de langue qui formalise ce critère de significativité en utilisant la notion d'émergence. L'émergence compare les fréquences d'un même descripteur dans les textes d'un registre cible et dans les textes d'un registre source. Un descripteur dit émergent est un descripteur plus fréquent dans le registre cible que dans le registre source.

2.3 Partitionnement de l'espace linguistique

Pour caractériser les registres de langue de manière comparative, l'espace linguistique doit être partitionné en différentes parties. Comme précédemment évoqué, nous avons choisi de considérer la tripartition suivante : les registres familier, courant et soutenu. Nous admettons qu'il existe un continuum entre ces trois registres. Cependant, le découpage en valeurs discrètes permet un traitement automatique plus aisé. Avant de motiver notre choix de partitionner l'espace linguistique en trois registres, nous détaillons comment cet espace est représenté.

2.3.1 Représentation de l'espace linguistique : entre modalités orale et écrite

(Koch et Oesterreicher 2001) représente l'espace linguistique comme un continuum communicatif dont les deux extrêmes sont la modalité orale et la modalité écrite. La modalité orale est caractérisée par l'immédiateté communicative, tandis qu'à l'inverse la modalité écrite est caractérisée par la distance communicative. Cette représentation de

12. La référence bibliographique indique d'où il est tiré, l'absence de référence indique que le descripteur est issu de notre analyse linguistique du corpus.

l'espace linguistique pour l'analyse des registres de langue indexe leur perception à une maîtrise plus ou moins bonne de l'écrit, puisque la norme grammaticale est donnée comme repère zéro pour évaluer les variations linguistiques. Dès lors, tout ce qui est de l'ordre de l'oral ne peut être ni courant ni soutenu, et inversement tout ce qui est de l'ordre de l'écrit ou bien littéraire est considéré comme courant ou soutenu. Cette indexation des registres par rapport à une maîtrise plus ou moins forte de l'écrit vient de la tradition scolaire qui enseigne la maîtrise de la langue à travers la modalité écrite et non orale. Si nous prenons l'exemple de la dictée, cet exercice emblématique de l'apprentissage de la langue n'existe que pour l'écrit, il n'existe pas d'exercice similaire pour l'oral.

Le titre des travaux de (Poisson 2012) « L'oral, l'écrit et les registres » illustre cette corrélation établie entre registres et modalités : ils adoptent une tripartition (les registres familier, courant et soutenu) et associent la modalité orale avec le registre familier et la modalité écrite avec les registres courant et soutenu. Nous pouvons également citer l'organisation de l'espace linguistique proposée par (Bourquin 1965) dont l'axe horizontal est un continuum entre la modalité écrite vers la modalité orale, et dont l'axe vertical est un continuum structuré par trois registres principaux : le registre guindé soigné, le registre neutre, et le registre familier relâché. Ce paradigme se retrouve également dans les travaux de (Stourdzé et Collet-Hassan 1969) qui proposent une partition multidimensionnelle avec deux axes de lecture. Cette partition met en opposition deux usages : un usage instinctif, un usage élaboré. À chacun de ces usages est associé une modalité, respectivement la modalité orale avec la langue populaire et la langue familière ; et la modalité écrite avec la langue soignée, la langue littéraire et la langue classique. L'idée que le rôle de l'éducation et de la maîtrise scolaire de la langue conditionneraient les locuteurs à user de tels ou tels registres est explicitement formulée dans (Stourdzé et Collet-Hassan 1969) :

> « Une langue populaire, parlée naturellement par certaines couches sociales, formées en gros par les Français qui n'ont pas fait d'études secondaires, constitue un instrument de communication dans lequel formes et constructions grammaticales en particulier ne semblent obéir à aucune norme : il suffit que l'interlocuteur paraisse avoir compris le message. »

Citons aussi l'étude présentée dans (Devolder 2007) dans laquelle « le registre soutenu est celui de l'écrit ou de la distinction sociale, le registre courant celui des échanges quotidiens, le registre familier, celui de la communication entre pairs ». Ce lien entre niveaux d'éducation/classes sociales et registres de langue se retrouve dans plusieurs travaux, parmi eux nous pouvons citer (Damourette et Pichon 1930) qui parlent par exemple

d'une opposition entre une « parlure bourgeoise » et une « parlure vulgaire » ; ou bien chez (Bernstein 1977) qui envisage la langue comme un code plus ou moins élaboré où l'on distingue un « code restreint » et un « code élaboré » avec une connotation négative pour le premier par rapport au second.

Deux points se dégagent de ces représentations de l'espace linguistique : il est vu comme un continuum allant de la modalité orale à la modalité écrite, les registres de langue y sont situés selon une maîtrise plus ou moins bonne de l'écrit dépendante d'un niveau social et scolaire. Nous contestons ces deux points. Tout d'abord, les critères caractérisant les modalités orale et écrite ne suffisent plus à les distinguer avec l'émergence de nouveaux moyens de communication. En effet, la modalité écrite ne peut plus être caractérisée par la distance communicative, dès lors qu'il est possible d'échanger par écrit dans un chat lors d'une visioconférence. À l'inverse, la modalité orale ne peut plus être caractérisée par l'immédiateté communicative, puisqu'il est possible d'envoyer des notes vocales sur WhatsApp pouvant être réécoutées et modifiées avant l'envoi par l'émetteur, et écoutées plus tard par le récepteur. Ensuite, nous ne pensons pas la classe socioculturelle comme variable prédominante dans l'usage des registres. Par exemple, un locuteur d'un milieu social aisé ayant suivi un cursus universitaire peut utiliser le registre familier dans certaines situations, et le registre soutenu dans d'autres. Ces exemples mettent en avant le besoin de séparer la modalité orale de la notion d'immédiateté, et la modalité écrite de la notion de distance ; et la nécessité de distinguer les classes socioculturelles des registres de langue.

C'est pourquoi nous proposons une représentation de l'espace linguistique structurée par deux axes sans prendre en compte la notion de classe socioculturelle. La figure 2.2 montre l'espace organisé par un axe vertical allant de la modalité écrite à la modalité orale, et un axe horizontal allant de situations de communication diachroniques [13] à synchroniques [14]. Elle montre comment cette représentation permet de prendre en compte des types de situations de communication nouvelles en distinguant la modalité écrite d'une situation diachronique, et la modalité orale d'une situation synchronique. Pour caractériser les registres de langue, nous avons choisi de partitionner cet espace linguistique en une tripartition : les registres familier, courant et soutenu. Cependant, selon l'angle d'étude privilégié, on observe dans la littérature linguistique diverses manières de partitionner l'espace linguistique en différents registres. La section suivante dresse un panorama des partitions possibles.

13. Situation de communication où les locuteurs ne sont pas dans le même cadre spatio-temporel.
14. Situation de communication où les locuteurs sont dans le même cadre spatio-temporel.

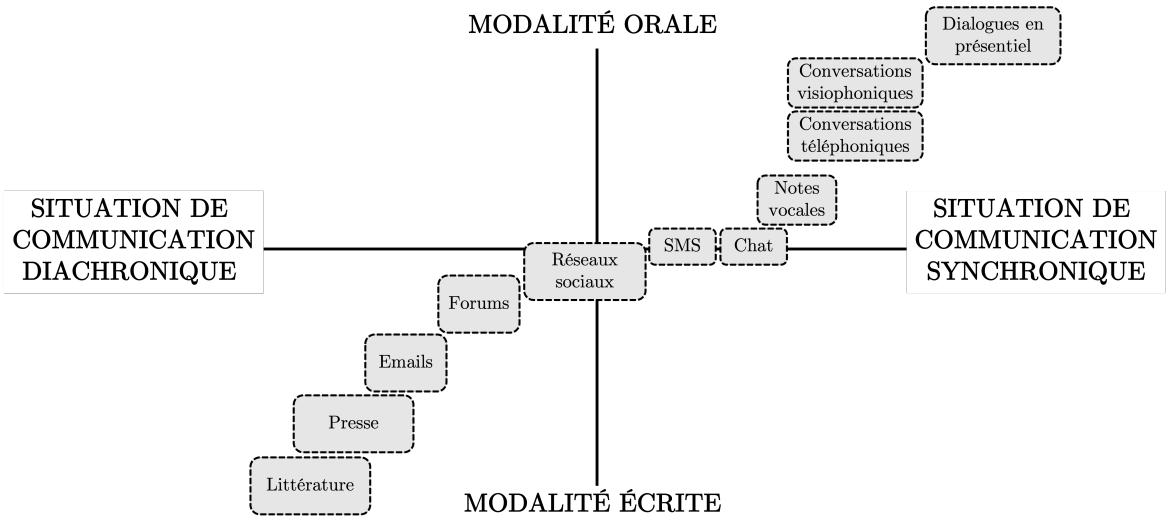

FIGURE 2.2 – Notre représentation de l'espace selon deux axes : un axe vertical allant de la modalité écrite à la modalité orale, un axe horizontal allant de situations de communication diachroniques à synchroniques. Des exemples de situations de communication sont répartis dans cet espace.

2.3.2 Partitionnement de l'espace linguistique

Dans nos travaux, nous avons choisi une tripartition de l'espace : les registres familier, courant et soutenu. Elle a été choisie car prédominante dans les travaux en sociolinguistique et linguistique, ainsi que dans les manuels scolaires. Nous montrons dans cette section des exemples de tripartitions dans les domaines de la sociolinguistique et linguistique, puis dans le milieu scolaire.

En sociolinguistique et linguistique Suivant l'objectif de l'étude, l'espace linguistique est segmenté de différentes façons. Par exemple, (ILMOLA 2012) propose de distinguer les registres familier, populaire et vulgaire dans des journaux satiriques. Les travaux de (BORZEIX et FRAENKEL 2005), quant à eux, catégorisent différentes situations de communication au travail en opposant, par exemple, la communication fonctionnelle à la communication relationnelle. Cependant, nous pouvons constater la prédominance de la tripartition (identifiant les registres familier, courant et soutenu) en linguistique dans la littérature scientifique. Cette tripartition que nous adoptons trouve ses origines chez les rhéteurs grecs qui considéraient trois catégories : le bas, le médiocre, le sublime (DANGEL 2007). De nombreux travaux la reprennent, citons par exemple (MONNERET et POLI 2020), qui distingue trois registres de langue :

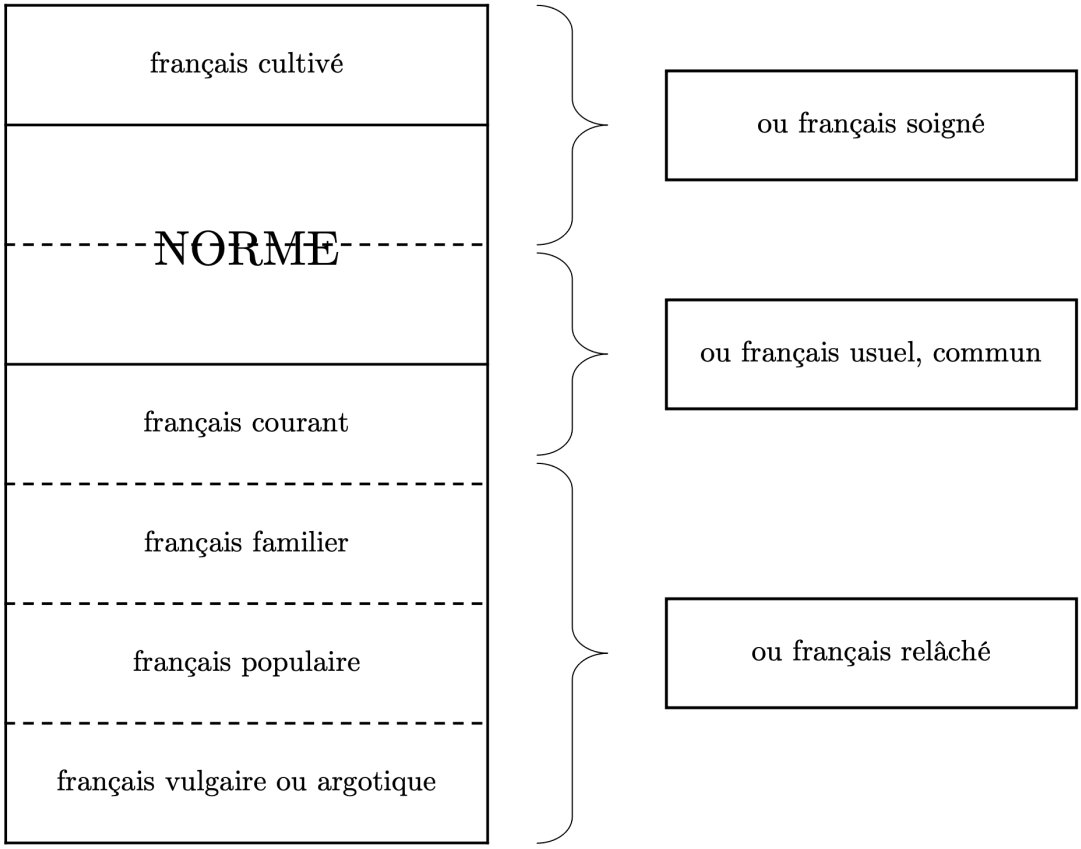

FIGURE 2.3 – Partition de l'espace linguistique extraite de (Bodo MULLER 1985)

> « On distingue usuellement trois registres de langue : familier (canasson), courant (cheval) et soutenu (destrier). D'autres registres sont possibles, bien sûr, dans un continuum qui va du très vulgaire au très soutenu. » (MONNERET et POLI 2020)

Cette tripartition se retrouve aussi dans les travaux présentés dans (Bodo MULLER 1985) avec le français relâché, le français usuel/commun et le français soigné. La figure 2.3 montre qu'à cette tripartition il ajoute une sous partition où une supernorme est formée par le français cultivé par rapport à la norme qu'il appelle le « niveau zéro » représentant le français courant. Toute production non standard ou dite relâchée est inférieure au « niveau zéro ». Ici, les pointillés indiquent un continuum possible seulement entre certains registres et montre de ce fait que le respect ou non de la norme linguistique stoppe le continuum. Ce choix met en exergue l'importance accordée à la norme linguistique dans la perception des registres de langue.

En milieu scolaire De même, cette tripartition de l'espace linguistique en registres familier, courant et soutenu, est prédominante dans les contenus scolaires. Parmi les contenus pédagogiques accessibles en ligne, nous trouvons cette tripartition dès la primaire dans diverses fiches de révision [15], dans des vidéos pédagogiques [16], ou encore dans des fiches d'exercices [17]. La tripartition est également retrouvée pour des niveaux scolaires plus avancées à travers l'étude des figures de style [18], ou encore dans les outils à maîtriser pour l'analyse et la rédaction de textes [19]. Ces quelques exemples montrent que l'espace linguistique est partitionné en trois registres principaux durant tous les niveaux scolaires. Dès lors, cette tripartition nous semble une segmentation robuste sur laquelle nous appuyer pour caractériser les registres de langue.

2.4 Conclusion

Pour caractériser les registres de langue renvoyant aux variations linguistiques, il faut s'appuyer sur une définition de ces derniers. Cette section a montré pourquoi la littérature scientifique ne s'accorde pas sur une définition stable. Comprendre quels sont les points de désaccord dans la littérature nous a permis de circonscrire notre objet d'étude, que sont les registres de langue, en positionnant notre travail.

Le premier point se trouve dans le choix de la norme linguistique considérée pour évaluer les variations linguistiques. Nous avons montré plusieurs types de normes allant d'une posture descriptive formalisant des règles issues de l'usage, à une posture prescriptive imposant des règles grammaticales. En soulignant les limites de ces types de normes considérées séparément, nous avons proposé une norme reposant à la fois sur une norme d'usage et une norme grammaticale. Notre définition de la norme évite, notamment, d'orienter la perception des registres de langue uniquement sur la notion de faute grammaticale.

La seconde difficulté réside dans la profusion des types de variables croisées avec les registres de langue lors de leurs analyses. Les études sociolinguistiques considérant des informations sur le locuteur parlent de « style » lorsqu'elles sont anglophones, ou de « niveau de langue » lorsqu'elles sont francophones. En revanche, les études linguistiques se

15. Lien pour accéder au contenu dédié aux CM1.
16. Lien pour accéder au contenu dédié au cycle 3.
17. Lien pour accéder au contenu dédié aux CE1.
18. Lien pour accéder au contenu dédié au niveau terminal.
19. Lien pour accéder à la fiche de révision pour le Bac.

focalisant sur des variables linguistiques parlent de « registre » ou de « style » lorsqu'elles sont, respectivement, anglophones et francophones. En outre, le panorama de ces études a révélé qu'aucune n'a travaillé sur les registres de manière comparative, si nous faisons exception de notre propre travail préliminaire. Face à cette profusion terminologique, nous avons choisi le terme de « registre de langue » principalement pour éviter toute notion hiérarchisante. Notre étude est linguistique, car elle considère divers niveaux d'analyse de la langue pour caractériser les registres de langue de manière comparative. En cela, notre approche est originale dans le paysage de la littérature linguistique francophone.

Enfin, le troisième obstacle à un consensus sur la définition des registres de langue est la représentation de l'espace linguistique considéré. Elle se base sur l'idée d'un continuum communicatif allant de la modalité écrite à la modalité orale. Nous avons exposé pourquoi cette représentation est problématique, à la fois pour l'analyse des registres s'indexant alors sur la maîtrise de l'écrit ; et pour la prise en compte de nouveaux moyens de communication brouillant la frontière entre ces deux modalités. Pour répondre à ces limites, nous avons introduit une représentation de l'espace linguistique organisée selon deux axes. Elle sépare les modalités orales et écrites, des notions de synchronie et de diachronie. En dernier lieu, nous avons justifié notre choix de partitionner cet espace en trois registres principaux (familier, courant et soutenu) en montrant la prédominance de cette tripartition dans la littérature scientifique, ainsi que dans les manuels scolaires.

Ainsi, cette section a borné notre objet d'étude en positionnant nos travaux par rapport à la littérature scientifique. Délimiter le cadre théorique dans lequel nous situons les registres de langue a été une étape primordiale dans notre travail. La prochaine étape dont l'objectif final est de caractériser ces registres par des descripteurs linguistiques est maintenant la collecte de textes. C'est précisément l'objectif du prochain chapitre.

Chapitre 3

Constitution de corpus

Sommaire

Ce chapitre développe les travaux menés pour constituer un ensemble de textes représentant les registres de langue. La construction de cet ensemble de textes s'inscrit dans l'approche utilisée en sciences du langage, ou bien en linguistique de corpus. Cette approche envisage sa construction comme une réponse à un objet de recherche, il y est désigné sous le terme de « corpus » et défini comme suit :

> « Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : de manière théorique réflexive en tenant compte des discours et des genres, et de manière pratique en vue d'une gamme d'applications. Tout corpus suppose en effet une préconception des applications, fussent-elles simplement documentaires, en vue desquelles il est rassemblé : elle détermine le choix des textes, mais aussi leur mode de « nettoyage », leur codage, leur étiquetage ; enfin, la structuration même du corpus. » (Rastier 2005)

Dès lors, notre objectif est de construire un corpus de textes devant répondre à la question suivante : comment représenter les registres de langue français ? Pour respecter notre positionnement, établi au chapitre 2, nous voulons représenter les registres en les séparant des modalités orale et écrite ainsi que des genres de textes. Nous souhaitons également obtenir un corpus volumineux pour représenter un large spectre des usages réels.

Dans la première section de ce chapitre, nous donnons une vue d'ensemble des travaux

dans la littérature scientifique se rapprochant de notre démarche. Cela nous donne l'occasion d'en souligner les limites notamment liées au biais associant registres et genres de textes ; et de motiver notre démarche de création d'un corpus de textes respectant notre positionnement sur la représentation des registres de langue français.

Constituer un corpus de textes représentant des registres de langue revient à attribuer un registre de langue à chaque texte. Pour cela, nous avons utilisé un modèle d'apprentissage automatique. Il permet de généraliser à un corpus entier l'annotation manuelle d'un sous-corpus. Comme le processus d'annotation manuelle est chronophage, le sous-corpus annoté manuellement représente une petite proportion du corpus total. Afin de contrebalancer ce déséquilibre, nous avons employé une technique d'apprentissage semi-supervisée augmentant itérativement les données d'entraînement. À chaque itération, des textes étiquetés automatiquement jugés fiables sont filtrés pour être ajoutés aux données d'entraînement. Notre technique d'apprentissage semi-supervisée adaptant celui de (Lecorvé et al. 2018) est présentée section 3.2.

La section 3.3 introduit des travaux préliminaires employant notre modèle d'apprentissage. Durant cette phase exploratoire, nous avons testé la possibilité de neutraliser les biais liés aux genres de textes, ainsi que d'avoir un étiquetage des registres plus réaliste que celle de (Lecorvé et al. 2018). La question des biais, liée aux genres de textes, a été traitée lors de la constitution du sous-corpus d'entraînement en considérant le même nombre de genres de textes pour chaque registre. Quant à l'étiquetage en registres de langue, nous le voulions plus réaliste en illustrant l'hétérogénéité de la langue : plusieurs registres peuvent être étiquetés dans un même texte. Les résultats obtenus ont validé la possibilité d'étiqueter un même texte avec plusieurs registres de langue. Cependant, ils ont également révélé certaines limites. La première est la persistance des biais liés aux genres de textes. La deuxième est la discordance entre l'unité textuelle étiquetée et celle considérée lors de l'extraction de motifs séquentiels émergents. Comme l'unité étiquetée est un texte de plusieurs phrases, si nous considérons une unité plus petite pour l'extraction de motifs, alors l'étiquetage pourrait être faussé. Par exemple, une phrase étiquetée comme appartenant au registre familier car provenant d'un texte étiqueté comme tel, pourrait en réalité appartenir au registre courant. Cette discordance dégraderait la qualité des motifs extraits.

Partant de ce constat, nous avons proposé la constitution d'un corpus de tweets appelé TREMoLo-Tweets. Pour supprimer tout biais associé aux genres de textes, nous avons adopté la stratégie inverse à celle mise en place dans nos travaux préliminaires. Un

seul genre de textes est utilisé : les tweets. Le choix des tweets repose, notamment, sur son format court permettant de faire correspondre l'unité étiquetée en registres avec celle considérée lors de l'extraction de motifs. Enfin, pour encadrer l'annotation manuelle, nous avons proposé un guide d'annotation introduisant un nouvel ensemble de descripteurs linguistiques pour analyser les CMO [1]. Une de nos contributions réside dans l'intégration d'éléments linguistiques traditionnellement écartés dans les travaux de TAL comme les hashtags, les URLs, ou encore les pictogrammes. Les expériences ont indiqué la qualité des prédictions. Les premières analyses du corpus ont, quant à elles, validé la pertinence de l'intégration de ces descripteurs pour l'analyse des CMO. Leurs répartitions dans le corpus montrent l'intégration à la norme grammaticale de certains d'entre eux comme les hashtags, ou bien les pictogrammes, qui ne sont plus réservés à un registre familier. Au contraire, ils sont utilisés dans un registre soutenu. De plus, nous avons observé l'émergence de descripteurs linguistiques pour les registres de langue. Nous avons sélectionné manuellement ces descripteurs lors de l'analyse linguistique d'une partie du corpus. Ces résultats valident la possibilité de caractériser les registres avec des traits choisis *a priori* et confortent notre démarche de les caractériser automatiquement sans *a priori* sur les descripteurs à découvrir. Elle assure également la possibilité d'utiliser TREMoLo-Tweets comme ressource textuelle pour caractériser automatiquement les registres de langue. Ces travaux, dont résulte le corpus TREMoLo-Tweets, terminent ce chapitre avec la section 3.4.

3.1 Contexte et motivations

Nous avons cherché dans la littérature scientifique si des corpus annotés en registres de langue, ou bien des classifieurs déjà entraînés pour la prédiction de registres, existaient déjà. L'état de l'art présenté dans (S. E. Argamon 2019) montre que les travaux ayant pour objet d'étude ce que nous appelons « registres de langue » ne recourent que de manière très marginale au terme « registre ». Ils utilisent préférentiellement celui de « style », de « genre » ou encore de « (degré de) formalité ». De fait, dans le contexte du TAL (au sens strict du terme), on ne trouve, à notre connaissance, aucune étude qui utilise le terme de « registres de langue ». Néanmoins, nous pouvons citer des travaux qui s'en rapprochent. Dans cette section, nous en dressons un panorama avec des tâches de prédiction de (degré de) formalité et les corpus produits. En mettant en avant les diffé-

1. CMO pour **C**ommunications **M**édiées par **O**rdinateur

rentes limites de ces travaux, nous montrons la nécessité de constituer un nouveau corpus pour la représentation des registres de langue français.

Travaux anglophones Pour l'anglais, (Peterson, Hohensee et Xia 2011) propose des techniques de classification de textes considérant quatre classes : très formel, peu formel, peu informel, très informel. Les auteurs prédisent un degré de formalité d'une phrase donnée. Le corpus utilisé est le corpus *EnronSent* de courriers électroniques composés de 96 107 emails (Styler 2011), dont 400 emails sont tirés aléatoirement afin de les annoter manuellement. Les auteurs n'ont pas donné de définition de la formalité aux annotateurs. Ils supposaient les registres connus de tous :

> « *Because formality is hard to define, we did not give annotators a concrete definition. Instead, we provided a few guidelines and asked annotators to follow the guidelines and their intuition.* » (Peterson, Hohensee et Xia 2011)

(Sheikha et Inkpen 2010) prédit également un degré de formalité en utilisant une régression à partir d'un corpus de textes catégorisés en deux classes : formel vs. informel. Au lieu d'annoter manuellement les données, ils ont tiré aléatoirement 1 000 textes de deux corpus qu'ils ont jugé soit formels soit informels. Plus précisément, pour les données dites formelles, 500 textes sont collectés à partir du :

- *Reuters Corpus*[2], qui contient des textes estimés formels tels que des textes journalistiques ;
- *Open American National Corpus : written texts*[3], qui contient des textes.

Pour les données dites informelles, 500 textes sont collectés à partir du :

- *Late Modern English Corpus*[4], qui contient des courriers personnels ;
- *Enron Email Corpus*[5], qui contient des emails ;
- *Open American National Corpus : spoken texts*[6], qui contient des retranscriptions d'oral.

L'utilisation de ces corpus pour obtenir des ensembles de textes étiquetés en registres selon leurs genres introduit le biais associant registres et genres de textes. Ce biais est également dans les travaux de (Pavlick et Tetreault 2016). Pour prédire un degré de formalité

2. `https://trec.nist.gov/data/reuters/reuters.html`
3. `https://www.anc.org/data/`
4. `https://perswww.kuleuven.be/~u0044428/clmet.html`
5. `https://wstyler.ucsd.edu/enronsent/`
6. `https://www.anc.org/data/`

d'une phrase, ils utilisent un corpus contenant quatre genres de textes différents : 4 977 phrases extraites de résultats du moteur de recherche Yahoo ![7], 1 701 phrases d'emails professionnels [8], 2 775 phrases de la presse écrite (Lahiri 2015), et 1 821 phrases de billets de blog (Ibid.). Ils ne définissent pas la notion de formalité et demandent aux annotateurs de suivre leurs intuitions. Les travaux de (Heylighen et Dewaele 1999) se différencient de par leur volonté de préciser la notion de formalité, mais introduisent un biais associant les modalités orale et écrite avec les registres de langue. Ils n'utilisent pas une technique de prédiction, mais appliquent une formule spécifique pour prédire un degré de formalité. La formule pondère positivement la fréquence de certains éléments grammaticaux tels que les noms communs, les adjectifs et les prépositions ; et pondère négativement celle des verbes, adverbes et interjections. Deux types de formalité sont pris en considération : l'une dite profonde (*"deep" formality*), l'autre dite de surface (*"surface" formality*). La première est définie comme l'évitement de l'ambiguïté en minimisant la dépendance au contexte et le flou des expressions, tandis que la seconde se caractérise par l'attention portée à la formalité dans le sens du respect des conventions du genre par exemple. Les expériences sont menées sur un corpus de productions orales et écrites d'étudiants dans trois situations différentes : une conversation informelle, un examen oral qui évalue leurs connaissances sur le langage, un essai écrit rédigé durant un partiel.

Certains travaux de prédiction de la formalité se concentrent sur les éléments lexicaux. Par exemple, les travaux de (Brooke, T. Wang et Hirst 2010) se basent sur la longueur des mots afin d'en prédire leur degré de formalité. Ils utilisent *The Brown Corpus* [9] (W. N. Francis et Kucera 1979) qu'ils divisent en trois classes principales : le corpus formel composé de productions écrites ; le corpus informel composé de productions orales ; le corpus dit mixte. Leurs données annotées se composent de 138 termes issus d'un dictionnaire d'argot en ligne [10], 105 termes issus d'une liste de connecteurs logiques et d'un lexique de sentiments dont les auteurs ne précisent pas la source. Tout comme pour les travaux précédents, la méthodologie présentée mélange les modalités orale et écrite en les associant respectivement aux registres informel et formel. Suite à leurs expériences, (Pavlick et Tetreault 2016) propose un classifieur que (Rao et Tetreault 2018) utilise pour constituer un corpus de phrases formelles et informelles alignées : le ***G**rammarly's **Y**ahoo **A**nswers **F**ormality **C**orpus* (GYAFC). Les phrases sont des paires

7. `https://fr.yahoo.com/`
8. `https://americanbridgepac.org/jeb-bushs-gubernatorial-email-archive/`
9. `http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM`
10. `http://onlineslangdictionary.com/`

de questions/réponses. Le corpus GYAFC compte au total 110 000 paires de phrases tirées du corpus *Yahoo Answers L6 corpus*[11]. La principale limite de ces travaux est la langue : ils traitent les registres de langue pour l'anglais et produisent un corpus de textes anglais avec un classifieur inadapté à la langue française. L'absence de guide d'annotation rend impossible toute reprise totale ou partielle de leurs travaux.

Travaux francophones Côté francophone à notre connaissance, seul le corpus web constitué par (Lecorvé et al. 2018) se rapproche de notre objet de recherche. Après avoir collecté automatiquement un ensemble de 800 000 textes pages web, à partir de requêtes rédigées selon deux lexiques spécialisés (le premier pour le registre familier, le second pour le soutenu), trois classes ont été considérées pour l'annotation : familier, courant, et soutenu. 435 textes ont été choisis manuellement et équitablement répartis entre les classes pour être annotés manuellement. Ce sous-corpus est appelé la graine. L'annotation manuelle a ensuite été généralisée à l'ensemble du corpus grâce à une approche qui procède par itérations en alternant l'apprentissage d'un classifieur et l'annotation par ce classifieur de nouveaux textes pour augmenter la graine. Le classifieur se base sur un ensemble de 46 traits linguistiques utilisés comme descripteurs d'apprentissage, issus d'une étude linguistique (Mekki, Battistelli, Lecorvé et al. 2018) sur les registres de langue. Ces traits linguistiques recouvrent divers niveaux d'abstraction de la langue (détails donnés table 3.1). La composition de la graine avec des genres de textes différents selon les registres de langue introduit le biais associant genres et registres. Ce biais se retrouve dans la majorité des approches travaillant sur les registres de langue.

Synthèse des travaux La figure 3.1 donne une synthèse des corpus existants pour la représentation des registres de langue anglais et français. Elle précise quels genres de textes ont été utilisés, si l'annotation manuelle a été encadrée par un guide d'annotation et quelles étiquettes ont été employées pour l'annotation en registres. Ces cinq propositions de corpus révèlent deux limites communes :

1. l'association de certains genres de textes avec certains registres, le registre familier est souvent associé à des genres tels que les SMS ou les billets de forum, tandis que les registres courant et soutenu sont associés à des genres littéraires ;
2. l'association de certaines modalités avec certains registres, le registre familier est associé à la modalité orale, tandis que les registres courant et soutenu sont associés

11. `https://webscope.sandbox.yahoo.com/catalog.php?datatype=l`

Lexique
- Onomatopées (« ah », « pff ». . .) : 125 éléments
- Termes du langage SMS (« slt », « lol », « tkt ». . .) : 540 éléments
- Anglicisme (lexique et syntaxe)
Phonétique
- Élision voyelle (« m'dame », « p'tit ». . .)
- Élision « r » (« vot' », « céleb' ». . .)
- Liaisons écrites « z » (« les zanimaux »)
Morphologie
- Répétitions de syllabes (« baba », « dodo ». . .)
- Répétitions de voyelles (« saluuuut »)
- Emploi de mots terminant en « -asse »
Morphosyntaxe
- Emplois des temps : impératif présent, indicatif futur, indicatif imparfait etc.
- Emploi des personnes : seconde pluriel (« vous ... »), seconde singulier (« tu ... »)
- Emploi de verbe du premier ou deuxième groupes
Syntaxe
- Redoublement de la possession (« son . . . à lui »)
- Structure « c'est ... qui »
- Emploi de la conjonction « et »

Table 3.1 – Liste d'exemples de descripteurs d'apprentissage issus de (Mekki, Battistelli, Lecorvé et al. 2018) utilisés par le classifieur dans (Lecorvé et al. 2018).

à la modalité écrite.

Or, notre positionnement sur la représentation des registres veut justement séparer les genres de textes et les modalités, des registres de langue. C'est pourquoi nous avons choisi de constituer notre propre corpus de textes représentant les registres en veillant à ne pas introduire ces biais. La section suivante présente notre démarche pour constituer le corpus.

3.2 Notre modèle d'apprentissage automatique

Constituer un corpus de textes représentant les registres de langue équivaut à attribuer pour un ensemble de textes un ou des registres à chacun d'entre d'eux. Comme nous avons voulu avoir une représentation proche des usages réels des registres de langue, nous avons souhaité créer un corpus volumineux. Dès lors, il aurait été difficile d'annoter manuellement tous les textes du corpus. Pour éviter ce travail chronophage, des techniques

ARTICLE	LANGUE	GENRES DE TEXTES	GUIDE D'ANNOTATION DISPONIBLE	ETIQUETTES POUR L'ANNOTATION
(Sheikha et Inkpen, 2010)	Anglais	• Lettres • Emails • Transcriptions de discussions orales...	Non	• Degré de formalité
(Peterson et al., 2011)	Anglais	• Emails	Non	• Très formel • Un peu formel • Un peu informel • Très informel
(Pavlick et Tetreault, 2016)	Anglais	• Emails • Presse écrite • Billets de blogs • Extraits de résultats de recherches Web	Non	• Degré de formalité
(Rao et Tetreault, 2018)	Anglais	• Questions/Réponses	Non	• Formel • Informel
(Lecorvé et al., 2018)	Français	• Œuvres littéraires • Billets de forums • Commentaires...	Non	• Familier • Courant • Soutenu

FIGURE 3.1 – Synthèse des corpus existants pour les registres de langue anglais et français

d'apprentissage peuvent être employées. Les modèles d'apprentissage supervisés réduisent le temps d'annotation manuelle en étiquetant automatiquement tout le corpus, après s'être entraînées sur un sous-corpus de textes annotés manuellement. D'autres types de modèles d'apprentissage sont possibles. Parmi eux, les techniques d'apprentissage semi-supervisées permettent d'utiliser un petit ensemble de données annotés manuellement. Le classifieur s'entraîne sur cet ensemble de textes annotés qu'il augmente itérativement de textes étiquetés automatiquement.

Dans cette section, nous introduisons la technique d'apprentissage semi-supervisée reprise des travaux de (LECORVÉ et al. 2018). Nous l'avons adaptée pour constituer notre corpus de textes représentant les registres de langue. Le choix de ce modèle d'apprentissage se base sur deux points. Premièrement, il compensait la faible quantité de textes annotés manuellement en y ajoutant itérativement des textes étiquetés automatiquement jugés fiables. Deuxièmement, sa robustesse pour la tâche a été validée par les résultats présentés dans (LECORVÉ et al. 2018). Nous donnons tout d'abord une vue d'ensemble de la technique d'apprentissage semi-supervisée, dont nous détaillons ensuite chaque étape en mettant en lumière leurs enjeux.

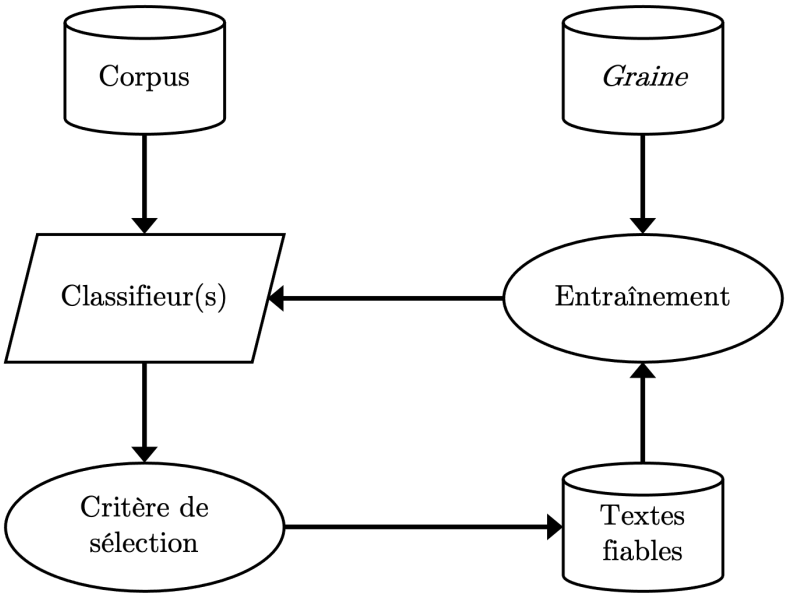

FIGURE 3.2 – technique d'apprentissage semi-supervisée repris de (LECORVÉ et al. 2018) utilisée dans nos travaux.

3.2.1 Approche générale du modèle

Notre technique d'apprentissage semi-supervisée (*self-training* en anglais) utilise un classifieur pour étiqueter les textes de notre corpus, c'est-à-dire leur attribuer un ou des registres de langue. Il s'entraîne à partir d'un jeu de données annotées manuellement qui est itérativement augmenté de données étiquetées automatiquement. Dans notre travail, ce jeu de données d'origine est appelé la « graine ». La figure 3.2 illustre toutes les briques de notre technique d'apprentissage semi-supervisée reprenant celle proposée par (LECORVÉ et al. 2018). Le classifieur étiquette les textes du corpus, après s'être entraîné à partir de la graine. Puis, nous filtrons les textes étiquetés pour sélectionner les textes dont l'étiquetage est jugé fiable selon un critère de sélection. Ces textes fiables sont ajoutés à la graine. Le classifieur initie un nouvel entraînement à partir de la graine augmentée et étiquette les textes restants. Le nombre d'itérations est fixé par l'utilisateur. Ce que nous appelons les « prédictions finales » correspond aux étiquetés lorsque tous les textes ont été étiquetés par le classifieur. Chaque étape de la figure 3.2 peut varier selon l'objectif des travaux l'utilisant. Dans la prochaine section, nous expliquons en quoi leurs variations peuvent avoir des conséquences linguistiques sur la représentation des registres de langue.

3.2.2 Enjeux et problématiques des étapes du modèle

Chaque étape de notre approche conditionne la manière dont les registres de langue sont représentés et étiquetés. Nous détaillons ci-dessous les enjeux et problématiques liés à chacune de ces étapes.

Représentation des registres La constitution du corpus détermine la représentation des registres en choisissant quels genres de textes et quelles modalités seront étiquetés en registres de langue. La graine, quant à elle, est un sous-corpus dont la constitution détermine la manière dont le classifieur apprend à distinguer les registres entre eux. Pour constituer la graine, nous devons sélectionner des textes pour représenter chaque registre de langue. Le classifieur considère les textes sélectionnés comme des exemples à suivre pour étiqueter les textes du corpus en registres de langue. Il faut représenter de manière équilibrée les registres dans la graine en ayant les mêmes quantités de textes pour chacun d'entre eux. Un registre sous-représenté dégraderait les prédictions du classifieur, car il n'aurait pas eu assez d'exemples lors de son entraînement pour distinguer ce registre. Enfin, pour ne pas introduire de biais dans la représentation des registres, il est nécessaire d'équilibrer les genres de textes les représentant. Par exemple, si le registre soutenu n'était représenté qu'avec des textes scientifiques, alors le classifieur n'apprendrait pas à distinguer le registre soutenu mais à reconnaître le genre de textes scientifiques. Dès lors, la constitution de la graine est le moyen d'illustrer notre positionnement théorique sur la représentation des registres de langue.

Choix du type de prédictions Plusieurs manières de prédire un ou des registres de langue d'un texte sont possibles. Celle proposée par (Lecorvé et al. 2018) attribue une seule étiquette par texte, cela veut dire qu'un seul registre est prédit par texte. Une autre manière de faire est d'attribuer plusieurs étiquettes à un même texte, afin de prédire plusieurs registres de langue pour ce dernier. Enfin, il est possible de prédire un degré d'appartenance à un registre en attribuant au texte une valeur estimant ce degré d'appartenance. Selon le type de prédictions souhaité, les classifieurs utilisés varient. D'un point de vue linguistique, choisir le type de prédiction revient à se positionner sur un usage homogène ou hétérogène des registres de la langue, ou encore sur la possibilité de mesurer la présence d'un registre.

Critère de sélection des textes fiables Choisir un critère de sélection pour filtrer les textes fiables revient à dire ce qu'est une « prédiction fiable » pour nous, c'est-à-dire une prédiction dans laquelle nous avons confiance. La méthode de sélection du texte fiable oriente la manière d'étiqueter du classifieur, puisqu'un texte qualifié de fiable est ajouté à la base d'entraînement. Selon les types de prédictions, plusieurs méthodes évaluant la fiabilité d'une prédiction sont applicables. Avec un classifieur multi-classes dont la somme des prédictions est égale à 1 nous pouvons, par exemple, vérifier la confiance du classifieur en regardant la valeur des probabilités d'appartenance à un registre. Plus les valeurs sont hautes, plus le classifieur est confiant. Cette méthode de filtrage oriente le classifieur à étiqueter les textes avec un registre dominant. Avec plusieurs classifieurs binaires, représentant chacun un registre, nous pouvons vérifier la cohérence des prédictions en regardant la somme des probabilités prédites. Cela aura pour effet de conserver un étiquetage plus nuancé. Plus la somme est supérieure à 100%, plus les prédictions sont incohérentes : un texte ne peut pas être à la fois 100% du registre courant, 100% du registre familier et 100% du registre soutenu.

Différentes expériences ont été menées en faisant varier la manière de représenter les registres en constituant le corpus et la graine, les types de prédictions pour étiqueter les textes et, enfin, les méthodes pour filtrer les textes fiables ajoutés aux données d'entraînement. Des travaux préliminaires ont tenté de neutraliser les biais liés aux genres de textes et aux modalités orale et écrite. Ils ont également exploré la possibilité d'étiqueter un même texte avec plusieurs étiquettes pour illustrer l'usage hétérogène des registres de langue. Ces travaux préliminaires sont détaillés à la suite de cette section.

3.3 Travaux préliminaires : corpus TREMoLo-Web

Une phase exploratoire a consisté à reprendre les travaux menés par (Lecorvé et al. 2018) pour constituer un corpus de textes étiquetés en registres de langue. Elle a deux objectifs : réduire les biais introduits dans leur représentation, améliorer l'étiquetage en le rendant plus réaliste. En effet, le modèle proposé par (Lecorvé et al. 2018) associe des genres de textes avec des registres et étiquette un seul registre par texte. Or, nous supposons l'usage des registres moins uniforme que cela. Par exemple, l'extrait de *San Antonio, Les années 1980* donné citation (9) illustre la co-présence des registres familier et soutenu : le registre familier est présent à travers le lexique employé, tout comme le registre soutenu avec une phrase finale complexe.

(9) Lola regarde les lignes à l'encre bleue. Mais elle ne comprend pas l'angliche.
A preuve, elle grince :
— Cette foutue garce te file un rendez-vous, hein ?
Je hausse les épaules. Comme quoi en jalousie, il n'existe pas de femme bien élevée.
— C'est ça, dis-je, et je te porterai des oranges au parloir.
Elle dort, comme jetée en travers du lit, avec juste un bout de drap chiffonné sur les mollets, abandonnée et belle dans une précaire innocence retrouvée.

Pour atteindre nos deux objectifs, nous avons agi sur différents éléments du modèle introduit par la section précédente. Tout d'abord, pour neutraliser les biais dans la représentation des registres, nous avons proposé une recomposition de la graine. Puis, pour améliorer le réalisme de l'étiquetage, nous avons proposé un nouveau procédé d'annotation manuelle de la graine, choisi un autre type d'étiquetage, et introduit une nouvelle méthode de sélection des textes fiables. Cette section développe chacun de ces éléments en commençant par détailler le corpus utilisé et la composition de la graine.

3.3.1 Corpus et sous-corpus

La constitution du corpus est présentée, avant de détailler les trois sous-corpus : les deux premiers sous-corpus ont servi à évaluer l'étiquetage du modèle, le troisième est celui annoté manuellement (et appelé « la graine ») à partir duquel le modèle s'est entraîné à distinguer les registres entre eux.

Le corpus web Un ensemble de 100 000 textes a été sélectionné à partir du corpus construit par (Lecorvé et al. 2018) présenté section 3.1. Il est composé de 50 000 textes étiquetés par le modèle de (Lecorvé et al. 2018) comme appartenant au registre familier, et de 50 000 textes étiquetés comme appartenant au registre soutenu. Les textes sélectionnés ont ensuite été nettoyés et segmentés sur des frontières de paragraphes en portions de 5 000 caractères pour éviter de créer un manque d'homogénéité dans les longues pages web telles que celles des forums par exemple et d'introduire des biais dûs aux disparités de longueur des textes telles que celles entre un billet de blog ou bien un commentaire.

Les sous-corpus d'évaluation Deux sous-corpus sont utilisés pour évaluer notre modèle d'apprentissage automatique. Le premier, noté E_{graine}, est un sous-ensemble de textes issus de la graine permettant d'évaluer la capacité des classifieurs à généraliser l'annotation manuelle. Le second, noté E_{web}, est un sous-corpus de 58 textes extraits du corpus

web et annotés manuellement par les mêmes annotateurs experts que pour la graine. Ce second ensemble d'évaluation permet de vérifier la capacité des classifieurs à s'adapter à différents types de textes.

Composition de la graine La stratégie mise en place, pour réduire les biais liés aux genres de textes et aux modalités orale et écrite, donne le même nombre de genres pour chaque registre. Cinq genres de textes sont utilisés : des textes issus de romans [12], de journaux [13], de discussions en ligne [14], de pages web [15] et de transcriptions orales [16]. La graine compte 538 textes totalisant environ 681 000 mots. Nous donnons trois exemples d'extraits de pages web (10), (11) et (12) pour, respectivement, les registres familier, courant et soutenu.

(10) Vu que j'avais fait un article y a deux ans pour revenir sur tout Fast & Furious et que y a le 6 qui sort mercredi et que je l'ai vu hier, je me suis dit "azy faut écrire un truc dessus, en plus ça fait un bail que t'as pas écrit sur ton putain de blog, les gens qui le lisent vont croire que t'es mort ou que t'as trouvé mieux à faire que raconter des conneries ici". Bref. Fast & Furious 6. Un cas intéressant car c'est un film de merde archi cool. Je veux dire par là que si c'était un film Marvel, ce serait une merde infâme et ce serait déjà au panthéon des films de merde de 2013 mais FF6 réussit là où les Marvel échouent.

(11) Faut pas rêver Les vacances ne sont plus une parenthèse enchantée dès lors qu'elles deviennent le lieu d'arnaques à la location. Pourtant nous abordons cette période avec l'âme neuve d'un écolier en rupture de classe, un peu effaré, désarçonné de cette liberté retrouvée d'aller et de venir où bon nous semble, dans l'Hexagone ou hors de l'Hexagone. Mais il est très rare, à moins d'avoir des jambes et un sommeil de vingt ans, de pouvoir prendre son sac à dos du jour au lendemain pour l'endroit de notre choix. Il nous faut donc louer, retenir, verser des arrhes et c'est là où les loups nous attendent.

12. textes extraits des romans suivants : *Kiffe kiffe demain* (Faïza Guène), *Albertine disparue* (Marcel Proust), *Les Mohicans de Paris* (Alexandre Dumas), *Les Bâtisseurs de ponts* (Rudyard Kipling), *Les misérables* (Victor Hugo)
13. articles extraits du journal L'Humanité
14. extraites de webchat
15. extraites et sélectionnées manuellement : elles ne proviennent pas de l'ensemble automatiquement collecté
16. transcriptions extraites du corpus CEFC-ORFEO

(12) Vous avez indiscutablement une bonne connaissance du fonctionnement du FMI et des enjeux qui sont à l'œuvre. Cependant, pour ma compréhension personnelle et celle, sans doute, de bien d'autres lecteurs de ce Bolg, je voudrais que vous nous éclairiez de vos lumières concernant certaines interrogations. Tout d'abord, lorsqu'il s'agit de Monsieur, DSK, je constate qu'en tant que directeur général du FMI, il soulève une sympathie et une adhésion de la part de bon nombre des français, comme s'il était une sorte de Messie qui aurait compris la manière, non seulement de venir en aide des pais déjà en difficultés, comme c'est le cas de la Grèce mais en plus, avec la représentation que les outils que son institution mets en œuvre pour venir en aide de ces pays solliciteurs seraient, sans conteste, les plus performants pour les tirer d'affaire.

L'exemple (10) montre un extrait majoritairement familier contenant quelques expressions du registre soutenu telles que « infâme » ou bien « panthéon ». À l'inverse, l'exemple (12) illustre un extrait majoritairement soutenu avec quelques expressions du registre familier comme « une sorte de Messie » ou encore « tirer d'affaire ». Ces deux exemples mettent en avant la possibilité d'avoir plusieurs registres présents dans un même texte.

Protocole d'annotation manuelle de la graine Pour que des registres de langue puissent être co-présents dans un texte, nous avons annoté les textes de la graine en proportions de registres de langue. L'annotation manuelle a limité les résultats de l'annotation à 100% du texte. Par exemple, le texte de l'exemple (10) peut être annoté comme 90% familier et 10% soutenu, et celui de la citation (12) 90% soutenu et 10% familier. Comme (Peterson, Hohensee et Xia 2011), nous avons supposé les registres connus de tous les annotateurs. Dès lors, l'annotation s'est basée sur l'intuition des cinq annotateurs experts. Pour cette tâche d'annotation manuelle, un outil de validation a été développé puis utilisé afin d'annoter la graine. La figure 3.3 est une capture d'écran de cet outil où l'on peut voir à gauche le texte à annoter, à droite un curseur par registre, et un espace commentaire en bas si l'annotateur veut faire une remarque. La figure montre que nous avons considéré trois registres de langue, les registres familier, courant et soutenu, auxquels nous avons ajouté une quatrième catégorie appelée « poubelle » permettant d'étiqueter les textes de mauvaise qualité, c'est-à-dire mal encodé ou tronqué.

Résultats de l'annotation manuelle de la graine Tous les textes de la graine ont été annotés par trois annotateurs experts. L'annotation manuelle d'un texte est composée de quatre valeurs renvoyant chacune à la moyenne des trois annotations pour chaque

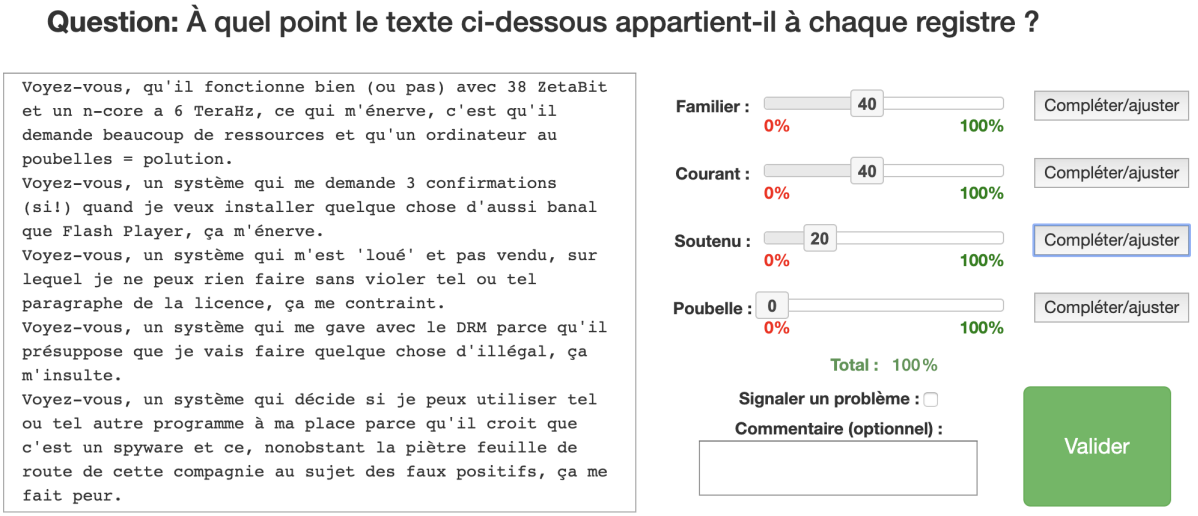

FIGURE 3.3 – Capture d'écran de l'outil utilisé pour annoter manuellement la graine employée pour l'annotation automatique du corpus TREMoLo-Web

registre. Sur les 538 textes de la graine, 57 ont été annotés comme appartenant à un seul registre de langue, 77% de ces textes mono-registre ont été annotés comme appartenant au registre courant. Plus de la moitié de la graine (297 textes) a été annotée avec deux registres présents. La majorité de ces textes (294 textes) a été annotée avec un registre dominant, c'est-à-dire avec une annotation dont la valeur est supérieure ou égale à 60% du texte. 176 textes ont été annotés avec trois registres différents, et 8 textes avec les quatre registres. L'exemple (13), est extrait d'un texte annoté comme appartenant à 53% au registre familier, 30% au registre courant, 10% au registre soutenu, et 7% à la catégorie poubelle.

(13) — Je vais faire une petite politesse à médème, me confie le Magistral ; j'sais bien qu'les chiches sont un peu contigus et qu'la place manque pour folâtrer, mais nous aut'qu'on a pas d'pétrole, on a des idées , pas vrai ?
Le paysage est à ce point sublime que je stoppe ma voiture de louage pour mieux le savourer.
Imagine des collines mauvies par la bruyère et sillonnées de ruisseaux qui courent approvisionner des lacs enchanteurs.
Des troupeaux de moutons sans berger, ovins blancs et têtes noires, paissent dans cette pastorale.
Ça et là, les tranchées des tourbières découpent le paysage en parcelles géométriques, lui donnant un caractère abstrait.

Cet exemple montre la difficulté de la tâche d'annotation manuelle, notamment due à la différence de registres entre la narration et les dialogues. Cette difficulté est illustrée par la dispersion des valeurs des annotations des trois annotateurs. En examinant ces valeurs pour chaque registre, nous constatons que cette dispersion est moins importante pour le registre familier. Les annotateurs ont donc plus souvent été d'accords pour le registre familier avec un écart type moyen de 9%, que pour les registres courant et soutenu avec, respectivement, des écarts types moyens de 18% et 12%. Ces résultats valident la pertinence d'une annotation en proportions de registres, puisque 90% des textes de la graine ont été annotés comme appartenant à au moins deux registres différents. Cette grande variabilité dans les registres doit pouvoir être prise en compte par notre modèle d'apprentissage.

3.3.2 Description du processus d'apprentissage semi-supervisé

Notre modèle d'apprentissage a été conçu dans l'objectif de généraliser l'annotation manuelle de la graine illustrant l'hétérogénéité de la langue. Il repose sur l'hypothèse qu'un ensemble de plusieurs classifieurs, chacun spécialisé dans l'étiquetage d'un seul registre, serait meilleur qu'un classifieur général pour étiqueter tous les registres. Ce processus d'apprentissage automatique nous a donné l'occasion de tester la pertinence des descripteurs linguistiques issus de l'étude (Mekki, Battistelli, Lecorvé et al. 2018) en les utilisant comme descripteurs d'apprentissage.

Descripteurs d'apprentissage En supposant qu'un ensemble de descripteurs pertinents conduit à des prédictions de qualité, nous avons comparé les prédictions obtenues en entraînant les classifieurs avec trois ensembles de descripteurs différents. L'objectif est de découvrir si un ensemble restreint mais expert (noté D_{exp}) peut être meilleur qu'un ensemble plus large mais générique (noté $D_{gén}$). Un troisième ensemble (noté $D_{exp+gen}$) est testé pour voir si la combinaison des deux premiers obtient de meilleures prédictions que les deux séparément. L'ensemble D_{exp} est constitué de 72 descripteurs experts issus de (Mekki, Battistelli, Lecorvé et al. 2018) couvrant divers niveaux d'analyse de la langue regroupés sous les catégories lexicale (16 descripteurs), morphologique (16), syntaxique (38) et phonétique (2). Des exemples de descripteurs sont donnés tableau 3.1 et l'ensemble complet est détaillé dans le tableau 1 de (Lecorvé et al. 2019). L'ensemble $D_{gén}$ regroupe 557 descripteurs plus génériques couvrant les niveaux phonétique avec, par

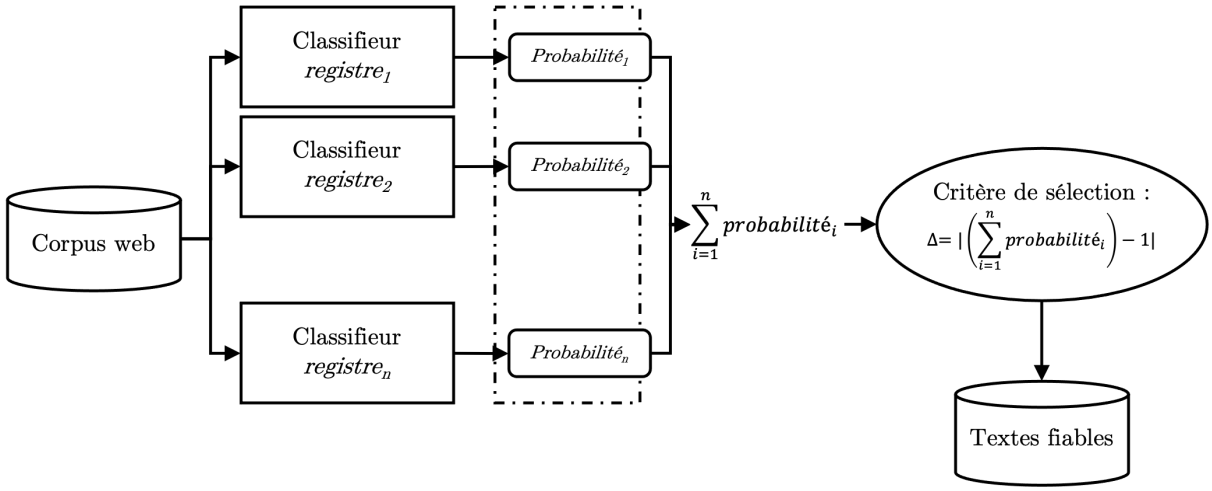

FIGURE 3.4 – Filtre des textes fiables avec un seuil sur la somme des probabilités des différents classifieurs.

exemple, le nombre et la diversité des phonèmes composant les mots [17] ; lexical avec les adverbes temporels et les connecteurs logiques ; morphosyntaxique avec la taille moyenne des mots, la proportion de verbes ; syntaxique avec la profondeur de l'arbre de dépendance et la variance du nombre de dépendances ; et enfin le niveau sémantique avec la moyenne des embeddings de mots composant les phrases, pour chacune des 500 coordonnées [18].

Classifieurs Pour illustrer la présence de plusieurs registres dans un même texte et avoir des probabilités dont la somme n'est pas égale à 1, nous avons étiqueté un texte avec autant de classifieurs que de registres de langue. Pour chacun d'eux, la première classe représente le registre et la seconde tous les autres. Chaque classifieur est un réseau neuronal entraîné à prédire un seul registre. Ainsi, un premier classificateur prédit le registre familier, un second le registre courant, et ainsi de suite. Ils comportent tous deux couches cachées de taille 10 avec *ReLU* comme fonction d'activation. La couche de sortie a une fonction d'activation *sigmoïde*. La première itération est entraînée à partir de la graine et les sessions d'entraînement successives sont faites par lots de 100 instances sur 20 *epochs*. Nous avons utilisé *rmsprop* comme algorithme d'optimisation afin d'implémenter les classifieurs. L'erreur absolue moyenne est utilisée comme fonction de perte.

17. par rapport aux phonèmes les plus courants
18. Les coordonnées ont été développées par (FAUCONNIER 2015) et apprises sur 1,6 milliard de mots (HMIDA et al. 2018).

Critère de sélection des textes fiables Nous avons choisi un critère de sélection, qui ne favorise pas la prédominance d'un registre dans le texte, nous permettant de respecter l'hétérogénéité des textes. C'est pourquoi le critère de sélection utilisé regarde la cohérence de l'ensemble des probabilités prédites par les classifieurs. Comme ces probabilités émanent de modèles indépendants, leur somme n'est pas égale à 100%. Si les classifieurs produisent des prédictions de qualité, ils doivent arriver à une sorte d'accord inter-classifieurs en étiquetant les textes de manière cohérente. Par exemple, un étiquetage cohérent serait un texte étiqueté par le classifieur dédié au registre familier comme étant à 50% familier, par le classifieur courant comme étant à 40% courant, par le classifieur soutenu comme étant à 10% soutenu, et enfin par le classifieur de la catégorie poubelle comme étant à 0% poubelle. L'étiquetage est cohérent car les proportions de registres de langue n'excède pas 100% du texte. À l'inverse, un étiquetage incohérent aurait été un étiquetage où la somme des proportions des registres de langue excéderait les 100% indiquant des prédictions de proportions très hautes pour tous les registres, c'est-à-dire qu'un texte serait à la fois complètement des registres familier, courant, et soutenu, ainsi que de très mauvaise qualité. De même, un texte avec un étiquetage dont la somme serait très inférieure à 100% serait un texte avec une grande proportion sans registre de langue indiquant un étiquetage incohérent. La figure 3.4 illustre comment le critère de sélection, noté Δ, filtre les textes fiables en s'appliquant sur la somme des prédictions de manière symétrique. Il s'applique sur la valeur absolue de la différence entre cette somme et 1 (équation 3.1) :

$$\Delta = |(\sum_{i=1}^{n} probabilite_i) - 1| \tag{3.1}$$

Plus la différence est grande, plus la prédiction est jugée incohérente donc non fiable ; à l'inverse plus la différence est faible, plus la prédiction est jugée cohérente donc fiable.

Ainsi, notre technique d'apprentissage semi-supervisée est construite afin d'étiqueter un texte en proportions de registres de langue sans prédominance d'un registre, tout en veillant à conserver cette variabilité des registres lors de la sélection des textes fiables ajoutés aux données d'entraînement. Les expériences, présentées à la suite de cette section, ont validé la robustesse de ce modèle.

3.3.3 Validation de la technique d'apprentissage semi-supervisée

Dans cette section, nous détaillons les paramètres des expériences avant de présenter les résultats obtenus validant la solidité de notre modèle d'apprentissage. Puis, nous présentons les résultats expérimentaux montrant la capacité du modèle à s'adapter à d'autres genres de textes. Ces résultats illustrent ainsi la possibilité de construire avec notre modèle d'autres types de corpus à partir de différentes graines. Toutes les expériences ont été développées en Python à l'aide de Keras [19].

Paramètres des expériences Le modèle a étiqueté les textes en considérant les trois registres familier, courant et soutenu, ainsi qu'en y ajoutant la catégorie poubelle. Dès lors, nous avons quatre classifieurs dont chacun d'eux doit prédire la proportion du texte appartenant à ces quatre catégories. Les trois ensembles de descripteurs d'apprentissage, D_{exp}, $D_{gén}$ et $D_{exp+gen}$, ont été utilisés afin d'en discerner le plus pertinent. Nous avons également fait varier la valeur du critère de sélection Δ, pour augmenter l'ensemble de textes de la graine : 0.20, 0.10, 0.15 et 0.01. Comme Δ s'applique sur la différence à 1 (équation 3.1), plus sa valeur est petite, plus Δ est sélectif. Faire varier les valeurs de Δ nous a permis de découvrir si un critère plus strict sélectionnant peu de textes mais avec des étiquetages très cohérents conduisait à de meilleures prédictions, qu'un critère plus lâche sélectionnant plus de textes mais avec des étiquetages moins cohérents. Enfin, nous avons utilisé les deux ensembles d'évaluation, E_{graine} et E_{web}, pour évaluer la qualité des prédictions obtenues. Toutes les expériences ont été réalisées en utilisant une validation croisée à 5 plis avec 1/5 de test représentant l'ensemble E_{graine}, 1/5 de validation, et 3/5 d'entraînement à partir du corpus web d'origine. Lors des itérations du processus semi-supervisé, pour un pli donné, les ensembles de test et de validation n'ont pas changé, seul l'ensemble d'entraînement a été augmenté avec de nouvelles observations.

Résultats de l'étiquetage automatique Pour évaluer la qualité des probabilités prédites, nous avons utilisé l'erreur quadratique moyenne (équation 3.2) indiquant la différence entre les valeurs prédites des étiquettes et les valeurs réelles des annotations.

$$EQM = \frac{1}{N} \sum_{i=1}^{N} (f_i - y_i)^2 \tag{3.2}$$

19. `https://keras.io/`

Valeur de Δ	Descripteurs d'apprentissage	EQM calculée avec E_{graine}	EQM calculée avec E_{web}
0.20	D_{exp}	0.07	**0.08**
	$D_{gén}$	**0.05**	0.09
	$D_{exp+gen}$	0.07	0.09
0.15	D_{exp}	0.07	**0.08**
	$D_{gén}$	**0.06**	0.09
	$D_{exp+gen}$	0.15	**0.08**
0.10	D_{exp}	0.08	**0.08**
	$D_{gén}$	**0.06**	0.09
	$D_{exp+gen}$	**0.06**	0.09
0.01	D_{exp}	**0.05**	**0.07**
	$D_{gén}$	**0.05**	0.09
	$D_{exp+gen}$	0.06	0.08

TABLE 3.2 – Résultats des évaluations de notre modèle selon la valeur de Δ, les descripteurs d'apprentissage utilisés, et l'ensemble de textes d'évaluation considéré.

où N est le nombre de la taille de l'ensemble de test, f_i la valeur renvoyée par le modèle et y_i la valeur réelle du point de données i. La table 3.2 montre les résultats obtenus en faisant varier la valeur de Δ, les descripteurs d'apprentissage et les ensembles de textes d'évaluation utilisés. Ces expériences ont mené à trois principaux résultats.

Tout d'abord, elles ont confirmé la possibilité d'étiqueter des textes en proportions de registres de langue avec de bons résultats validant la robustesse de notre modèle d'apprentissage. Elles ont prouvé sa capacité à généraliser les annotations manuelles de la graine en l'évaluant avec E_{graine}, ainsi que sa capacité à s'adapter à d'autres genres de textes en l'évaluant avec E_{web}.

Puis, les expériences ont montré que la valeur de Δ n'est pas déterminante pour la qualité de l'étiquetage automatique. Nous pouvons tout de même constater une légère amélioration des résultats avec une valeur stricte indiquant un impact de la valeur de Δ sur les prédictions des classifieurs. Ces résultats confirment la pertinence d'évaluer la cohérence des prédictions, mais ne nous permettent pas de déterminer nettement une valeur de Δ meilleure qu'une autre.

Enfin, en considérant la totalité des expériences menées avec E_{graine} et E_{web}, il est difficile de départager avec des résultats très proches. Notons tout de même que l'ensemble D_{exp} conduit cinq fois aux meilleurs résultats, contre quatre fois pour $D_{gén}$ et deux fois pour $D_{exp+gen}$. En revanche, en considérant les expériences uniquement évaluées

avec E_{graine}, l'ensemble $D_{gén}$ est systématiquement meilleur. À l'inverse, l'ensemble D_{exp} est tout le temps meilleur lorsque les prédictions sont évaluées avec E_{web}. Ces résultats indiquent qu'un ensemble volumineux composé de descripteurs génériques d'apprentissage permet aux classifieurs de bien reproduire une annotation manuelle, mais qu'il est moins robuste face à des genres de textes inconnus. Au contraire, un ensemble réduit de descripteurs linguistiques experts est meilleur pour généraliser l'annotation manuelle à de nouveaux genres de textes. Dans le cadre de notre travail, cela signifie que l'ensemble $D_{gén}$, prenant en compte beaucoup de traits linguistiques, rajoute du bruit dans la représentation des registres en introduisant par exemple des biais liés aux genres de textes. Quant à l'ensemble D_{exp}, considérant seulement des traits linguistiques ciblés, il évite ce bruit d'où sa robustesse face à de nouveaux genres de textes.

Ainsi, ces résultats montrent la qualité de l'étiquetage en proportions de registres de langue validant notre technique d'apprentissage semi-supervisée. Une seconde manière de le valider est d'explorer le corpus TREMoLo-Web, pour vérifier manuellement la qualité de l'étiquetage automatique.

3.3.4 Exploration linguistique du corpus TREMoLo-Web

Dans un premier temps, l'exploration linguistique du corpus TREMoLo-Web nous a confirmé la qualité de l'étiquetage. Dans un second temps, nous soulignons toutefois les limites de notre approche rendant difficile l'utilisation de TREMoLo-Web comme ressource textuelle pour caractériser automatiquement les registres de langue.

Confirmation de la qualité de l'étiquetage automatique En regardant manuellement l'étiquetage automatique, il semble de qualité avec des prédictions de proportions de registres fines. L'exemple (14) est un premier exemple de texte étiqueté comme 60% du registre familier, 25% du registre courant, 6% du registre soutenu et 7% de la classe poubelle. Cet étiquetage montre que le modèle est assez fin pour différencier la proportion d'appartenance à un registre, de la proportion textuelle portant ce registre. En effet, l'exemple (14) a été étiqueté comme appartenant majoritairement au registre familier. Cependant, nous pouvons constater que le registre familier est porté par des descripteurs linguistiques quantitativement minoritaires dans le texte, tels que l'abréviation de certains mots, la construction du futur avec le verbe « aller », ou encore l'utilisation d'anglicisme dans une métaphore. Les segments textuels portant ces descripteurs sont soulignés dans

les exemples donnés.

(14) Tout bascule lors d'une visite au musée, quand Mrs Dodds, la prof de maths, se transforme en une horrible créature et agresse Percy. J'ai ouvert ce livre surtout par obligation et en craignant le pire. En fin de compte, je l'ai suffisamment apprécié pour avoir envie de lire la suite. Comme j'ai quelques réserves, on va commencer par là. Déjà, le livre ne brille pas par son originalité. J'ai passé la première partie avec la désagréable sensation de lire un remake de Harry Potter à la sauce grecque.

L'exemple (15), quant à lui, semble montrer la capacité du modèle d'apprentissage à différencier les éléments cités entre guillemets, des éléments faisant partie du corps narratif. En effet, il a été étiqueté comme appartenant à 58% au registre familier, 32% au registre courant, 3% au registre soutenu et 5% à la classe poubelle. La proportion du texte perçue comme appartenant au registre familier pourrait être plus élevée, si les surnoms n'avaient pas été mis à distance par le locuteur avec les guillemets.

(15) Et les noms les plus improbables sont passé en revu, entre "Poil noir" et "Dédé" en passant par des trucs improbables, dont "Elo" ou "Mangetou" jusqu'à leur arrivé au pigeonnier à la nuit tombante. Quelques piécettes et 3 missives envoyées plus tard, le chemin du retour est pris, entre deux Pouic et quelques éclats de rire.
Acquiescement énergique de Minimoyette.
Du pain en taverne Municipale.
Un trajet de retour presque aussi calme qu'à l'aller.
Les triple D habitent pas vraiment l'arrondissement des bourges.

L'exemple (16) est étiqueté comme étant à 37% du registre familier, 20% du registre courant, 32% du registre soutenu et 9% de la classe poubelle ; il montre qu'un registre de langue peut se manifester par peu de traits linguistiques, mais avoir un impact fort sur la perception du lecteur. Ici, les répétitions contiguës de « . » suffisent à donner au texte une impression d'appartenance au registre familier, malgré la présence majoritaire d'un lexique courant et soutenu.

(16) Benyamin Nétanyahou, dans un communiqué, s'est pour sa part distancié de propos "inappropriés, qui ne reflètent pas mes positions ni les politiques du gouvernement".....Le premier ministre israélien entretient des relations notoirement exécrables avec Barack Obama, mais les deux hommes soulignent régulièrement que leurs différends personnels ne nuisent en rien à la qualité du lien entre Israël et les États-Unis .....

Enfin, les exemples (17) et (18) sont des textes étiquetés par le modèle comme appartenant à un registre fortement dominant. L'exemple (17) est étiqueté comme appartenant à 1%

au registre familier, 66% au registre courant et 93% au registre soutenu et l'exemple (17) à 3% au registre courant et 97% au registre soutenu. Ces deux exemples montrent une forte homogénéité du registre soutenu : seul le registre soutenu semble présent avec un lexique employé soutenu, ainsi que des phrases longues et syntaxiquement complexes. Aucun élément textuel pouvant être perçu comme familier n'est présent.

(17) Hâtez-vous de prendre place, sitôt qu'il vous en sera possible, car il semble bien que le moment tant attendu soit arrivé : les cors au repos se sont dressés à l'horizon et les tocsins commencent à se faire entendre en tout lieu des Terres de l'Échiquier. Nous y sommes donc enfin ! La bataille, croyez-le bien, est définitivement engagée, sans retour possible, et surtout avec un seul objectif désormais : ad victoriam ("vers la victoire") !

(18) La conversation que j'avais eue avec Albertine en rentrant du Bois avant cette dernière soirée Verdurin, je ne me fusse pas consolé qu'elle n'eût pas eu lieu, cette conversation qui avait un peu mêlé Albertine à la vie de mon intelligence et en certaines parcelles nous avait faits identiques l'un à l'autre. Car sans doute son intelligence, sa gentillesse pour moi, si j'y revenais avec attendrissement, ce n'est pas qu'elles eussent été plus grandes que celles d'autres personnes que j'avais connues. Mme de Cambremer ne m'avait-elle pas dit à Balbec : « Comment ! vous pourriez passer vos journées avec Elstir qui est un homme de génie et vous les passez avec votre cousine ! » L'intelligence d'Albertine me plaisait parce que, par association, elle éveillait en moi ce que j'appelais sa douceur, comme nous appelons douceur d'un fruit une certaine sensation qui n'est que dans notre palais.

Ces exemples de textes étiquetés automatiquement montrent la robustesse de notre modèle d'apprentissage. Mais, ils soulignent également quelques défauts au regard d'une utilisation du corpus TREMoLo-Web comme ressource textuelle, pour caractériser des registres de langue. Ces limites sont détaillées dans le paragraphe suivant.

Limites du corpus TREMoLo-Web Trois principales limites restreignent la qualité du corpus TREMoLo-Web pour la caractérisation des registres. La première limite est liée au choix de corpus composé de différents genres de textes. Malgré notre stratégie visant à neutraliser les biais liés aux genres de textes en composant une graine avec le même nombre de genres par registre, les genres bruitent encore la représentation des registres de langue. Par exemple, comment être certain que les caractéristiques, telles que les temps verbaux utilisés ou bien l'utilisation des discours rapportés, des textes des exemples (17) et (18) sont liées au registre soutenu et non au genre romanesque ? La seconde limite est inhérente

aux contraintes posées par les techniques de caractérisation des registres. Pour des raisons de complexité algorithmique, ces techniques ne peuvent pas considérer une unité textuelle aussi grande que celle étiquetée par notre modèle. En d'autres termes, l'unité considérée comme un texte, pour la caractérisation des registres, sera plus petite que les textes du corpus TREMoLo-Web. Dès lors, l'étiquetage automatique devient instable. La phrase suivante « J'ai ouvert ce livre surtout par obligation et en craignant le pire. » issue de l'exemple (14) sera étiquetée comme étant à 60% du registre familier, 25% du registre courant et 6% du registre soutenu, car extraite d'un texte étiqueté comme tel, alors qu'elle appartient à 100% au registre courant. Enfin, la troisième limite est liée aux descripteurs d'apprentissage que nous avons choisis comme pertinents pour distinguer les registres. Ne seront-ils pas similaires aux motifs extraits par les techniques de caractérisation des registres, car justement permettant de distinguer les registres entre eux ? Par exemple, en disant aux classifieurs que la présence du terme « ça » permet de distinguer le registre familier du registre soutenu, nous risquons lors de la caractérisation du registre familier d'extraire ce même terme.

Cette section a détaillé les expériences confirmant la pertinence d'un étiquetage illustrant la variabilité des registres, et validant un ensemble de descripteurs linguistiques experts (issus de Mekki, Battistelli, Lecorvé et al. 2018) ainsi que la robustesse de notre modèle. Elle a également souligné trois principales limites détaillées ci-dessus. Pour répondre à chacune d'elles, nous avons adapté notre technique d'apprentissage semi-supervisée dans le but de constituer un corpus de tweets étiquetés en proportions de registres de langue. La section 3.4 présente nos motivations linguistiques pour la constitution de ce corpus, appelé TREMoLo-Tweets, et les travaux conduits dans le cadre de sa création.

3.4 Corpus de référence pour les registres de langue français : TREMoLo-Tweets

Nous basant sur la même approche, nous avons modifié plusieurs éléments de notre modèle pour répondre aux limites mises au jour avec les travaux préliminaires exposés dans la section précédente. Ces modifications y répondent :

— en choisissant un seul genre de textes, dont le format court permet d'avoir la même unité lors de l'étiquetage en registres, et lors de la caractérisation des registres ;

— en utilisant un classifieur pré-entraîné, en l'occurrence le modèle CamemBERT (Martin, Benjamin Muller, Suárez et al. 2019), nous évitant de sélectionner des descripteurs d'apprentissage.

Dans cette section, nous motivons la constitution de ce second corpus, avant d'introduire le nouveau protocole d'annotation manuelle. Nous présentons ensuite le processus semi-supervisé d'étiquetage automatique généralisant les résultats de cette tâche d'annotation. Ces travaux ont fait l'objet de deux publications (Mekki, Battistelli, Béchet et al. 2021 ; Mekki, Lecorvé et al. 2021).

3.4.1 Constitution du corpus TREMoLo-Tweets

La constitution d'un corpus de textes écrits représentatif de l'usage réel des registres de langue présente deux difficultés majeures : tout d'abord le lien bi-univoque fort entre certains registres et certains types de textes (par exemple le soutenu associé à des romans de la littérature classique, le familier aux forums de discussion, et le courant à des dépêches journalistiques) ; ensuite l'association quasi immédiate de la modalité orale avec le registre familier d'une part, et de la modalité écrite avec les registres courant ou soutenu d'autre part (Gadet 2000 ; Rebourcet 2008).

Pour répondre à ces biais, nous avons choisi de construire notre corpus à partir d'un seul genre de textes issu des CMO [20] définis comme « toute communication humaine qui se produit à travers l'utilisation de deux ou plusieurs dispositifs électroniques » (McQuail 2010). Un des intérêts des CMO sur le plan linguistique réside dans le fait qu'ils contribuent à créer un « parlécrit » (Jacques 1999) par le caractère instantané des échanges qu'ils matérialisent ; l'intérêt des tweets en particulier parmi les CMO est leur limite à 280 caractères, imposée par Twitter, ce qui homogénéise la taille des textes produits et analysés. L'extraction automatique des tweets a été conduite en s'appuyant sur l'hypothèse qu'en collectant les tweets qui contiennent les hashtags les plus utilisés à un moment donné (noté « SA » pour **S**ujet d'**A**ctualité) dans une zone géographique donnée, la diversité des productions devrait être représentative des différentes fonctions du langage et de différents registres de langue. L'API de Twitter [21] permet, à partir d'un identifiant de lieu (dans notre cas celui de Paris), de récupérer automatiquement 50 SA. Pour chaque SA, une extraction a recherché tous les tweets le mentionnant. Afin de couvrir le plus grand nombre d'usages et donc de sujets différents, 10 extractions ont été faites à 10

20. Pour **C**ommunications **M**édiées par **O**rdinateurs
21. `https://developer.twitter.com/en/docs`

dates différentes pour couvrir la totalité du mois d'août 2020. Les tweets non français ont été repérés grâce à un module Python [22] qui, pour un texte donné, prédit la probabilité de chaque langue possible. Si la probabilité est supérieure à $0,9$ pour le français alors le texte est conservé dans le corpus, si non il en est exclu. Nous avons fixé la valeur de $0,9$ pour filtrer les textes afin de garder ceux avec la présence de quelques termes non français intéressants tels que « lol », « dead », « stan »... Quant aux tweets tronqués, ils ont été repérés grâce au caractère unicode U+2026, « … », et ont été supprimés. Finalement, après ces filtrages, le corpus compte 228 505 tweets, pour 6 201 339 mots. La totalité du corpus a été traitée en remplaçant les identifiants des utilisateurs par *@X*, et les URLs par *url_path*.

3.4.2 Annotation manuelle de la graine

La première étape de nos travaux d'annotation de la graine a été de comprendre les codes linguistiques propres aux tweets en analysant une partie du corpus. Ces codes reposent notamment sur l'utilisation d'éléments linguistiques particuliers comme les identifiants d'utilisateurs, les hashtags, les URLs ou encore les pictogrammes. Suivant notre définition des registres basée sur le respect partiel ou total des normes d'usage et grammaticale, cette analyse du corpus a été nécessaire pour lister les usages linguistiques spécifiques aux tweets. Une de nos contributions est de les intégrer à un ensemble de descripteurs linguistiques pour analyser les registres de langue à partir d'un corpus de CMO [23], au lieu de les écarter comme dans (Agarwal et al. 2011 ; Pak et Paroubek 2010 ; Go, Bhayani et L. Huang 2009). Ensuite, nous avons mis en place un nouveau protocole d'annotation manuelle permettant d'attribuer des proportions de registres dont les valeurs ne sont pas fixées par l'annotateur, mais par une méthode reposant sur un système de rangs hiérarchisant la présence des registres dans un même texte. Ce nouveau protocole est proposé car les résultats de l'annotation manuelle de la graine du corpus TREMoLo-Web nous ont montré que la perception des proportions de registres ne correspondait pas systématiquement à la proportion textuelle portant ce registre. Ce décalage rend l'attribution d'une valeur de proportion de registre trop dépendante de la subjectivité de l'annotateur.

Nous introduisons, dans le paragraphe suivant, les principaux descripteurs linguistiques propres aux tweets que nous avons intégrés à l'ensemble de descripteurs pour

22. `https://pypi.org/project/langdetect/`
23. Cet ensemble de descripteurs est accessible dans notre guide d'annotation.

l'analyse des CMO. Puis, nous détaillons le protocole d'annotation de la graine, avant de développer la campagne d'annotation appliquant ce protocole. Enfin, nous concluons cette section en présentant les annotations manuelles résultantes de ce travail.

Descripteurs linguistiques pour l'analyse des CMO Un des enjeux linguistiques de l'intégration des formes propres aux tweets est de leur donner des fonctions linguistiques afin de pouvoir les décrire. Ces formes sont désignées par (Paveau 2013) avec le terme « technomorphèmes ». Parmi elles, les hashtags sont définis comme un ou plusieurs mots contigus précédés d'un # (par exemple « #Rentrée2020 »). Certaines typologies de hashtags mettent l'accent sur leur fonction d'indexation (Jackiewicz et Vidak 2014). En plus d'intégrer à notre analyse ce type de fonction pour les hashtags nous pensons qu'elle joue un rôle dans la perception de registres dans les tweets, nous avons proposé d'intégrer en outre le degré d'intégration syntaxique plus ou moins fort des hashtags selon trois cas. Nous donnons ci-dessous chaque cas avec un exemple de tweet l'illustrant :

1. intégration syntaxique, c'est-à-dire qu'il assume une fonction syntaxique au sein de la phrase ;

   (19) Essor de la #télémédecine, attractivité des métiers et établissements, enseignements et perspectives de la crise #underline, bilan du #SegurDeLaSante ... Autant de sujets présents aux conférences de #SANTEXPO 2020. RDV du 7-9 oct à Paris ! 👉 le programme 👉 url_path url_path

2. indépendance syntaxique, c'est-à-dire qu'il n'assume pas de fonction syntaxique au sein de la phrase ;

   (20) RT @X : Caleb Ewan est tombé dans cette chute ! Un des favoris du jour. #TDF2020

3. sans rapport syntaxique entre eux, lorsqu'ils sont écrits côte à côte sans avoir de rapport syntaxique entre eux. Cela crée un effet de juxtaposition c'est-à-dire qu'il n'y a pas de rapport syntaxique mais que la liaison syntaxique se fait « par simple rapprochement excluant la coordination et la subordination » [24].

   (21) Marche arrière toute direction la Terre ! 🧐 #espace #astronomie url_path

24. https://www.cnrtl.fr/definition/juxtaposition

Un autre type de technomorphèmes est le pictogramme qui se réfère à la fois à un « émoticône »[25] et à un « emoji »[26]. Nous avons utilisé les trois fonctions de la typologie proposée par (Magué, Rossi-Gensane et Halté 2020) en les adaptant à l'analyse de notre corpus. Nous les présentons ci-dessous, chacune d'elle est suivie d'un exemple l'illustrant :

1. la fonction de remplacement, quand un pictogramme remplace un syntagme ;

   (22) RT @X : La 🇫🇷 aux côtés de nos amis libanais... #Beyrouth #Liban url_path

2. la fonction d'illustration, quand il a une fonction référentielle ;

   (23) Westbrook a quasi 32, il a autant d'énergie qu'à 25. Je ne sais pas ce qu'il bouffe. Du feu 🔥 peut-être. #nbaextra

3. la fonction de modalisation, quand il indique l'émotion ou l'attitude énonciative de l'auteur.

   (24) @X Si on rajoute les gens qui crient pendant qu'ils parlent (non j'ai pas été voir BTS en concert mais j'ai déjà vue des concerts) 😭 😭 😭 😭

Nous avons ajouté à cette typologie une autre fonction :

4. la fonction d'encadrement/structuration, lorsque le pictogramme entoure ou pointe vers une information.

   (25) ➡️ Vous souhaitez préparer un BTS Communication en alternance ? Notre partenaire dans le secteur automobile propose un poste de chargé.e de communication, adressez votre candidature au plus vite ! 👉🏼 url_path #apprentissage #formation #alternance #recrutement url_path

En mettant à jour une liste issue d'une étude qui avait déjà identifié des descripteurs dans la littérature scientifique pour les registres de langue (Mekki, Battistelli, Lecorvé et al. 2018) avec ces traits spécifiques aux technomorphèmes, notre annotation prend au final en compte un ensemble de 47 descripteurs experts (dont certains sont présentés tables 3.4 et 3.5) répartis dans divers niveaux d'analyse. La table 3.3 montre la répartition de ces descripteurs dans les niveaux d'analyse de la langue, en donnant le nombre de descripteurs par niveau d'analyse.

25. Un émoticône est un signe graphique ressemblant à une émotion Beccucci 2018.
26. Un emoji est un symbole répertorié dans une base de données (ibid.).

Niveau d'analyse	**Nombre de descripteurs**
Syntaxique	15
Lexico-syntaxe	5
Discursif	6
Lexical	9
Morphologique	9
Phonologique	3
Total	**47**

TABLE 3.3 – Détails quantitatifs des descripteurs par niveau d'analyse de la langue extraits de (MEKKI, BÉCHET et al. 2020).

Protocole d'annotation de la graine Pour rendre l'attribution de la valeur des proportions de registres moins dépendante de la subjectivité des annotateurs, nous proposons un protocole demandant aux annotateurs d'attribuer un rang aux registres de langue et non une proportion de registres. Ces annotations hiérarchisant la présence des registres sont ensuite converties en proportions de registres. Dans notre protocole d'annotation, l'annotateur doit ordonner les registres en fonction de leur prédominance dans un texte en leur attribuant un rang [27]. À chaque fois qu'il attribue un rang, il doit le justifier par la présence d'au moins un descripteur de la liste présentée dans le guide d'annotation issue de notre analyse linguistique du corpus. Chaque rang est ensuite transformé en proportion de registre. Soit :

- R, l'ensemble des registres présents dans un texte ;
- *rang*, la correspondance qui associe à tout registre $r \in R$ son rang ;

$$\begin{aligned} rang : \quad & R \longmapsto [\![1, Card(R)]\!] \\ & r \longrightarrow rang(r) \end{aligned}$$

- *inv-rang*, la correspondance entre un registre r et son rang inversé, c'est-à-dire depuis le dernier rang possible :

$$inv\text{-}rang(r) = Card(R) - rang(r) + 1$$

- S, la somme des rangs présents dans le texte.

27. À noter que ne pas mettre de rang signifie que le registre n'est pas présent dans le texte selon l'annotateur.

$$S = \sum_{r \in R} rang(r)$$

Nous définissons alors la proportion $prop(r)$ de chaque registre du texte comme suit :

$$prop(r) = \frac{inv\text{-}rang(r)}{S}.$$

Par exemple, si un annotateur décide qu'un tweet appartient aux registres familier, soutenu et courant, mais que le registre familier est plus présent que le registre courant, lui même plus présent que le registre soutenu ; alors on obtient :

- R ={*familier, courant, soutenu*} ;
- $rang(familier) = 1$;
- $rang(courant) = 2$;
- $rang(soutenu) = 3$;
- *inv-rang(familier)*=3 ;
- *inv-rang(courant)*=2 ;
- *inv-rang(soutenu)*=1 ;
- $S = 6$;
- $prop(familier) = \frac{3}{6} = 50\%$;
- $prop(courant) = \frac{2}{6} = 33\%$;
- $prop(soutenu) = \frac{1}{6} = 17\%$.

Les proportions des registres sur cet exemple sont donc familier 50%, courant 33% et soutenu 17%.

Campagne d'annotation manuelle de la graine Pour annoter manuellement la graine, nous avons sélectionné au hasard 4 000 tweets à partir du corpus de tweets. Comme pour les travaux préliminaires, nous avons considéré trois registres familier, courant et soutenu auxquels une catégorie « poubelle » est ajoutée pour les tweets mal encodés ou incompréhensibles. Chaque texte a été annoté par deux annotateurs experts [28]. N'ont été conservées que les étiquettes présentes dans l'intersection des deux annotations. Pour chaque étiquette, la moyenne de sa proportion a été calculée. Sur les 4 000 tweets annotés

28. Les annotateurs sont des doctorant.e.s ou chercheur.e.s en sciences du langage spécialisé.e.s en écrits numériques ou en TAL.

manuellement par quatre annotateurs, 976 textes ne se sont pas retrouvés dans l'intersection entre les deux annotations. Une seconde annotation a alors été faite par un 5e annotateur expert. Après cette seconde annotation, la totalité des 976 tweets s'est retrouvée dans l'intersection. Au final, 3 269 tweets annotés manuellement sont conservés, car dans l'intersection des annotations de deux annotateurs différents, soit 81,73% des textes initialement sélectionnés pour constituer la graine.

Résultats de l'annotation manuelle de la graine Puisque l'annotateur doit justifier l'attribution d'un registre en sélectionnant des descripteurs linguistiques présents dans le texte, la campagne d'annotation manuelle a produit des tweets annotés en registres de langue, mais aussi des descripteurs catégorisés en registres. Pour le premier ensemble, nous avons regardé les proportions des textes selon leurs registres dominants. Nous avons constaté que les résultats de l'annotation manuelle sont dominés par le registre courant (51% de la graine), puis par le familier (39%), et enfin le soutenu (10%). Pour le second ensemble, nous l'avons exploré plus finement en regardant les descripteurs caractéristiques de chaque registre. Afin de faire cela, nous avons caractérisé un registre de langue cible r_c par rapport aux autres registres combinés ensembles (notés r_s pour registre source), en mesurant l'importance de chaque descripteur D vu dans r_c à partir du calcul de son taux de croissance (TC) obtenu à partir de sa fréquence relative, notée $freq$, dans r_c et r_s :

$$TC(D_{r_c|r_s}) = \begin{cases} \infty, & sifreq_{r_s}(D) = 0 \\ \frac{freq_{r_c}(D)}{freq_{r_s}(D)}, & sinon \end{cases} \tag{3.3}$$

Si $TC(D_{r_c|r_s}) > 1$, alors D est émergent pour r_c puisque sa présence est plus importante dans r_c que dans r_s. La table 3.4 présente les trois descripteurs les plus émergents caractérisant les registres cibles dans la graine annotée manuellement. Pour chacun d'eux, son taux de croissance est donné avec le détail des fréquences relatives associées, un exemple illustre le descripteur. Nous pouvons constater que tous les TC du registre courant sont inférieurs à ceux du registre familier et soutenu, ces valeurs soulignent ses limites floues avec les autres registres. Au contraire, le familier présente des TC aux valeurs élevées qui indiquent la présence de formes très spécifiques pour ce registre. De plus, la présence de technomorphèmes pour les registres courant et soutenu signifie que certains éléments spécifiques aux CMO ont été intégrés à la norme grammaticale.

Familier	*TC*	**F vs. Autres**	**Exemple**
Remplacement du « il » par « y »*	33,2	6,6% / 0,4%	*Y* sont pas sérieux
Répétitions de caractères*	29,5	11,3% / 0,5%	L*ooo*l
Onomatopées	22,5	11,7% / 0,7%	oh
Courant	*TC*	**C vs. Autres**	**Exemple**
Présent comme unique temps*	3,2	13,8% / 3,1%	on lui *coupe* le pied, il *doit* jouer
Agglutination en nom propre*	2,8	2,9% / 0,7%	*#ThierryBodson* revient sur
# sans relation syntaxique*	2,7	7,5% / 2,0%	*#fastfashion #slowfashion*
Soutenu	*TC*	**S vs. Autres**	**Exemple**
Inversion sujet verbe	12,8	20,0% / 1,5%	*As tu* lu X
Diversité des connecteurs logiques	7,4	20,0% / 2,6%	*car* [...] *et*
Discours rapporté*	6,8	38,2% / 5,3%	Merci chère amie, *dit-elle*

TABLE 3.4 – Top 3 des descripteurs (* : issus de notre analyse linguistique) qui caractérisent les registres dans la graine annotée manuellement. Chaque descripteur est donné avec son taux de croissance et fréquences relatives associées, ainsi qu'un exemple. Le taux de croissance, noté *TC*, est introduit par l'équation 3.3.

3.4.3 Étiquetage automatique du corpus TREMoLo-Tweets

Notre objectif est d'obtenir un corpus entièrement étiqueté en registres. Or, nous ne disposons que de 3 269 textes annotés manuellement de la graine. Comme déjà fait précédemment, notre approche vise à augmenter l'ensemble de données d'entraînement à partir duquel un classifieur étiquette automatiquement la totalité du corpus tel que décrit en section 3.2. Cette section détaille le classifieur multi-étiquettes utilisé et les critères de sélection employés pour filtrer les textes ajoutés aux données d'entraînement.

Classifieur Nous avons le modèle de langage pré-entraîné CamemBERT, que nous avons affiné sur une tâche de classification multi-étiquettes, afin de ne pas biaiser la caractérisation des registres, en choisissant des descripteurs d'apprentissage. Nous avons pris le modèle CamemBERT-base [29] entraîné sur le corpus OSCAR [30] pour nos expériences.

Critère de sélection des textes fiables Deux critères de sélection sont introduits : soit un seuil sur la probabilité d'appartenance à un registre (noté $T1$), soit un seuil sur le nombre maximum de textes à ajouter (noté $T2$). Le premier seuil $T1$ filtre les textes en regardant si le classifieur est sûr de lui en attribuant une haute probabilité d'appartenance à un registre. $T1$ est un seuil à dépasser : tous les textes ayant une probabilité qui dépasse ce seuil sont jugés fiables et sont ajoutés aux données d'entraînement. $T1$ garde la même

29. `https://huggingface.co/camembert-base`
30. `https://oscar-corpus.com/`

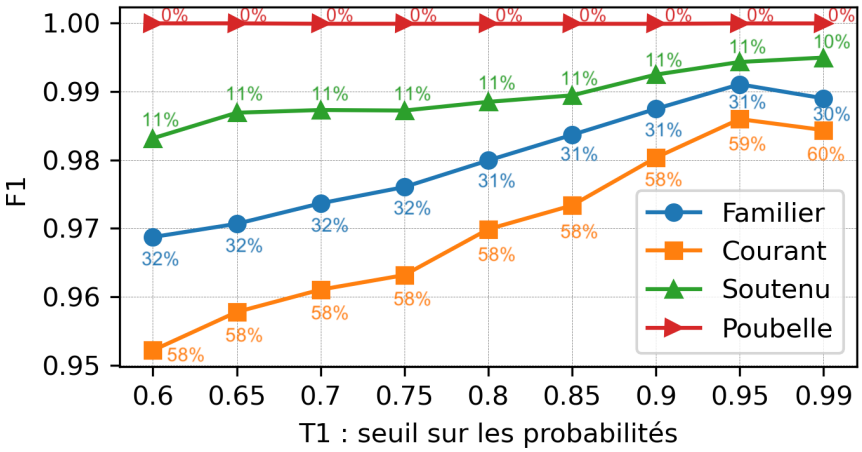

FIGURE 3.5 – $F1$ pour chaque registre et leur pourcentage dans l'ensemble des textes à ajouter à chaque valeur de $T1$

valeur pour les trois registres de langue et la classe poubelle. Le nombre de textes ajoutés varie selon les probabilités prédites. À l'inverse, le second seuil $T2$ sélectionne toujours le même nombre de textes. Pour filtrer les textes avec $T2$, toutes les probabilités sont triées par ordre décroissant et pour chacun des quatre registres les $\frac{T2}{4}$ premiers textes sont ajoutés. Autrement dit, $T2$ filtre les textes assumant que les textes fiables sont ceux dont les probabilités ont les valeurs les plus hautes parmi toutes celles prédites.

Paramètres des expériences Nous avons discrétisé les proportions de chaque registre pour chaque tweet, à savoir si la proportion est inférieure à 50% le registre est jugé absent, si non il est jugé présent. Lors de nos expériences, le processus d'apprentissage semi-supervisé n'a fait qu'une itération : après avoir initié l'entraînement à partir de la graine, le classifieur étiquette le corpus, les textes fiables sont filtrés, le classifieur commence un second apprentissage à partir de la graine augmentée, les étiquettes prédites après ce second entraînement sont considérées comme les étiquettes finales. Pour comparer la qualité des étiquettes prédites par rapport aux annotations de la graine, nous avons utilisé la F-mesure (notée $F1$). Les paramètres sont fixés à 10^{-4} pour le taux d'apprentissage, 8 pour le nombre d'*epochs*, et une division de la graine 90%/10% d'entraînement/test. Toutes les expériences ont été implémentées avec K-train [31].

31. `https://github.com/amaiya/ktrain`

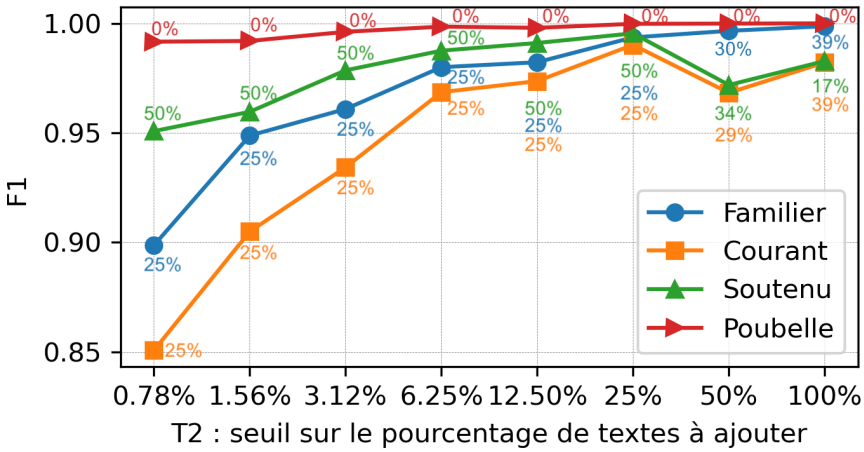

FIGURE 3.6 – $F1$ pour chaque registre et leur pourcentage de textes avec une probabilité ≥ 0.90 pour 1 des 4 registres à chaque valeur de $T2$

Résultats de l'étiquetage automatique Nous avons fait varier les valeurs de $T1$ et $T2$, afin de découvrir si une de ces valeurs amenait de meilleures prédictions. La figure 3.5 montre les résultats des expériences faisant varier le seuil $T1$: l'axe horizontal donne les différentes valeurs de $T1$, l'axe vertical indique les valeurs de la mesure $F1$. Pour chaque expérience, les résultats obtenus sont détaillés en précisant par registre son score $F1$, avec le pourcentage de textes du corpus appartenant au registre, par rapport à l'ensemble total des textes du corpus. Considérant qu'un texte appartient à un registre lorsque la probabilité associée est strictement supérieure à 50%, et que l'étiquetage est multi-étiquettes : un texte peut être étiqueté comme appartenant à plusieurs registres à la fois, ou à l'inverse à aucun registre. C'est pourquoi la somme des proportions de textes du corpus par registre peut être supérieure ou inférieure à 100%. Nous pouvons constater sur la figure 3.5 une détérioration des résultats entre 0,95 et 0,99 : le déséquilibre de la répartition des registres dans les données d'entraînement s'accentue légèrement à 0,99 (60% de textes courant, 30% familier et 10% soutenu). La figure 3.6 donne les résultats des expériences avec les différentes valeurs de $T2$ indiquées par l'axe horizontal, et les scores de $F1$ précisés par l'axe vertical. Une échelle logarithmique est prise pour faire varier $T2$. Les meilleurs scores sont obtenus lorsque $T2$ est fixé à 25%. La légère dégradation des $F1$ au delà de 25% peut être due à la baisse du pourcentage de textes ayant une probabilité $\geq 0,90$ pour un des quatre registres : il est de 100% lorsque $T2$ est à 25% et décroît à 93% et 95% lorsque $T2$ passe à 50% puis 100%. $T2$ à 25% (66 369 textes) semble donc être un bon

équilibre entre la quantité et la qualité des données. Enfin, la distribution des registres dans le corpus finalement obtenu de manière semi-supervisée confirme la qualité de l'étiquetage, puisqu'elle est relativement similaire à celle de la graine annotée manuellement : 31% familier, 59% courant et 10% soutenu.

L'évaluation de l'étiquetage automatique du corpus TREMoLo-Tweets a confirmé la robustesse de notre processus d'apprentissage semi-supervisé déjà validé par nos travaux préliminaires sur le corpus TREMoLo-Web. Pour vérifier la qualité de l'étiquetage automatique, nous avons également exploré manuellement le corpus TREMoLo-Tweets.

3.4.4 Exploration linguistique du corpus TREMoLo-Tweets

L'exploration linguistique du corpus TREMoLo-Tweets a deux objectifs : confirmer la qualité de l'étiquetage automatique en registres de langue, montrer que les descripteurs linguistiques proposés sont pertinents pour l'analyse des registres de langue à partir de CMO. Après avoir donné une vision d'ensemble des descripteurs caractéristiques des registres, nous donnons pour chaque registre des exemples de textes montrant la pertinence des descripteurs linguistiques intégrant des technomorphèmes.

Comparaison des descripteurs linguistiques selon les registres de langue Nous voulions découvrir si des descripteurs linguistiques, issus de notre analyse linguistique d'une partie du corpus, étaient caractéristiques d'un registre particulier à partir du corpus étiqueté automatiquement. Pour cela, nous avons implémenté des règles symboliques pour chaque descripteur cherchant à les identifier dans les tweets afin de calculer leurs fréquences relatives par sous-corpus de registres [32]. À partir de ces fréquences, nous avons calculé les taux de croissance de tous les descripteurs linguistiques pour les registres cibles familier, courant et soutenu. La table 3.5 présente les trois descripteurs les plus émergents caractérisant les registres cibles dans le corpus étiqueté automatiquement. Pour chacun d'eux, son taux de croissance est donné avec le détail des fréquences relatives associées, un exemple illustre le descripteur. Pour les registres courant et soutenu, plusieurs traits émergents contiennent des technomorphèmes tels que des hashtags ou bien des pictogrammes. Ces résultats semblent valider la pertinence des descripteurs linguistiques que

32. le nombre d'occurrences du descripteur dans le sous-corpus divisé par le nombre total de textes du sous-corpus

Familier	*TC*	**F vs. Autres**	**Exemple**
Orthographe électronique	8.3	6.7% / 0.8%	Ha *ptdrrrr*
Remplacement du « il » par « y »	2.5	6.5% / 2.2%	*Y'*en a le 25
Motif « juste »*	2.1	0.5% / 0.2%	*Juste* comme ça
Courant	*TC*	**C vs. Autres**	**Exemple**
Absence d'un item attendu	2.2	0.1% / 0.07%	ils ⊘ vont quand même pas
# sans relation syntaxique*	1.6	11.4% / 7.3%	*#stress #bonheur*
# indépendant syntaxiquement*	1.4	12.5% / 8.7%	[...] . *#MondayMotivation*
Soutenu	*TC*	**S vs. Autres**	**Exemple**
Fonction d'encadrement ou de structuration du pictogramme*	6.7	2.2% / 0.3%	🌲 [ECOLOGIE] 🌳 À #Montréal, 🔴 #X banni de #Facebook 🚩 [Webinar] J-1 « Le bilan à 6 ans »
Phrase avec ponctuation*	2.3	57.1% / 24.3%	VIDEO. Crise des transports *:*
# intégré syntaxiquement*	2.3	10.8% / 4.7%	les *#ViolencesPolicieres* ne sont pas

TABLE 3.5 – Top 3 des descripteurs (* : issus de notre analyse linguistique) qui caractérisent les registres dans le corpus étiqueté automatiquement. Chaque descripteur est donné avec son taux de croissance et fréquences relatives associées, ainsi qu'un exemple.

nous avons proposés. Les paragraphes suivants avec des exemples de textes illustrant ces nouveaux descripteurs confirment leur pertinence.

Analyse des textes du registre courant Le registre courant montre des usages commerciaux (exemples (26) et (27)), ou de communications institutionnelles ou marketing (respectivement, exemples (28) et (29)) des tweets qui mettent à profit la fonction d'indexation des hashtags afin de les rendre investigables :

(26) Bonne nouvelle pour les amoureux des versions boites. Le jeu #MonstrumGame de @X sort en physique le 23 octobre sur Nintendo Switch, PlayStation 4 et Xbox One ⬇️ http ://path.com

(27) @X @X Vive le jeu #UltimatePlay

(28) RT @X : L'appel de #LionelJospin au rassemblement de « la #gauche écologiste » pour #2022. http ://path.com

(29) Retrouvez-moi dans le 12/13 et le 19/20 de @X. Pour la #RentreeScolaire, le @X dote les collégiens de masques. #RENTREE2020 http ://path.com

Le phénomène d'agglutination pour des noms propres sert également à créer de nouveaux mots qui réfèrent à de nouveaux produits (exemples (30) et (31)), ou à des évènements (exemple (32)) :

(30) 🎉 Mario Kart dans ton salon, c'est le 16 octobre ! Voila, en plus de #PokemonGO, une nouvelle application incroyable de la #réalitéaugmentée http ://path.com

(31) @X Trop tôt pour mon portefeuille avec la sortie de #NBA2k21 du calme nintendo 🙄

(32) 🎮 GAMING | Le 13 septembre, le plus célèbre des plombiers du monde du jeu vidéo aura 35 ans ! 🍄 Pour célébrer cela comme il se doit, #Nintendo a tout prévu. • Le programme complet est détaillé dans un nouveau #SuperMario35 Anniversary Direct à voir dès maintenant. ⬇️ http ://path.com

Enfin, mentionnons l'utilisation des pictogrammes afin de renvoyer le lecteur à des liens hypertextes qui mènent à des sites extérieurs à Twitter (comme dans les exemples (33) à (36)) :

(33) Console Game & Watch : Super Mario Bros est de retour en précommande à 44,99€ ➡️ Cdiscount http ://path.com ➡️ Leclerc (49,90€) http ://path.com ➡️ Fnac (49,99€) http ://path.com ➡️ Cultura (49,99€) http ://path.com

(34) 📚 🤹 Culture @X met à disposition de nombreuses ressources culturelles #confinement : 🔗 http ://path.com

(35) Droits TV : la Ligue des champions sur Téléfoot et RMC Sport cette saison ? 👉 http ://path.com http ://path.com

(36) Je viens de poster mon tout premier remix sur #Soundcloud, n'hésitez pas à me donner vos avis 🎵 👇 http ://path.com

Analyse des textes du registre soutenu Le registre soutenu, quant à lui, intègre syntaxiquement les hashtags en les utilisant comme des *mots classiques* dans une phrase (cf. exemples (37) à (40)) :

(37) « Ne pas nommer les choses c'est ajouter du malheur au monde » A.Camus Les violences vécues en #France ne sont pas des #incivilites comme le dit @X mais des meurtres Les mots ne suffisent plus il faut passer aux actes et renforcer l'efficacité de notre chaîne pénale

(38) Sondage riche d'enseignements de #CharlieHebdo alors que s'ouvre le procès des attentats terroristes de janvier 2015 : « cinq ans après, quel regard portent les musulmans sur les attentats ? » http ://path.com

(39) À quelques jours de son élection à la tête de la #FGTB, #ThierryBodson revient sur le blocage politique au Fédéral et fixe des lignes rouges à l'approche de la rentrée sociale. http ://path.com

(40) #Communication #digitale : le #slowcontent, une voie d'avenir ? Une réflexion ultra pertinente, amenant à penser une « #écologie éditoriale », moins mais mieux. Via @X @X http ://path.com

De même, les pictogrammes sont utilisés comme des *mots ou ponctuations classiques*, c'est-à-dire employés comme du lexique traditionnel (cf. exemples (41) à (43)) :

(41) La #Covid_19 va-t-elle bousculer les habitudes de #consommation en 🇫🇷 ? La moitié des Français ne sont pas prêts à réduire leur consommation, malgré la crise sanitaire selon @X. Les plus jeunes sont les plus motivés pour changer leurs habitudes http ://path.com

(42) #LeSaviezVous ❓ Pour 2020 et 2021, @X a décidé d'augmenter son enveloppe d'investissement de 4️⃣ 0️⃣ M € afin d'accompagner une relance économique forte et rapide. http ://path.com

(43) Sauf que rapidement il dévie sur des propos antisémites, négationnistes voire même islamistes. D'après lui, la shoah est un complot monté par les juifs pour renforcer leur "emprise sur le 🌍 ".

Nous avons constaté que cet usage des pictogrammes comme lexique sert particulièrement lors des procédés de discours rapportés. Les pictogrammes sont soit placés en début de tweets et opèrent comme des signaux graphiques qui annoncent un discours rapporté à suivre (exemples (44), (45) et (46)) ; soit utilisés comme des verbes de paroles (exemples (47) et (48)).

(44) 🗣️ 💬 Thierry Gomez « Si on joue à huis clos, ce n'est pas viable sur le plan économique. On se doit de trouver des solutions et passer par cette étape d'accueillir 5 000 personnes est un premier pas » OF & @X http ://path.com

(45) 🎙️ 💬 "Il a un vide à combler depuis qu'il n'est plus à la tête du RCT. Il n'a pas fait le deuil de ce passage exceptionnel de sa vie." 🙄 Denis Charvet s'inquiète pour la crédibilité de Mourad Boudjellal, qui continue de pousser pour un rachat de l'OM par Ajroudi. #rmclive http ://path.com

(46) ▶️ #Castex à #Lille. 🗣️ Pr @X : "Je ne sais pas s'il faut rendre le #MasqueObligatoire partout mais faire une recommandation, transformer cela en une règle sociale et l'imposer comme une règle administrative : pourquoi pas". 📺 #LaMatinaleLCI @X. http ://path.com

(47) #FranceRelance "On a besoin de sauver les entreprises (...) S'il y a un plan très conditionné, ça ne marchera pas" 🗣 @X, député de l'Oise, président de la commission des Finances de l'Assemblée ▶ #Les4V http ://path.com

(48) Antonio Conte 💬 : « Les rumeurs sur Leo Messi à l'Inter sont totalement fausses. Ne faites pas confiance à ces fake news. Il ne rejoindra pas l'Inter ! » #Inter #Conte #Messi http ://path.com

Enfin, les pictogrammes servent également à structurer les tweets : soit en mettant en avant des éléments de titre (exemples (49) et (50)), soit en organisant des énumérations/listes (exemples (51) et (52)).

(49) 🔴 Covid-19 / Gaza en danger 🔴 👉 L'enclave palestinienne est à son tour touchée par le Covid-19, plus de 400 cas recensés, ce qui laisse craindre le pire pour sa population qui subit déjà les pénuries et un manque criant d'accès aux soins. #Gaza #Palestine #Covid_19 http ://path.com

(50) 🚨 INFO MIDI OLYMPIQUE 🚨 Cette semaine, plusieurs réunions sont prévues en Biterre afin de savoir si les investisseurs des Émirats, qui viennent d'échouer dans leur projet de rachat de l'ASBH, pourraient à court terme en devenir les actionnaires. http ://path.com

(51) [PLAN DE RELANCE] Le Gouvernement a annoncé le plan #FranceRelance, un plan de long terme : 100 milliards pour préparer la France de 2030 autour de trois piliers : 1️⃣ Transition écologique 2️⃣ Souveraineté et compétitivité économique 3️⃣ Cohésion sociale et territoriale http ://path.com

(52) Notre première journée de travail s'articule autour de 3 grands axes : 📌 Enjeux du groupe pour la session 2020/2021 à l'@X 📌 Actualité et enjeux internationaux 📌 #FranceRelance #journéesparlementaires http ://path.com

Analyse des textes du registre Familier Enfin, le registre familier est utilisé pour dialoguer entre utilisateurs avec des marqueurs de l'oral qui se traduisent notamment en onomatopées (cf. éléments soulignés dans les exemples (53) à (57)) :

(53) @X C'est exactement ça la fachosphère, lr, lrem, rn s'est empressée de relayer la vidéo sans rien vérifier et bim c'est le clip d'un groupe de rap...

	Familier		Courant		Soutenu	
	#	%	#	%	#	%
Tweets qui se terminent par un ou des pictogramme(s)	6 227	8,92	7 582	5,65	304	1,25
Tweets qui commencent par un ou des pictogramme(s)	162	0,23	2 728	2,03	990	5,10

TABLE 3.6 – La table donne pour chaque registre le nombre de tweets soit se terminant, soit commençant par un ou des pictogrammes, avec le pourcentage qu'il représente dans l'ensemble du sous-corpus du registre.

(54) @X (t'as rt donc je prends ça pour un fav) jte mets avec @X pcq ça se voit à ton visage que t'es gentille et selena est gentille aussi donc bam quoi

(55) Lorsque tu penses que 2020 ne peut plus te surprendre et boum Leris nous annonce que Landorly est son fils 😨 http ://path.com

(56) @X @X @X @X rholala c'est vraiment grave a quelle point les commu manga/anime elle sont toxiques.

(57) RT @X : popopopoooow ce but des Lyonnaises

Dans ce contexte de discussion, les pictogrammes sont utilisés comme des modalisateurs, c'est-à-dire des moyens linguistiques à travers lesquels le locuteur exprime son point de vue par rapport à un contenu (cf. éléments soulignés dans les exemples (58) à (60)) :

(58) RT @XXX : Ils ont rajouté de la moquette et ils ont pris une photo un jour ensoleillé et c'est du Glow up jsuis explosé 😭 😭 😭

(59) C'est bien comme ça on catalogue ceux qui sont différents ou malade comme vous préférez 😂 😂 😂 Pfff toujours obligé de vous faire remarquer pour que vous soyez acceptés... #AlwaysProud

(60) lady gaga elle va faire deux performances du coup Rain On Me et un medley YESSSS ❤️ ❤️ ❤️ ❤️

La multiplication des items modalisateurs, que sont les pictogrammes, fait écho aux signes de ponctuation classiques qui sont également multipliés afin de marquer l'intensité de la modalisation. La table (3.6) détaille pour chaque registre le nombre de tweets soit se terminant, soit commençant, par un ou des pictogrammes, et précise le pourcentage qu'il représente dans l'ensemble du sous-corpus étiqueté automatiquement du registre associé.

Nous pouvons y constater qu'un trait linguistique saillant du registre familier semble être la position finale des pictogrammes remplaçant les signes de ponctuation traditionnels (tels que les ".", "?", "!" et "..."), puisqu'il est proportionnellement plus présent dans le sous-corpus familier que dans ceux des autres registres. Les exemples (58), (60), (61), (62) et (63) illustrent ce trait linguistique.

(61) @X Bizarre l'ambiance qd ca va remettre le trophée mvp a giannis après qu'il se soit fait 4-0 💀

(62) @X enfin quelqu'un d'honnête 😋

(63) Lena la best j'ai eu de merveilleux cadeau d'anniv 😋

Le phénomène inverse, c'est-à-dire des pictogrammes qui sont positionnés en tout début de tweet (comme les exemples (44), (45) et (46)), est observé : le registre soutenu en contient un pourcentage plus important que pour les autres registres.

Aussi, l'utilisation de la répétition contiguë des pictogrammes représente un second trait saillant du registre familier par rapport aux registres courant et soutenu. La figure 3.7 détaille les 10 séquences contiguës de pictogrammes les plus fréquentes par registre, en donnant leurs fréquences absolues associées aux pourcentages qu'elles représentent dans chaque registre. Seul le registre familier contient des séquences de plusieurs pictogrammes accolés. Remarquons également, qu'à l'instar des ponctuations classiques, les pictogrammes répétés sont les mêmes au sein d'une séquence.

Synthèse des analyses Les analyses linguistiques ont validé la pertinence de l'ensemble de descripteurs proposé pour l'analyse des CMO en registres, en montrant qu'ils permettent de les distinguer entre eux avec des descripteurs spécifiques à certains registres. Cette validation a été l'occasion de découvrir que des technomorphèmes tels que les pictogrammes ou les hashtags, ont été intégrés aux normes d'usage et grammaticale. En effet, si intuitivement l'utilisation des technomorphèmes est associée à un registre plus familier que soutenu, ces premières analyses ont montré au contraire que les technomorphèmes ont été intégrés à la norme et qu'ils ne sont plus uniquement caractéristiques du registre familier. Ils peuvent tout aussi bien marquer un discours soutenu. Plus précisément, l'intégration à la norme grammaticale de certains technomorphèmes a mis en exergue des usages différents selon les registres de langue. Pour le registre courant, les hashtags sont souvent utilisés afin de tirer partie de leurs fonctions d'indexation et renvoyer à d'autres sites que Twitter. De même, les pictogrammes servent à visuellement mettre en

	Familier	Freq abs	%	Courant	Freq abs	%	Soutenu	Freq abs	%
1	😭	1 403	8,62 %	➡️	1 105	2,45 %	👉	756	8,90 %
2	😂	1 049	6,44 %	👉	1 079	2,39 %	➡️	428	5,04 %
3	😭😭	456	2,80 %	😂	1 006	2,23 %	🇫🇷	307	3,61 %
4	💀	330	2,03 %	😉	895	1,98 %	🤔	201	2,37 %
5	😂😂	329	2,02 %	🇫🇷	781	1,73 %	🔴	193	2,27 %
6	😂😂😂	304	1,87 %	🤔	728	1,61 %	✅	176	2,07 %
7	🤣	301	1,85 %	😭	676	1,49 %	👇	171	2,01 %
8	😭😭😭	301	1,85 %	🔴	617	1,36 %	↪️	157	1,85 %
9	🥺	249	1,53 %	😍	569	1,26 %	😊	150	1,77 %
10	😅	212	1,30 %	😅	566	1,25 %	😉	110	1,29 %

FIGURE 3.7 – Top 10 des séquences contiguës de pictogrammes par registre, pour chacune d'elles nous précisons leurs fréquences absolues ainsi que le pourcentage qu'elles représentent au sein de leur registre.

avant un ou des lien(s) hypertextes vers d'autres contenus. Pour le registre soutenu, les pictogrammes sont utilisés comme des éléments lexicaux classiques et sont parfaitement intégrés syntaxiquement aux phrases ; ou comme éléments de ponctuation classiques pour palier à la linéarité des tweets en donnant une structure au texte avec la mise en avant de titres par exemple. Enfin, pour le registre familier, les pictogrammes sont majoritairement utilisés pour leur fonction de modalisateurs et opèrent comme des signes de ponctuation pouvant être répétés de manière contiguë pour illustrer l'intensité de la modalisation.

Ainsi, l'intégration à la norme des technomorphèmes par les registres courant et soutenu ramène les tweets vers un genre de textes plus traditionnel, respectant des conventions rédactionnelles, comme des brèves de presse par exemple. Cela illustre l'assimilation, par les locuteurs, de l'idée que les registres de langue français sont axés sur une norme linguistique associée à l'écrit. Au contraire, le registre familier présentant des formes plus orales n'intègre pas les technomorphèmes à la norme grammaticale, il les utilise en opérant des écarts à cette dernière.

3.5 Conclusion

Dans ce chapitre, nous avons présenté une technique d'apprentissage semi-supervisée dont l'objectif est d'étiqueter de manière réaliste des textes en registres de langue. Pour

cela, il étiquette automatiquement tous les textes d'un corpus en proportions de registres de langue en généralisant l'annotation manuelle d'une graine. Ce modèle a été validé par la constitution de deux corpus : TREMoLo-Web (section 3.3) et TREMoLo-Tweets (section 3.4). Ces travaux ont validé :

— la possibilité d'étiqueter les textes en représentant la variabilité des registres de langue ;

— la possibilité de neutraliser les biais liés aux genres de textes et aux modalités orale et écrite ;

— la pertinence de juger la fiabilité d'un texte : soit selon la confiance du classifieur dans sa prédiction illustrée par la valeur de la probabilité prédite, soit selon la cohérence de la prédiction en regardant la somme des probabilités prédites ;

— la pertinence de deux ensembles de descripteurs linguistiques : le premier pour l'analyse de genres de textes classiques issus de (Mekki, Battistelli, Lecorvé et al. 2018), le second pour l'analyse des CMO.

Le corpus TREMoLo-Web obtenu lors de travaux préliminaires présente deux principales limites l'empêchant d'être utilisé comme ressource textuelle pour caractériser automatiquement les registres de langue : la différence entre la taille de l'unité annotée, avec celle considérée lors de la caractérisation des registres, rend l'étiquetage incertain ; les descripteurs d'apprentissage peuvent biaiser l'extraction des motifs caractéristiques des registres en décidant en amont quels motifs permettent de distinguer les registres entre eux. C'est pourquoi nous avons proposé un second corpus, TREMoLo-Tweets, reprenant la même technique d'apprentissage semi-supervisée, mais la modifiant afin de répondre à ces limites. Pour cela, TREMoLo-Tweets est composé d'un seul genre de textes courts, des tweets ; et nous avons utilisé un classifieur pré-entraîné ne nécessitant pas le choix de descripteurs d'apprentissage. Les travaux ont conduit à plusieurs contributions : la rédaction d'un guide d'annotation pour les registres de langue français proposant un nouvel ensemble de descripteurs linguistiques pour l'analyse des CMO, la constitution d'un large corpus de tweets étiquetés en registres dont la qualité a été validée par des expériences ainsi que des analyses linguistiques. Cette exploration linguistique a mis au jour l'intégration à la norme grammaticale de certains technomorphèmes. Ces premiers résultats assurent la possibilité d'utiliser TREMoLo-Tweets comme ressource textuelle pour caractériser automatiquement les registres de langue. Nous présentons justement dans le chapitre suivant la fouille de motifs séquentiels émergents que nous avons choisie comme outil automatique afin de caractériser les registres à divers niveaux d'analyse de la langue.

Chapitre 4

La fouille de motifs comme outil automatique

Sommaire

En cherchant à caractériser les registres à partir d'un corpus de textes, nous cherchons à découvrir des régularités dans des données symboliques séquentielles (en l'occurrence, des phrases ou des textes). Le chapitre 3 a détaillé la constitution du corpus TREMoLo-Tweets illustrant les registres de langue avec un ensemble de tweets étiquetés en registre. TREMoLo-Tweets est divisé en sous-corpus, représentant chacun un registre, à partir desquels nous voulons comparer la présence de descripteurs linguistiques [1] afin de caractériser les registres. Pour illustrer la problématique, prenons l'exemple de deux corpus — l'un pour le registre familier, l'autre pour le registre soutenu — limités chacun à deux tweets, tel que l'illustre la table 4.1. Admettons alors que notre objectif est de découvrir automatiquement des descripteurs linguistiques, permettant de distinguer le registre soutenu du registre familier. Pour chercher ces descripteurs nous avons fait deux hypothèses. La première est de travailler au niveau des mots comme unité de parcours de la phrase. La seconde est de représenter chaque mot via un ensemble de niveaux d'analyse de la langue comme par exemple le lemme, la partie grammaticale du discours et les trois dernières

1. Nous rappelons qu'un descripteur linguistique est un patron linguistique décrivant une séquence de une ou plusieurs unités linguistiques, avec des traits linguistiques de divers niveaux d'analyse de la langue.

ID	Exemples du registre soutenu
1	Un joueur triplement décisif (1 but, 2 passes décisives), un grand Caleta-Car (héroïque défensivement, doublé) et un succès précieux malgré la fatigue après une préparation totalement tronquée. L'OM pouvait difficilement faire mieux. Brest valeureux et plaisant mais maladroit
2	Je ne doute pas du pouvoir résilient de tous les acteurs et actrices de cet incroyable microcosme qu'est l'école. Belle rentrée à toutes et tous (pensée spéciale pour mes collègues) #rentreescolaire2020

ID	Exemples du registre familier
3	Mdrrr en rajoute pas non plus la url_path
4	Mdrrr il vient de sous entendre que Ademo vend encore url_path

TABLE 4.1 – La table donne deux exemples de tweets pour le registre soutenu (ID 1 et 2) et familier (ID 3 et 4).

lettres (des exemples sont donnés table 4.2).

La question abordée dans ce chapitre est maintenant la suivante : quel type d'approche formelle est préférable de solliciter pour découvrir automatiquement ces descripteurs caractéristiques des registres ? Comme nous supposons que les registres touchent tous les niveaux d'analyse de la langue sans prédominance et que leur croisement rend l'interprétation des descripteurs plus robustes, ces approches formelles doivent déterminer si un descripteur distingue un registre A cible d'un registre B source en observant trois contraintes :

1. découvrir des descripteurs pouvant être composés de divers niveaux d'analyse de la langue ;
2. ne pas poser d'*a priori* linguistiques sur les descripteurs à extraire, par exemple ne pas juger le niveau lexical plus important que le niveau morphologique ;
3. fouiller un large ensemble de données pour obtenir des descripteurs représentatifs de l'usage réel des registres.

Parmi les approches proposées dans la littérature scientifique, les techniques de fouille de motifs (AGRAWAL, SRIKANT et al. 1994) semblent pertinentes puisqu'elles permettent d'extraire des « motifs langagiers », correspondant à ce que nous appelons « descripteurs linguistiques », à partir d'un large corpus de textes représentatifs des registres de langue

ID	Exemples du registre soutenu
1	$\begin{pmatrix} un \\ det \\ -un \end{pmatrix} \begin{pmatrix} joueur \\ nom\ commun \\ -eur \end{pmatrix} \begin{pmatrix} triplement \\ adverbe \\ -ent \end{pmatrix} ...$
2	$\begin{pmatrix} je \\ pronom\ personnel \\ -je \end{pmatrix} \begin{pmatrix} ne \\ adverbe \\ -ne \end{pmatrix} \begin{pmatrix} douter \\ verbe \\ -ter \end{pmatrix} ...$

ID	Exemples du registre familier
3	$\begin{pmatrix} mdr \\ adverbe \\ -rrr \end{pmatrix} \begin{pmatrix} en \\ préposition \\ -en \end{pmatrix} \begin{pmatrix} rajouter \\ verbe \\ -ter \end{pmatrix} ...$
4	$\begin{pmatrix} mdr \\ adverbe \\ -rrr \end{pmatrix} \begin{pmatrix} il \\ pronom\ personnel \\ -il \end{pmatrix} \begin{pmatrix} venir \\ verbe \\ -nir \end{pmatrix} ...$

TABLE 4.2 – La table donne deux extraits de tweets dont chaque mot est représenté par son lemme, sa partie grammaticale du discours et ses trois dernières lettres pour le registre soutenu (ID 1 et 2) et familier (ID 3 et 4).

française. Historiquement, ces techniques de fouille de motifs ont été proposées pour explorer des bases de données de transactions de clients, afin de repérer des ensembles d'objets fréquemment achetés ensemble. Ces ensembles sont appelés des *motifs fréquents*. Ensuite, pour répondre à d'autres cas d'usage, d'autres techniques de fouilles de motifs ont été introduites en proposant divers types d'ensembles de motifs et différentes manières de les filtrer en sous-ensembles. Parmi elles, la fouille de motifs séquentiels émergents (Guozhu DONG et J. LI 1999) est particulièrement adaptée à notre objectif, dès lors qu'elle répond aux trois contraintes posées ci-dessus :

1. la formalisation des motifs permet une représentation multi-factorielle des données en les décrivant via différentes propriétés (en l'occurrence des traits linguistiques) ;
2. selon les algorithmes employés il est possible de ne pas poser d'*a priori* linguistiques quant aux motifs à découvrir ;
3. les ressources textuelles considérées peuvent être volumineuses.

Par exemple, à partir de la représentation des données illustrée table 4.2, la fouille de motifs séquentiels émergents cherche des descripteurs linguistiques caractéristiques du registre soutenu par rapport au registre familier et inversement. Par exemple, pour le registre soutenu qui semble marqué par l'emploi d'adverbe se terminant en *–ent* et l'utilisation d'adjectifs qualificatifs, les descripteurs linguistiques caractéristiques sont formalisés par

les motifs M_1=⟨{pos :adverbe, 3der :*–ent*}⟩ et M_2=⟨{pos :nom commun, pos :adjectif}⟩. Ces deux exemples de motifs séquentiels montrent :

- la combinaison de traits relevant de différents niveaux d'analyse de la langue avec M_1 ;
- la possibilité d'avoir des motifs multi-mots avec M_1 et M_2 ;
- l'importance des choix des niveaux de langue, par exemple le choix des trois dernières lettres n'est pas pertinent, il aurait été meilleur avec les quatre dernières lettres pour capturer la terminaison *–ement* plus précise ;
- la question de la contiguïté des mots avec M_2 qui capture des passages si un écart de 1 est autorisé.

L'avantage de la fouille de motifs séquentiels émergents est de découvrir des descripteurs linguistiques multi-mots inconnus caractéristiques des registres de langue en spécifiant avec quels traits nous voulons décrire ces mots et quel écart nous autorisons entre eux. Si les techniques de fouille de motifs séquentiels émergents présentent des avantages pour notre tâche, elles présentent également certains désavantages. Ces désavantages ne sont pas spécifiques à la tâche de caractérisation des registres de langue, mais ils sont communs à toutes les approches de fouille de motifs. Le calcul des motifs fréquents peut rapidement avoir un coût algorithmique élevé. De plus, le nombre élevé des motifs découverts affecte la qualité des résultats : les motifs sont trop nombreux et redondants entre eux pour être compréhensibles. C'est pourquoi, toutes les techniques de fouille de motifs tendent à réduire la complexité algorithmique, et mettent en place des filtres pour limiter la quantité résultats. Dans le cadre de notre travail, nous n'avons pas visé une contribution fondamentale dans le domaine des algorithmes d'extraction de motifs en proposant une nouvelle approche. Nous avons cherché, parmi les approches existantes, un outil automatique adapté à notre objectif de caractérisation des registres permettant d'obtenir un ensemble de motifs linguistiques exploitables.

Dans ce chapitre, en dressant un panorama des différentes approches proposées dans la littérature scientifique, nous motivons notre choix des MSE pour la caractérisation des registres de langue. La première section présente les notions de base issues des travaux à l'origine de la fouille de motifs et introduit les premiers algorithmes proposés pour découvrir des motifs dits « ensemblistes ». Nous y verrons, également, les principaux types de sous-ensembles de motifs ensemblistes qui sont repris et adaptés pour filtrer d'autres types de motifs, tels que les motifs dit « séquentiels » qui nous intéresseront particulièrement et seront introduits section 4.2. Enfin, dans la section 4.3, nous proposons

de détailler plusieurs applications des techniques de fouilles de motifs ensemblistes et séquentiels pour montrer leurs robustesses face à des corpus de données réelles.

4.1 Fouille de motifs ensemblistes

L'approche historique de fouille de motifs, dont découle celle que nous utiliserons mais aussi de nombreuses autres, s'intéresse à la recherche de motifs dit *ensemblistes* (ME). L'exemple typique est la recherche des produits fréquemment présents en association dans les paniers de courses des consommateurs. Un motif ensembliste est alors un sous-ensemble de l'ensemble des produits d'un panier. Comme l'énumération de tous les sous-ensembles possibles d'un ensemble est d'une combinatoire exponentielle, différentes stratégies ou étapes gravitent autour de la problématique de fouille, comme la figure 4.1 en donne un exemple. Parmi celles-ci, le critère de fréquence et particulièrement utile pour réduire l'espace des possibilités. On parle alors de ME fréquents dès lors qu'ils dépassent une certaine fréquence absolue ou relative. D'autres filtrages peuvent encore venir restreindre la fouille de ME.

Ces techniques de fouille de ME dégagent des informations permettant de mieux comprendre des données, dans le but de prendre des décisions à partir des connaissances extraites. Appliquées à notre cas d'étude, la fouille de ME a pour principale limite de ne pas prendre en compte l'ordre des mots entre eux puisqu'elle modélise chaque phrase ou chaque texte, selon l'unité choisie, comme des ensembles de mots. Cependant, d'importantes notions utilisées dans nos travaux, telles que celle de la fréquence, de la clôture ou de l'émergence viennent de la fouille de ME. C'est pourquoi, nous présentons dans la première partie de cette section le cadre théorique guidant la fouille de ME fréquents. Nous présentons ensuite les principaux algorithmes proposés dans la littérature partant de ce cadre formel pour extraire l'ensemble des ME fréquents.

4.1.1 Cadre théorique

Cette section pose le cadre théorique déterminant les trois étapes principales de la fouille de ME. La première transforme les données pour les convertir en ME si les données ne sont pas déjà des ensembles ; la deuxième explore les ME pour découvrir les ME fréquents ; enfin, la troisième filtre ces derniers selon différentes approches pour en réduire le nombre et la redondance.

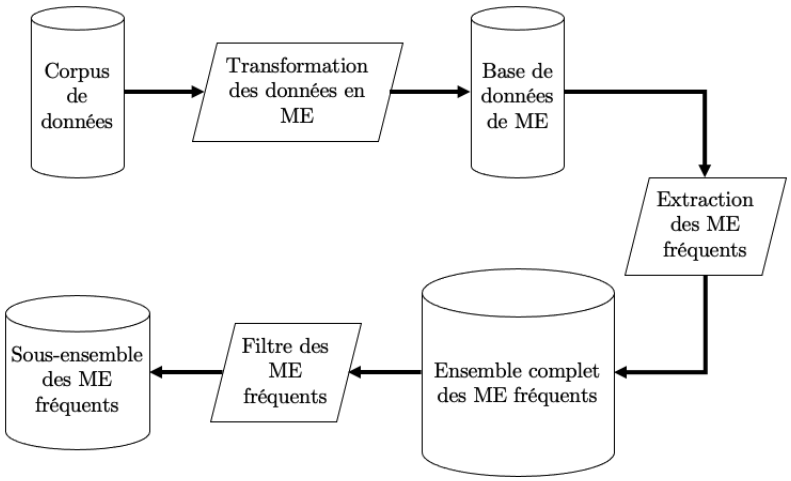

FIGURE 4.1 – Chaîne de traitement pour l'extraction de l'ensemble complet de ME fréquents, et de sa réduction à un sous-ensemble de ME fréquents.

tid	**Itemset**
1	{*Mdrrr, en, rajoute, pas, non, plus, la, url_path*}
2	{*Mdrrr, il, vient, de, sous, entendre, que, Ademo, vend, encore, url_path*}

TABLE 4.3 – Exemple de base de données, notée D, avec deux tuples associant *tid* et itemsets, pour représenter le sous-corpus du registre familier de la table 4.1 contenant deux textes.

Transformation des données en motifs ensemblistes La fouille de ME s'appuie sur le concept-clé d'itemset. Un itemset est défini comme un ensemble de littéraux appelés items $I = \{i_1, ..., i_k\}$, par exemple l'ensemble des produits qui composent un panier. Un itemset peut être également appelé $|k|$-*itemset* si I contient $|k|$ *items*. La notion d'itemset sert à représenter tant les transactions de la base de données sur laquelle opérer la fouille que les ME qui émergent de cette fouille. Chaque itemset est mémorisé dans une base de données grâce à un tuple $T = (tid, I)$, où un identifiant unique *tid* est associé à l'itemset I. S'appuyant sur la notion d'ensemble, l'espace des itemsets dispose de la relation d'inclusion, notée $\subseteq$. En particulier, si un itemset X est inclus dans une transaction T ($X \subseteq T$), on dit que « T support X ». Par exemple, $X = \{a, c\}$ et $T = \{a, b, c\}$.

Appliqué à notre cas d'usage, cette modélisation représente le sous-corpus du registre familier de la table 4.1 en une base de données notée D. La table 4.3 montre D, où chaque texte du sous-corpus est représenté par un tuple T contenant un *tid* et un itemset I, dans lequel chaque mot du texte est illustré par un item i. Par exemple, le tweet 3 de la table

4.1 contenant 8 mots y est représenté par le tuple $T_1 = (1, I_1)$, où I_1=$\{$*Mdrrr, en, rajoute, pas, non, plus, la, url_path*$\}$ contient 8 items. T_1 supporte l'itemset $\{$*en, rajoute, pas*$\}$, car $\{$*en, rajoute, pas*$\} \subseteq I_1$.

Extraction des motifs ensemblistes fréquents Après avoir converti un corpus en une base de données de transactions, comment en extraire les ME fréquents ? Pour cela, il faut définir la manière dont la fréquence d'un ME, noté X, est calculée dans une base de données D. Nous détaillons ici les trois étapes nécessaires à son calcul. La première étape est de mesurer la *couverture* de X dans la base de données D, notée $couverture(X, D)$.

$$couverture(X, D) := \{tid \mid (tid, I) \in D, X \subseteq I\} \tag{4.1}$$

La $couverture(X, D)$, définie par l'équation 4.1, représente l'ensemble des *tid* dont les tuples supportent X. Par exemple, la couverture de $\{$*url_path*$\}$, noté X_1, est l'ensemble {1,2} car X_1 est supporté par les tuples T_1 et T_2. La couverture de $\{$*Ademo*$\}$, noté X_2, est l'ensemble {2} car seul le tuple T_2 supporte X_2. La deuxième étape est le calcul du *support* de X dans D, noté $support(X, D)$, à partir de la couverture de X.

$$support(X, D) := Card(couverture(X, D)) \tag{4.2}$$

Comme défini par l'équation 4.2, le support de X est le nombre de *tid* contenus dans la couverture de X dans D, autrement dit le nombre de tuples supportant X dans D. Par exemple, le support de X_1 est $support(X_1, D) = 2$, car 2 tuples supportent X_1. Enfin, la troisième étape est le calcul de la fréquence de X, notée $frequence(X, D)$, à partir du support de X.

$$frequence(X, D) := \frac{support(X, D)}{Card(D)} \tag{4.3}$$

L'équation 4.3 estime que la fréquence de X dans D représente la probabilité d'occurrence de X dans un tuple $T \in D$. La fréquence de X_1 est de 100% car tous les tuples de D supportent X_1, tandis que la fréquence de X_2 est de 50% car un tuple sur deux supporte X_2. Un ME est dit « fréquent », lorsque sa fréquence est supérieure ou égale à un seuil fixé par l'utilisateur. Ce seuil, noté *minsup*, est une fréquence minimale. Ainsi, si nous considérons un $minsup = 50\%$, alors X_1 et X_2 sont tous deux des ME fréquents. En revanche, si nous considérons un $minsup = 75\%$, alors seulement X_1 est un ME fréquent.

Filtre des motifs ensemblistes fréquents Plusieurs manières de filtrer l'ensemble des ME fréquents sont proposées dans la littérature. Elles permettent d'obtenir des sous-ensemble de divers types de ME. Parmi eux, nous en avons retenu trois : les ME maximaux, les ME clos et les ME émergents. Nous les avons sélectionnés car ils ont été généralisés au delà des ME, notamment aux motifs séquentiels que nous utiliserons pour caractériser les registres. Les deux premiers sous-ensembles ont pour objectif de réduire les ME fréquents d'une même base de données en les condensant entre eux. Les ME maximaux de D sont des ME fréquents qui ne sont inclus dans aucun autre ME fréquents de D. Autrement dit, un ME dans D, noté X, est dit maximal si tous ses sur-ensembles sont non fréquents et tous ses sous-ensembles sont fréquents dans D. Par exemple, le ME fréquent $\{url\_path\}$ n'est pas maximal, car il existe le ME fréquent $\{Mdrrr,\ il,\ vient,\ de,\ sous,\ entendre,\ que,\ Ademo,\ vend,\ encore,\ url\_path\}$ qui est son sur-ensemble. Cette approche a pour principale limite de perdre des informations sur les fréquences des ME compris dans un ME maximal. Nous savons seulement que leurs fréquences sont supérieures ou égales au *minsup*. Les ME clos répondent à cette limite en conservant uniquement les ME fréquents dont aucun sur-ensemble n'a la même fréquence. En d'autres termes, un ME fréquent X est dit clos s'il n'existe aucun ME fréquent X' dans D tel que $X \subseteq X'$ et $frequence(X, D) = frequence(X', D)$. Dès lors, nous pouvons en déduire que tous les ME compris dans un ME clos ont la même fréquence que lui : les ME clos condensent les ME fréquents sans perte d'information. Par exemple, le ME fréquent $\{Mdrrr\}$ noté X_3 n'est pas clos, car il existe le ME fréquent $\{Mdrrr,\ en,\ rajoute,\ pas,\ non,\ plus,\ la,\ url\_path\}$ qui est son sur-ensemble et qui a la même fréquence que lui.

Le troisième sous-ensemble, les ME émergents, a deux objectifs combinés : réduire les ME fréquents de D et ne conserver que les ME fréquents caractéristiques de D, c'est-à-dire distinguant D d'une autre base de données. C'est pourquoi les ME émergents considèrent deux bases de données, D et D', à partir desquelles ils comparent les fréquences des ME fréquents : seuls les ME fréquents de D, dont les fréquences augmentent de manière significative de D à D', sont conservés. Cette augmentation se mesure via le *taux de croissance* défini comme suit :

$$TauxCroissance(X_{D|D'}) = \begin{cases} \infty, & \text{si} frequence(X, D') = 0 \\ \frac{frequence(X,D)}{frequence(X,D')}, & \text{sinon} \end{cases} \tag{4.4}$$

Un ME fréquent X est dit émergent de D par rapport à D' si son taux de croissance est strictement supérieur à un seuil fixé par l'utilisateur, noté ρ. Par exemple, si nous

fixons $\rho = 1$ et considérons D avec une seconde base de données D' représentant le sous-corpus du registre soutenu (de la table 4.1) ; alors le ME fréquent X_3 est émergent puisque son taux de croissance est supérieur à ρ avec $TauxCroissance(X_{3_{D|D'}}) = \frac{100}{0} = \infty$. Cette notion d'émergence, mesurée via le taux de croissance, répond à notre besoin de distinguer des descripteurs linguistiques d'un registre cible par rapport à un registre source. C'est pourquoi nous avons utilisé ce type de sous-ensemble dans nos travaux. Cependant, comme précisé dans l'introduction, nous l'avons appliqué aux motifs séquentiels, qui présentent l'avantage de considérer l'ordre entre les mots, plutôt qu'ensemblistes.

4.1.2 Algorithmes de fouille de motifs ensemblistes

Nous présentons les algorithmes de fouille de ME car les problématiques algorithmiques et les notions utilisées se retrouvent dans les algorithmes de fouille de motifs séquentiels.

Pour découvrir des ME fréquents, l'approche générale génère et élague des motifs candidats aussi appelés itemsets candidats. Un motif candidat est généré puis testé pour voir s'il est fréquent ou non, s'il est non fréquent il est élagué de l'espace de recherche. Un motif candidat est un motif créé en joignant deux motifs. Par exemple, $S_3 = \{a, b, c\}$ est un motif candidat résultant de la jointure de $S_1 = \{a\}$ et $S_2 = \{b, c\}$. Le coût algorithmique de cette étape de génération et test de candidats est exponentiel, car dépendant de la taille de la base de données considérée et du nombre d'items qu'elle contient : pour p items, 2^p itemsets candidats doivent être générés. Par exemple, considérant $I = \{a, b, c\}$ contenant 3 items, 8 itemsets candidats sont générés pour calculer leur fréquence et découvrir s'ils sont fréquents ou non : $\{a\}$, $\{b\}$, $\{c\}$, $\{a, b\}$, $\{a, c\}$, $\{b, a\}$, $\{b, c\}$, $\{a, b, c\}$. Afin de limiter ce coût algorithmique, les techniques de fouille de ME tirent partie de trois éléments (1) le choix des structures de données stockant les candidats lors de la génération pour structurer l'espace de recherche (2) l'élagage et (3) le parcours de ces structures de données pour réduire l'espace de recherche. Nous dressons dans cette section un panorama des algorithmes proposés dans la littérature. Comme nous n'avons pas utilisé les ME dans nos travaux, ce panorama n'a pas pour but d'être exhaustif mais de présenter différentes stratégies visant à restreindre la complexité algorithmique. Nous présentons à la suite de cette introduction plusieurs algorithmes découvrant les ME fréquents, avant d'introduire ceux qui en dérivent des sous-ensembles.

Algorithmes pour l'extraction de ME fréquents Le premier algorithme que nous présentons est celui introduit dans les années 1990 par l'article fondateur (Agrawal,

Srikant et al. 1994) : l'algorithme *Apriori*. Son objectif est de découvrir les ME fréquents d'une base de données D en réduisant le nombre d'itemsets candidats à générer. Pour cela, *Apriori* génère uniquement les candidats pour lesquels tous les sous-ensembles sont fréquents en se reposant sur la propriété antimonotone de la fréquence d'un itemset : tous les sur-ensembles d'un itemset non-fréquent sont non-fréquents ; tous les sous-ensembles d'un itemset fréquent sont fréquents. *Apriori* procède itérativement en suivant l'ordre croissant de la longueur des itemsets candidats [2], notée k, en partant de $k = 1$:

1. Génération des itemsets candidats de longueur k.
2. Test des candidats en calculant leur fréquence.
3. Jointure des itemsets candidats fréquents deux par deux.
4. Élagage des itemsets candidats non-fréquents.
5. Initiation de l'itération suivante en passant aux itemsets candidats fréquents de longueur $k = k + 1$.

La principale limite de cette approche est la génération d'itemsets candidats non-fréquents. Pour réduire le coût du calcul de cette étape, il est souvent nécessaire de fixer une valeur élevée de *minsup*. Cela a pour conséquence de découvrir uniquement des ME dont la fréquence est élevée et de passer à côté d'autres ME intéressants malgré une fréquence moins élevée. Pour répondre à cette limite, (J. Han, Pei, Yin et Mao 2004) propose un algorithme évitant la génération ainsi que le test de candidats : *FP Growth*. Il représente une base de données D avec un type de structure de données réduisant le coût de l'exploration des ME fréquents en compressant l'information : les arbres de hachage. Chaque nœud représente un ensemble d'items permettant ainsi aux arbres de hachage de condenser D en stockant les motifs dans ses feuilles. La figure 4.2, extraite de (Masseglia, Teisseire et Poncelet 2004), donne un exemple d'arbre de hachage. Il stocke en mémoire cinq séquences dans ses trois feuilles. Les séquences contiennent quatre items, A, B, C et D, respectivement représentés par un nœud et trois feuilles de l'arbre. Si un arbre de hachage permet de condenser l'information, sa construction est une opération coûteuse pouvant ne pas aboutir lorsque D est volumineuse. C'est pourquoi, d'autres types de structures de données ont été proposées. Citons, par exemple, (Mohammed Javeed Zaki 2000) qui présente un nouveau format de base de données, appelé *tid-liste*, où chaque itemset est associé à une liste de tuples dans laquelle il apparaît. Dès lors, tous les itemsets fréquents peuvent être énumérés grâce à des intersections de la *tid-liste*. Mentionnons également

2. La longueur d'un itemset est le nombre d'items qu'il contient.

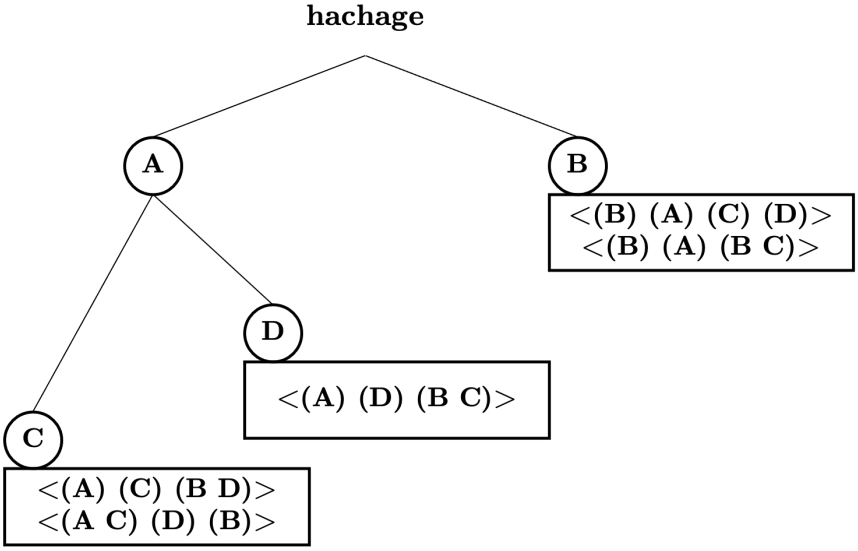

FIGURE 4.2 – Exemple d'un arbre de hachage composé d'un nœud et trois feuilles. Il stocke cinq séquences composées de quatre items : A, B, C et D. L'arbre est extrait de (MASSEGLIA, TEISSEIRE et PONCELET 2004).

(PEI, J. HAN, H. LU et al. 2001) qui propose une structure de données notée *H-struct*. *H-struct* permet à partir de la liste ordonnée des ME de projeter les ME fréquents afin de condenser l'information. Les données au format *H-struct* permettent une extraction plus rapide en réduisant le nombre de balayages de la base de données. Ces quelques exemples montrent que pour remédier au coût de calcul exponentiel de la découverte des ME fréquents, les approches utilisent des structures de données cherchant à compresser D pour limiter la complexité algorithmique liée à ses balayages.

Algorithmes pour l'extraction de ME maximaux et clos En précisant le type de ME fréquents que l'on cherche à découvrir, la fouille élague plus efficacement l'espace de recherche et réduit son coût algorithmique. Par exemple, l'algorithme *MaxMiner* proposé par (BAYARDO JR 1998) recherche des ME maximaux en limitant le nombre de chemins empruntés dans l'arbre et réduit ainsi l'espace de recherche considéré. L'algorithme proposé par (UNO, T. ASAI et al. 2003), quant à lui, recherche les ME clos en utilisant des chemins transversaux composés uniquement de motifs clos à travers les arbres de hachage. Ces algorithmes présentent donc le double avantage de réduire la complexité algorithmique lors de la fouille de ME maximaux ou clos ainsi que la quantité de ME extraits rendant les résultats plus interprétables.

Algorithmes pour l'extraction de ME émergents Enfin, nous présentons un dernier type de sous-ensemble introduit par (Guozhu Dong et J. Li 1999) : les ME émergents. Comme détaillé section 4.1.1, les ME émergents filtrent les ME fréquents en comparant une base de données D représentant une classe cible, à une base de données D' représentant une classe source. Seuls les ME caractéristiques de D par rapport à D' sont conservés. Ce sous-ensemble nous intéresse puisqu'il permet de découvrir des motifs distinguant une classe cible d'une classe source (en l'occurrence un registre cible d'un registre source). Un type de ME émergent est particulièrement intéressant : les ***J**umping **E**mergent **P**atterns* (JEP). Les JEP sont les ME émergents les plus caractéristiques de la classe cible : leur fréquence est non nulle dans D, mais est égale à 0 dans D'. Suivant la définition du taux de croissance, donnée équation 4.4, le taux de croissance d'un JEP tend vers ∞. Appliqué à notre cas d'usage, un JEP nous informe qu'une forme linguistique est très spécifique au registre cible par rapport au registre source, puisque présente dans les textes du premier et absente de ceux du second. Cependant, cette valeur de ∞ peut être problématique lors de l'interprétation des JEP puisqu'ils deviennent tous égaux quelques soient leurs fréquences dans le registre cible. Un JEP ayant une fréquence de 80% dans le registre cible aura le même taux de croissance de ∞ qu'un JEP avec une fréquence de 10% dans le registre cible si les deux sont non-fréquents dans le registre source. Or, un JEP avec une fréquence plus haute dans le registre cible n'est il pas plus caractéristique qu'un JEP avec une fréquence inférieure ? Les JEP sont donc intéressants mais peu aisés à interpréter dans notre cas d'étude. Suivant cette idée que les JEP sont particulièrement intéressants pour l'utilisateur, certains algorithmes les utilisent pour filtrer les ME à extraire. Par exemple, (Fan et Ramamohanarao 2003) les utilise pour réduire le temps de recherche. Il propose l'algorithme *iEPMiner* (pour ***I**nteresting **E**merging **P**attern **Miner***), qui sélectionne uniquement les JEP en les filtrant selon l'intérêt qu'ils représentent pour l'utilisateur, en se basant sur des termes objectifs. Un JEP est intéressant si ce dernier a une fréquence minimale et s'il a un taux de croissance minimal ainsi qu'un taux de croissance supérieur à celui de son sous-ensemble et enfin qu'il est fortement corrélé selon des mesures statistiques courantes telles que la valeur du χ^2. D'autres algorithmes de fouille de ME émergents sont proposés dans la littérature. Cependant, nous ne les détaillons pas ici car nous ne les avons pas utilisés dans nos travaux ; et les sections précédentes ont déjà évoqué les principales stratégies mises en place, pour la réduction du coût algorithmique de la fouille de ME fréquents, maximaux et clos, qui s'appliquent également aux motifs émergents.

4.1.3 Synthèse

Dans cette section, nous avons présenté formellement les ME en introduisant des notions centrales dans notre travail telles que la fréquence, la clôture et l'émergence. Puis, nous avons décrit certains algorithmiques fouillant les ME fréquents, maximaux, clos et émergents, dans le but de montrer comment les algorithmes répondent à leur coût algorithmique exponentiel et à la quantité de ME fréquents extraits. Pour répondre à la première limite, les travaux dans la littérature proposent des structures de données particulières qui compressent les bases de données pour en limiter les balayages lors de la fouille. Certains algorithmes réduisent également l'espace de recherche en parcourant les structures de manière partielle. Quant à la seconde limite, les sous-ensembles de certains types de ME condensent l'ensemble de ME fréquents selon des critères de fréquence et d'inclusion. Les ME clos et émergents ont semblé particulièrement intéressants pour notre objectif de caractérisation de registres de langue : les premiers condensent les ME sans perte d'information ; les seconds distinguent des ME caractéristiques d'un ensemble de données par rapport à un autre.

Cependant, les ME ont pour principale limite de ne pas garder la notion d'ordre entre les objets représentés. Or, cette notion est importante pour nos travaux puisqu'ils analysent des textes dont l'ordre des mots peut en changer le sens. La section suivante introduit la fouille de motifs séquentiels, qui, justement, considèrent la notion d'ordre et nous serviront dans les chapitres suivants.

4.2 Fouille de motifs séquentiels

Les **M**otifs **S**équentiels (noté MS) sont un type de motifs dérivé des ME. Les MS présentent deux avantages notables par rapport aux ME : donner une représentation multifactorielle des données avec des itemsets où chaque item renvoie à un facteur différent ; rendre compte de l'ordre des objets représentés avec des séquences d'itemsets. Les MS sont adaptés à notre tâche visant à caractériser les registres de langue à partir d'un corpus de textes puisqu'ils répondent à nos contraintes (cf. introduction) en représentant chaque mot avec un ensemble de traits linguistiques (comme le lemme, les fonctions grammaticale et syntaxique, etc.) et en conservant la notion d'ordre entre les mots.

De manière générique, l'objectif de la fouille de MS est de découvrir des sous-séquences intéressantes dans un ensemble de séquences. L'intérêt d'une sous-séquence peut être mesuré en fonction de différents critères : sa fréquence, sa longueur, son contenu, etc. Pour

cette tâche, l'approche est similaire à celle mise en place pour la fouille de ME. Un corpus de données est transformé en MS stockés dans une base de données D. Des algorithmes cherchent l'ensemble complet de MS intéressants à partir de D selon les critères fixés par l'utilisateur. Pour notre étude, un MS intéressant est un MS caractéristique d'une D par rapport à une seconde base de données. C'est pourquoi nous avons utilisé la fouille de motifs séquentiels émergents (MSE), à partir d'un large corpus de textes divisés en trois sous-corpus représentant les registres familier, courant et soutenu.

Cependant, si pour notre objectif ces techniques de fouilles de MSE présentent des avantages certains par rapport aux ME, elles présentent également les mêmes désavantages que pour la fouille de ME, mais accentués par la complexité supplémentaire introduite avec les séquences d'itemsets. Cette section a pour but de montrer pourquoi les MSE sont les plus adaptés à notre étude et comment les techniques de fouille limitent le coût algorithmique lié à leur découverte rendant leur utilisation possible pour une grande D. Comme dans la section précédente, nous introduisons tout d'abord le cadre théorique à partir duquel les algorithmes de fouille, présentés à la suite du cadre théorique, se basent pour découvrir les MS fréquents. Enfin, nous terminons en détaillant des approches filtrant les MS fréquents en divers sous-ensembles.

4.2.1 Cadre théorique

Le cadre théorique guidant la fouille de MS est dérivé de celui des ME. Beaucoup de notions comme celles de la fréquence, de la clôture ou bien de l'émergence sont les mêmes. Le principal changement est la formalisation d'un motif. Par conséquent, nous commençons par décrire formellement les MS. Puis, nous présentons deux manières de les représenter soit dans une base de données, soit dans un arbre de préfixes. Divers filtres pour obtenir des sous-ensembles de MS fréquents sont présentés en fin de section, ils viennent s'ajouter à ceux filtrant les ME fréquents présentés section 4.1.1 en s'adaptant aux MS.

Formalisation d'un MS Ce que nous appelons un MS est soit une séquence complète, soit un sous-ensemble d'une séquence appelé sous-séquence. En pratique, le terme MS est employé pour désigner une sous-séquence, tandis que le terme de séquence est employé pour désigner une séquence stockée dans une base de données séquentielles.

Definition 4.2.1 (Séquence)**.** Une séquence S est une liste ordonnée d'itemsets notée

$S = \langle I_1 \dots I_m \rangle$. Par exemple, la séquence $S = \langle \{a, b, c\}\{a, d\}\{a, b\} \rangle$ est une séquence de trois itemsets, chacun composé respectivement de trois, deux et deux items.

Definition 4.2.2 (Sous-séquence)**.** Une séquence $S_1 = \langle I_1, I_2, ...I_n \rangle$ est une sous-séquence de $S_2 = \langle I'_1, I'_2, ...I'_m \rangle$ s'il existe des entiers $1 \leq j_1 < \dots < j_n \leq m$ tels que $I_1 \subseteq I'_{j_1}, \dots, I_n \subseteq I'_{j_n}$. La notation est $S_1 \leqslant S_2$. Par exemple, $S_1 = \langle \{a\}\{d\} \rangle$ est une sous-séquence de $S_2 = \langle \{a, b, c\}\{c, d\}\{a, d\} \rangle$ car $\{a\} \subseteq \{a, b, c\}$ et $\{d\} \subseteq \{c, d\}$, alors $S_1 \leqslant S_2$ dès lors que $\langle \{a\}\{d\} \rangle \leqslant \langle \{a, b, c\}\{c, d\}\{a, d\} \rangle$.

Par exemple, suivant la définition 4.2.1 d'une séquence et considérant deux traits linguistiques, le mot et la fonction grammaticale, la phrase « Il chante bien ! » de quatre mots[3] est transformée en une séquence S de quatre itemsets composés de deux items : $\langle$(*mot:Il, pos:pronom personnel*) (*mot:chante, pos:verbe*) (*mot:bien, pos:adverbe*) (*mot: !, pos:ponctuation forte*)$\rangle$. Suivant la définition 4.2.2, la séquence $S' = \langle$(*mot:Il*) (*pos:verbe*)$\rangle$ est une sous-séquence de S.

Structures de données pour le stockage des MS Comme pour les ME, chaque séquence est stockée dans une base de données D sous la forme d'un tuple $T = (tid, S)$. La table 4.4 montre D, où chaque texte du sous-corpus familier de la table 4.1 est représenté par un tuple T. La première colonne de la table donne la valeur de *tid*, la seconde colonne donne une séquence S représentant le texte. Chaque mot est représenté par un itemset et chaque item représente un trait linguistique (en l'occurrence le mot et sa fonction grammaticale). En comparant la table de ME 4.3 avec la table de MS 4.4, nous voyons l'intérêt des MS qui donnent une représentation plus riche des données textuelles. Dans l'absolu, c'est-à-dire sans considérer les limites liées à la complexité algorithmique, il n'y a pas de limite au nombre d'items contenus dans un itemset : nous pourrions représenter chaque mot par autant de traits linguistiques que souhaités.

Definition 4.2.3 (Ordre lexicographique)**.** L'ordre lexicographique, noté $\succ_{lex}$, est défini comme tout ordre total de I. Cet ordre global peut être soit l'ordre alphabétique, soit l'ordre croissant. On suppose dans ce qui suit que tous les itemsets sont ordonnés en fonction de $\succ_{lex}$.

Definition 4.2.4 (Préfixe)**.** Une séquence $S_1 = \langle I_1, I_2, ..., I_n \rangle$ est un préfixe de la séquence $S_2 = \langle I'_1, I'_2, ..., I'_m \rangle$, $\forall n < m$, si et seulement si $I_1 = I'_1, I_2 = I'_2, ..., I_{n-1} = I'_{n-1}$ et les $|I_n|$ premiers items de I'_n selon $\succ_{lex}$ sont égaux à ceux de I_n

3. Les signes de ponctuation sont considérés comme des mots.

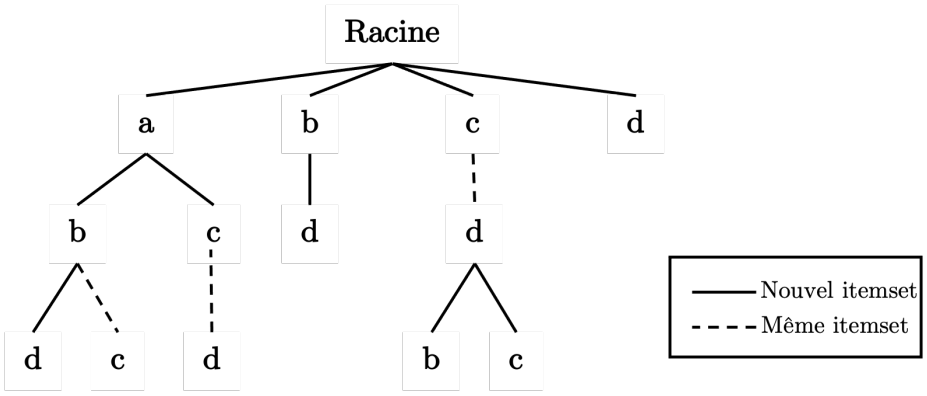

FIGURE 4.3 – Arbre de préfixes, noté A_{D_1}, représentant la base de données D_1 de cinq séquences donnée table 4.5.

À partir de D, nous pouvons construire un arbre de préfixes afin de compresser les informations contenus dans D. Le préfixe d'une séquence S est un type de sous-séquence particulier, qu'il est possible d'identifier après avoir ordonné S selon l'ordre lexicographique (définition 4.2.3). Un arbre de préfixes est une structure de données compressant l'information contenue dans D en factorisant les séquences de D en fonction de leur préfixe et en représentant un changement d'itemset avec deux types de branches, le premier pour relier deux items d'un même itemset, le second pour marquer un changement d'itemset entre deux items. Chaque chemin partant de la racine vers une feuille représente une séquence. La table 4.5 présente une base de données séquentielles D_1 composée de cinq tuples dont les séquences sont ordonnées selon $\succ_{lex}$ (en l'occurrence l'ordre alphabétique). À partir de D_1, nous pouvons construire l'arbre de préfixes A_{D_1} illustré figure 4.3. La figure 4.3 montre que A_{D_1} est un arbre de préfixes factorisant les séquences de D_1 selon leur préfixe. Les feuilles ne stockent pas la séquence comme dans un arbre de hachage. Les séquences sont stockées dans les chemins depuis la racine reliant les feuilles, représentant chacune un item, avec deux types de branches. Ainsi, en parcourant A_{D_1} depuis la racine nous pouvons retrouver toutes les séquences de D_1. L'avantage d'un arbre de préfixe par rapport à une base de données est qu'il permet de réduire le coût de calcul des MS fréquents car en condensant l'information il évite des balayages répétés de la base de données.

Contraintes pour extraire des MS « intéressants » L'objectif de la fouille de MS est d'extraire des MS jugés intéressants par l'utilisateur. L'intérêt d'un MS est mesuré avec des contraintes fixées par l'utilisateur qu'il doit remplir. Comme la définition d'un

tid	**Séquence**
1	⟨(*mot:Mdrrr*, *pos:adjectif*) (*mot:en*, *pos:préposition*) (*mot:rajoute*, *pos:verbe*) (*mot:pas*, *pos:adverbe de négation*) (*mot:non*, *pos:adverbe de négation*) (*mot:plus*, *pos:adverbe de négation*) (*mot:là*, *pos:adverbe*) (*mot:url_path*, *pos:nom propre*)⟩
2	⟨(*mot:Mdrrr*, *pos:adjectif*) (*mot:il*, *pos:pronom personnel*) (*mot:vient*, *pos:verbe*) (*mot:de*, *pos:préposition*) (*mot:sous-entendre*, *pos:verbe infinitif*) (*mot:que*, *pos:conjonction de coordination*) (*mot:Ademo*, *pos:nom propre*) (*mot:vend*, *pos:verbe*) (*mot:encore*, *pos:adverbe*) (*mot:url_path*, *pos:nom propre*)⟩

TABLE 4.4 – Exemple de base de données, notée D, avec deux tuples associant *tid* et séquences, pour représenter le sous-corpus du registre familier de la table 4.1 contenant deux textes. Chaque mot est représenté par deux traits linguistiques : le mot lui même et sa fonction grammaticale.

motif intéressant est dépendante de l'objectif pour lequel la fouille de MS est utilisée, il existe une multitude de contraintes pour l'extraction des MS intéressants.

Definition 4.2.5 (Contrainte de gap)**.** Un motif avec un gap $[M, N]$, noté $S_{[M,N]}$, est un motif dont chaque couple d'itemsets est séparé par au moins $M - 1$ itemsets et au plus $N - 1$ itemsets. Par exemple, $S_{[1,3]} = \langle\{a\}\{d\}\rangle$ est un motif qui apparaît dans les tuples T_1 et T_2 de D_1 donné table 4.5. En revanche, $S_{[1,1]} = \langle\{a\}\{d\}\rangle$ est un motif qui apparaît seulement dans le tuple T_1.

Parmi ces contraintes, certaines sont plus génériques que d'autres car pouvant s'appliquer à un grand nombre de cas d'usages. Citons entre autres celles dérivées des contraintes pour les ME introduites section 4.1.1 : la fréquence, la maximalité, la clôture ou encore l'émergence. Leurs définitions restent les mêmes mais s'appliquent à des MS au lieu de ME. À ces dernières, ajoutons la contrainte de gap présentée à l'instant qui permet d'extraire des séquences dont les itemsets peuvent être non consécutifs. La tolérance d'un ou

tid	**Séquence**
1	$\langle\{a\}\{c,d\}\rangle$
2	$\langle\{a\}\{b\}\{d\}\rangle$
3	$\langle\{a\}\{b,c\}\rangle$
4	$\langle\{c,d\}\{b\}\rangle$
5	$\langle\{c,d\}\{c\}\rangle$

TABLE 4.5 – Base de données séquentielles D_1, avec cinq tuples associant *tid* et séquences.

plusieurs itemsets inconnus entre des itemsets connus augmente la dimension générique d'un MS. La majorité des algorithmes utilise la contrainte de fréquence en la croisant avec d'autres contraintes pour extraire des sous-ensembles de MS fréquents. Nous dressons à la suite de cette section un panorama des algorithmes dont l'objectif est de découvrir différents types de MS fréquents.

4.2.2 Algorithme de fouille de motifs séquentiels

Les algorithmes de fouille de MS font face à trois enjeux majeurs (1) découvrir l'ensemble complet des MS intéressants (2) être suffisamment efficaces pour permettre le passage à l'échelle, c'est-à-dire le traitement d'un corpus de données réelles et (3) considérer diverses contraintes définies par l'utilisateur. L'approche générale est la même que pour les ME (générer/élaguer), cependant la complexité algorithmique liée à l'espace de recherche considéré est augmentée lors de la fouille de MS. En effet, l'espace de recherche est plus grand que celui considéré pour la fouille de ME : le nombre de MS candidats est supérieur au nombre de ME candidats. Pour générer un MS candidat, la jointure entre deux MS se fait avec deux types d'extension de séquences : la S-Extension et la I-Extension. Considérant une séquence S, une S-Extension de S par un item e ajoute un nouvel itemset à S, une I-Extension de S par e ajoute un nouvel item à S. Nous disons que e est un suffixe des MS générés. Par exemple, la séquence $S = \langle\{a,c\}\{b\}\rangle$ devient la séquence $S_1 = \langle\{a,c\}\{b\}\{c\}\rangle$ avec une S-Extension, ou bien la séquence $S_2 = \langle\{a,c\}\{b,c\}\rangle$ avec une I-Extension ; $\{c\}$ est le suffixe de S_1 et S_2. Pour une séquence de k items, nous avons $2^k - 1$ sous-séquences distinctes. Ainsi, pour une D contenant n séquences, nous avons $n(2^k - 1)$ séquences candidates à générer. La complexité algorithmique lors de la fouille de MS est liée au nombre de MS contenus dans l'espace de recherche, mais également au coût des opérations pour les générer et les tester. Pour limiter ce coût algorithmique, les approches cherchent à réduire le nombre de candidats à générer et les stocker avec des

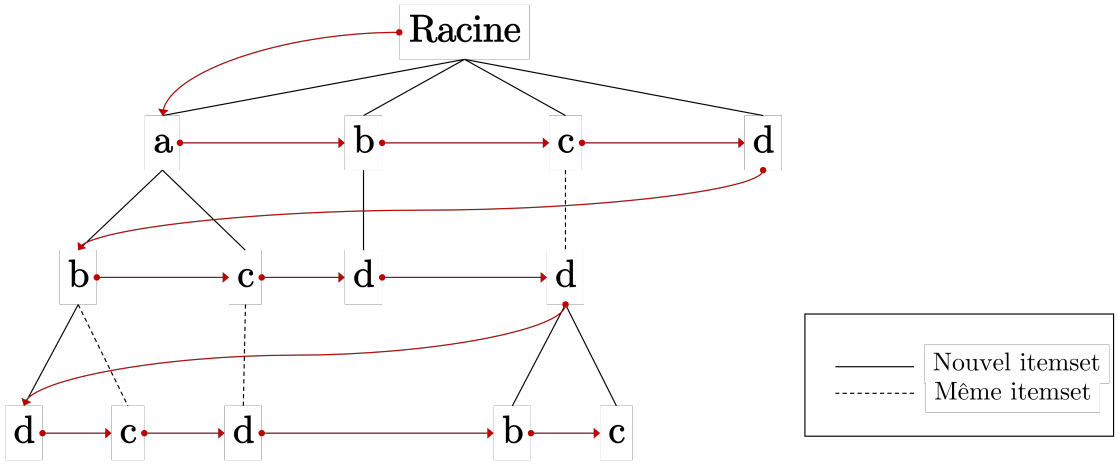

FIGURE 4.4 – Parcours en largeur de l'arbre A_{D_1}

structures de données condensées tout en limitant le nombre de balayages de ces structures par l'algorithme. Cette section dresse un panorama des algorithmes d'extraction de MS découvrant l'ensemble complet des MS fréquents, puis ceux utilisant des contraintes particulières pour les filtrer en sous-ensembles.

4.2.2.1 Fouille de motifs séquentiels fréquents

Un algorithme de fouille de MS fréquents a pour but d'extraire tous les MS dans une base de données D dont la fréquence est supérieure ou égale au seuil *minsup* fixé par l'utilisateur. Appliqué à notre travail, la fouille de MS fréquents découvre des motifs linguistiques fréquents dans un corpus en conservant la notion d'ordre entre les mots. Nous détaillons dans cette partie les algorithmes de fouille de MS fréquents, selon deux catégories de méthode de recherche : la recherche en largeur et la recherche en profondeur. Ces deux méthodes diffèrent dans leur façon d'explorer un arbre, en l'occurrence un arbre de motifs.

Recherche en largeur Un algorithme suivant une recherche en largeur explore un arbre par ordre de profondeur. La figure 4.4 montre l'arbre A_{D1} parcouru par un algorithme l'explorant en largeur, les flèches rouges illustrent les chemins empruntés. Les contenus des noeuds sont explorés dans l'ordre suivant : $\emptyset, a, b, c, d, b, c, d, d, d, c, d, b$ et c. L'algorithme de base pour la fouille de MS fréquent est l'algorithme *AprioriAll* introduit par (AGRAWAL et SRIKANT 1995). *AprioriAll* adopte une recherche en largeur avec une

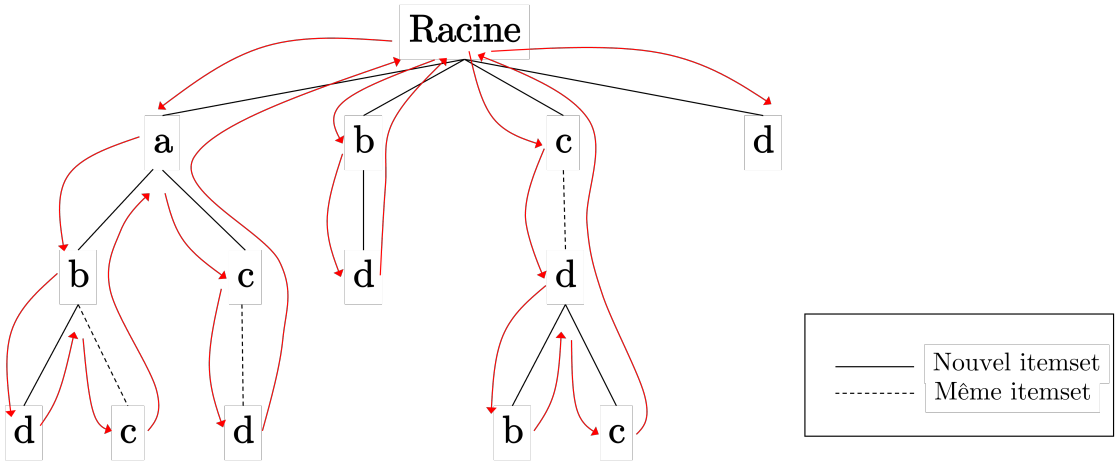

FIGURE 4.5 – Parcours en profondeur de l'arbre de A_{D_1}

représentation horizontale des données puisqu'il génère des MS candidats dans l'ordre croissant de leur taille. La taille d'un MS est son nombre d'items, par exemple $\langle\{a\}\{b\}\rangle$ et $\langle\{a,b\}\rangle$ sont tous les deux de la taille 2. *AprioriAll* procède en suivant les mêmes étapes que l'algorithme *Apriori* (présenté section 4.1.2), mais en générant des MS candidats au lieu de ME candidats. *GSP*, proposé par (SRIKANT et AGRAWAL 1996), adopte également une exploration en largeur des données puisqu'il généralise l'algorithme *AprioriAll* en y intégrant des taxonomies [4], une contrainte de fenêtre glissante et une contrainte de gap. La recherche en largeur présente trois principales limites :

1. le temps d'exécution dû au balayage de la base de données dans sa totalité à chaque taille de MS candidats ;
2. la génération de candidats fictifs non présents dans la base de données ;
3. le coût en mémoire provoqué par le nombre élevé des MS candidats de taille k devant être maintenus en mémoire pour pouvoir générer les candidats de taille $k+1$.

Recherche en profondeur Un algorithme suivant une recherche en profondeur explore un arbre de manière récursive. En partant de la racine de l'arbre, il explore un premier chemin jusqu'à sa dernière feuille, il repart alors en cherchant le sommet le plus proche qu'il n'a pas déjà visité depuis lequel il redescend jusqu'à la dernière feuille, et ainsi de suite. La figure 4.5 montre l'arbre A_{D1} parcouru par un algorithme l'explorant en profondeur, les

4. Selon le TLFi, une *taxonomie* est une « classification d'éléments ou bien une suite d'éléments formant des listes qui concernent un domaine, une science ».

a

tid	itemsets
1	1
2	1
3	1

b

tid	itemsets
2	2
3	2
4	2

c

tid	itemsets
1	2
3	2
4	1
5	1,2

d

tid	itemsets
1	2
2	2
4	1
5	1,2

FIGURE 4.6 – Représentation verticale de D_1 pour les quatre items fréquents d'après un *minsup* à 50% : a, b, c et d.

flèches rouges illustrent les chemins empruntés. Deux types de représentation des données sont utilisés pour la recherche en profondeur : une représentation verticale des données et une représentation en motifs croissants.

Représentation verticale des données La représentation verticale des données stocke chaque item i d'une D en l'associant avec la liste des séquences et itemsets où i apparaît. Pour mémoriser ces informations, la structure de données utilisée est appelée $IDListe$. Pour chaque item i fréquent dans une D, une $IDListe$ mémorise l'identifiant de la séquence *tid* de la séquence où i apparaît et l'horodatage de l'itemset [5] contenant i. Par exemple, considérant D_1 (donné table 4.5) et $minsup = 75\%$, les items a et b sont fréquents. Leurs $IDListes$ sont notées $IDListe(\{a\})$ et $IDListe(\{b\})$: $IDListe(\{a\}) = \{(1,1)(2,1)(3,1)\}$, car a est contenu dans les trois premières séquences de D_1 et toujours dans le premier itemset des séquences ; $IDListe(\{b\}) = \{(2,2)(3,2)(4,2)\}$, car b est contenu dans les deuxième, troisième et quatrième séquences de D_1, et toujours dans les seconds itemsets des séquences. Chaque $IDListe$ est illustrée par une table dans laquelle les tuples sont représentés par une ligne. La première colonne de la table stocke le *tid*, la seconde colonne stocke l'horodatage des itemsets. Par exemple, la figure 4.6 donne la représentation verticale de D_1 où chaque item fréquent est illustré par une table contenant son $IDListe$. Lorsqu'un item fréquent est présent dans deux itemsets d'une même séquence, les deux horodatages correspondants sont notés dans la seconde colonne en étant séparés par une virgule. Cette représentation de l'espace de recherche permet de le diviser en sous-espaces indépendants. L'algorithme *SPADE* proposé par (Mohammed J ZAKI 2001), ayant pour but d'extraire les MS fréquents d'une base de données D, réduit son coût algorithmique en tirant partie de cette représentation sur deux points notables :

5. L'horodatage d'un itemset dans une séquence est sa place dans la séquence : premier itemset, deuxième, troisième, etc.

item	vecteur de bits
a	10100100000
b	00010010100
c	01000011011
d	01001001010

FIGURE 4.7 – Représentation verticale de D_1 avec des vecteurs bits pour les quatre items fréquents d'après un *minsup* à 50% : a, b, c et d.

1. la génération des candidats utilise les sous-espaces de D : les MS candidats sont générés en opérant une jointure entre les $IDListes$ du préfixe considéré et de l'item pris en compte pour l'extension (S-Extension ou I-Extension) sans balayer à nouveau D ;
2. le calcul de la fréquence des candidats est simplifié : le support permettant de calculer la fréquence est obtenu en comptant le nombre de séquences distinctes contenues dans l'$IDListe$ du candidat.

Par exemple, pour générer le candidat $\langle\{a\}\{b\}\rangle$ avec une S-extension, nous joignons les $IDlistes$ de a et b . Nous obtenons alors $IDliste(\{a\}\{b\}) = \{(2,2)(3,2)\}$. Comme $IDliste(\{a\}\{b\})$ contient deux séquences distinctes, la fréquence de $\langle\{a\}\{b\}\rangle$ est de 40% car $frequence(\langle\{a\}\{b\}\rangle, D) = \frac{2}{5} = 0,4$. Un autre algorithme fouillant les MS fréquents utilise cette représentation verticale des données en améliorant le système d'$IDlistes$: l'algorithme *SPAM* proposé par (AYRES et al. 2002). *SPAM* réduit le temps nécessaire pour joindre deux $IDlistes$ en encodant les $IDlistes$ en *vecteurs de bits*. Considérant n itemsets contenu dans D, chaque item est encodé par un vecteur de bit composé de n coordonnées, où chaque coordonnée binaire représente l'absence ou la présence de l'item dans l'itemset considéré. Par exemple, la vecteur de bits de l'item a de D_1 est le suivant : 10100100000. Il est composé de 11 coordonnées binaires, car D_1 contient 11 itemsets : 1 indique la présence de a dans l'itemset et 0 son absence. Les vecteurs de bits sont stockés dans une table où chaque ligne représente un item fréquent : la première colonne de la table contient l'item fréquent, la seconde colonne son vecteur de bits. La figure 4.7 montre la table représentant les quatre items fréquents de D_1, en considérant *minsup* à 50%, via leur vecteur de bits. L'avantage de *SPAM* par rapport à *SPADE* est de réduire le temps nécessaire pour la génération des candidats en transformant l'opération de jointure en une comparaison entre deux vecteurs de chiffres. En revanche, l'inconvénient de *SPAM* est qu'il n'utilise pas de sous-espaces de recherche indépendants contrairement à *SPADE* :

a

tid	séquence
1	⟨{c, d}⟩
2	⟨{b}{d}⟩
3	⟨{b, c}⟩

c

tid	séquence
1	⟨{_, d}⟩
4	⟨{_,d}{b}⟩
5	⟨{_, d}{c}⟩

FIGURE 4.8 – Deux bases de données projetées pour les préfixes a et c de la base de données D_1 illustré table 4.5.

il a donc besoin de plus de mémoire. Pour répondre à cette limite, l'algorithme *bitSPADE* proposé par (ASEERVATHAM, OSMANI et VIENNET 2006) utilise l'encodage en vecteurs de bits tout en considérant un espace de recherche composé de sous-espace indépendants. L'usage de ces deux techniques permet à *bitSPADE* d'être plus rapide que *SPAM* tout en consommant moins de mémoire que *SPADE*.

Cependant, ces trois algorithmes génèrent des candidats fictifs. Cette étape restant la plus coûteuse lors de la fouille de MS fréquents, d'autres approches ont cherché en à limiter le coût en supprimant la génération de ces candidats fictifs.

Représentation en motifs croissants La recherche en profondeur, utilisant une représentation en motifs croissants (ou *growing pattern*), évite la génération de candidats fictifs absents de la base de données D à fouiller. Une représentation en motifs croissants est basée sur la notion de *bases de données projetées* introduite par (J. HAN, PEI et YIN 2000). La base de données projetée d'un MS est l'ensemble des suffixes suivants la première occurrence de ce MS. Le MS représente alors le préfixe de toutes les séquences contenues dans la base de données projetée. Lorsque le suffixe est ajouté avec une I-extension, c'est-à-dire qu'il commence dans le même itemset que le préfixe, le suffixe est précédé du symbole « _ ». La figure 4.8 donne deux exemples de bases de données projetées : la première pour le préfixe a, la seconde pour le préfixe c. Les bases de données projetées sont composées de deux colonnes, la première donne le *tid* de la séquence contenant le suffixe donné par la seconde colonne. Chaque ligne de la base projetée représente une séquence. Dans la base projetée de a, aucun suffixe ne commence dans un itemset partagé avec a, à l'inverse de la base projetée de c où les trois suffixes commencent dans le même itemset que le préfixe c. *FreeSpan* proposé par (J. HAN, PEI, MORTAZAVI-ASL et al. 2000) est le premier algorithme a avoir utilisé les bases de données projetées pour :

1. éviter la génération de candidats fictifs : afin d'étendre un MS noté S, *FreeSpan*

fouille la base projetée de S dans le but de trouver les suffixes candidats ;

2. accélérer le calcul de la fréquence d'un candidat : la fréquence d'un candidat est calculée dans le cadre de la base de données projetée et non plus dans la base entière.

Bien que ces bases de données projetées réduisent le coût algorithmique de l'étape de générations et tests des MS candidats, leurs constructions nécessitent plusieurs balayages de la base de données. C'est pourquoi (Pei, J. Han, Mortazavi-Asl, J. Wang et al. 2004) a proposé l'algorithme *PrefixSpan* remplaçant les bases de données projetées par des *pseudo bases de données projetées*. *PrefixSpan* crée pour chaque préfixe considéré lors de la génération de MS candidats, des pseudo bases de données projetées, c'est-à-dire des pointeurs contenant les positions du préfixe dans la base de données D. Cette stratégie permet à *PrefixSpan* d'être plus efficace que *FreeSpan* en termes de temps d'exécution et de mémoire utilisée.

Nous pouvons conclure ce panorama des algorithmes pour la fouille de MS fréquents en établissant que les approches utilisant une recherche en largeur sont moins efficaces tant en termes de temps que de mémoire, que celles utilisant une recherche en profondeur. Parmi les algorithmes faisant une recherche en profondeur, ceux adoptant une représentation verticale des données sont les plus rapides mais les plus coûteux en mémoire, tandis que ceux adoptant une représentation en motifs croissants sont plus lents mais moins coûteux en mémoire. Enfin, quelque soit l'algorithme utilisé, une des limites de ces techniques de fouille de MS fréquents est la quantité de résultats retournés : le poids des fichiers de résultats peut excéder la capacité de stockage de la machine et le nombre de MS peut rendre impossible l'interprétation des résultats par un utilisateur. Par conséquent, certains travaux ont proposé de fouiller les MS fréquents en ajoutant des contraintes afin d'en obtenir des sous-ensembles plus exploitables.

4.2.2.2 Fouille de motifs séquentiels fréquents sous contraintes

Nous rappelons que l'objectif principal de la fouille de MS est de découvrir des MS « intéressants » pour l'utilisateur. Si la fréquence est une première contrainte pour filtrer les MS, elle est insuffisante pour garantir à l'utilisateur que tous les MS fréquents seront intéressants. En outre, il est impossible pour l'utilisateur de chercher manuellement les MS intéressants parmi les MS fréquents car trop nombreux. C'est pourquoi, afin d'extraire seulement les MS intéressants pour l'utilisateur, il est possible de contraindre la fouille de MS avec d'autres contraintes que la fréquence. La fouille de MS fréquents sous contraintes

présente un double intérêt, celui de réduire le nombre de MS extraits ainsi que l'espace de recherche considéré en précisant d'avantage le type de MS à extraire.

Dans le cadre de notre étude, nous voulons caractériser les registres de langue à partir d'un large corpus de textes en posant le moins d'*a priori* linguistique possible afin d'obtenir une chaîne de traitement pouvant caractériser d'autres phénomènes linguistiques et avec des MS interprétables et permettant de distinguer deux registres de langue différents. L'objectif de cette section est de montrer différentes contraintes permettant de découvrir divers types de MS afin de motiver nos choix de contraintes utilisées dans notre travail qui sont la clôture et l'émergence.

Fouille de motifs séquentiels maximaux Les MS maximaux condensent les MS fréquents, puisque la fouille de MS maximaux a pour objectif de découvrir l'ensemble complet de MS fréquents non inclus dans un autre MS fréquent. L'avantage de ce sous-ensemble pour notre travail est qu'il retourne des MS plutôt longs, c'est-à-dire avec plusieurs itemsets et items. Comme nous traitons du langage naturel, un MS long est plus facile à interpréter car porteur d'un contexte plus étendu qu'un MS court. Par exemple, le MS $\langle$(*pos:pronom personnel*)$\rangle$ est plus difficilement interprétable que le MS $\langle$(*mot:il*, *pos:pronom personnel*)(*pos:verbe*)(*mot:mal*, *pos:adverbe*)$\rangle$. Plusieurs algorithmes ont pour tâche d'extraire les MS maximaux, ils reprennent les deux types de recherche que nous avons détaillés dans la section précédente. Par exemple, les algorithmes *DIMASP* proposé par (Garcia-Hernandez, Martinez-Trinidad et Carrasco-Ochoa 2006) et FMMSP proposé par (N. P. Lin, Hao et al. 2007) adoptent une recherche en profondeur avec une représentation en motifs croissants plus efficace qu'une recherche en largeur. D'autres algorithmes mettent en place des stratégies d'échantillonnage pour limiter le coût algorithmique de la fouille de MS maximaux. Par exemple, *MSPX* proposé par (Luo et Chung 2005) utilise des échantillons multiples pour exclure efficacement les candidats non fréquents.

Si nous n'avons pas utilisé ce type de MS c'est qu'ils perdent l'information de la fréquence des MS inclus dans un MS maximal contrairement au MS clos introduit à la suite de ce paragraphe.

Fouille de motifs séquentiels clos La tâche de fouille de MS clos est d'extraire tout MS fréquent qui n'a aucun sur-ensemble le contenant et ayant la même fréquence. Les MS clos condensent les MS fréquents sans perte d'information sur la fréquence des MS inclus

dans un MS clos. C'est pourquoi nous avons choisi ce type de MS : ils permettent de découvrir des MS interprétables du fait de leur longueur sans perte d'information. Parmi les algorithmes proposés dans la littérature plusieurs réduisent le coût algorithmiques en adoptant des stratégies évoquées dans la section 4.2.2 : *CloSpan* introduit par (Yan, J. Han et Afshar 2003) adopte une représentation en motifs en croissants ; *BIDE* présenté par (J. Wang et J. Han 2004) introduit un nouveau test de la clôture qui réduit le nombre de candidats à maintenir en mémoire ; ou encore *ClaSP* et *CM-ClaSP* proposés par, respectivement, (Gomariz et al. 2013) et (Fournier-Viger, Gomariz et al. 2014) qui utilisent une représentation verticale des données ; etc. Pour fouiller les MS clos à partir de nos données, nous avons utilisé l'algorithme *CloSPEC* développé par (Béchet et al. 2015) car il met en évidence l'intérêt de pouvoir croiser différents types de contraintes. De fait, la majorité des approches présentées introduisent des contraintes essentiellement numériques. Pour répondre à cette limite, les auteurs proposent *CloSPEC* qui gère également des contraintes de type symbolique ou bien syntaxique. Au total, *CloSPEC* permet de combiner simultanément sept contraintes d'origines diverses :

1. la fréquence minimale : afin d'extraire des motifs dont le support est supérieur à un seuil de fréquence fixé par l'utilisateur ;
2. la taille minimale : seulement les motifs dont la taille est supérieure à ce seuil seront extraits, cela permet de retirer les motifs trop courts et donc peu interprétables ;
3. la contrainte d'appartenance : seuls les motifs qui contiennent des items sélectionnés par l'utilisateur seront extraits ;
4. la contrainte d'association : tous les motifs séquentiels extraits doivent satisfaire une contrainte de relation entre deux classes d'items ;
5. la portée maximale : seuls les motifs dont la valeur maximale de la portée linguistique d'un motif respecte un seuil fixé par l'utilisateur sont extraits ;
6. le gap : seulement les motifs dont les itemsets sont séparés par un intervalle précisé par l'utilisateur sont conservés (voir la définition 4.2.5) ;
7. la clôture : les motifs séquentiels conservés sont des séquences dont aucune n'est incluse dans une autre séquence ayant le même support.

Bien qu'une de nos contraintes était de ne pas poser d'*a priori* linguistiques trop forts pour l'extraction de MS intéressants, la possibilité de pouvoir croiser ces diverses contraintes nous paraissait intéressante pour explorer nos données. Dès lors, les MS clos réduisent les MS fréquents à un ensemble plus interprétable car moins volumineux et moins redondants.

Toutefois, ils ne permettent pas de distinguer deux classes, en l'occurrence deux registres de langue.

Fouille de motifs séquentiels émergents Les travaux de (Plantevit et Crémilleux 2009) introduisent le principe de filtrer les MS fréquents selon des mesures permettant de conserver les MS qui maximisent des règles de classification. Autrement dit, ces MS doivent distinguer deux classes entre elles. Nous appelons ce type de motifs des motifs séquentiels émergents (MSE). Appliqué à notre cas d'usage, la fouille de MSE permet de découvrir des MS représentant des descripteurs linguistiques composés de divers traits linguistiques et permettant de caractériser un registre de langue cible par rapport à un registre de langue source. Par exemple, pour caractériser le registre cible familier par rapport au registre source courant, nous souhaitons découvrir le motif $S = \langle(mot{:}Mdrrrr, pos{:}adverbe)\rangle$ car il est fortement présent dans les textes du registre cible et à l'inverse absent des textes du registre source. Dès lors, la présence de S dans un texte doit permettre à l'utilisateur de le percevoir comme du registre familier et non du registre courant. Parmi les mesures proposées dans (Plantevit et Crémilleux 2009), nous avons choisi le taux de croissance (présenté équation 4.4 page 96) car il rend compte du rapport de la présence d'un même MS dans un ensemble de données D_1 et de sa présence dans un ensemble de données D_2.

Autres sous-ensembles de motifs séquentiels D'autres types de contraintes existent dans la littérature, elles permettent de découvrir d'autres MS intéressants pour l'utilisateur. Ne les ayant pas utilisées dans nos travaux, nous les mentionnons simplement dans ce paragraphe pour donner au lecteur une idée de la diversité des contraintes existantes. Les MS *δ-libres*, proposés par (Holat, Plantevit et al. 2014), condensent l'information en rassemblant les MS selon leurs supports plus ou moins proches, c'est-à-dire à dire avec une tolérance de plus ou moins δ. D'autres comme les MS Top-k sélectionnent les k meilleurs MS selon leurs fréquences ce qui évite à l'utilisateur de fixer un *minsup* ; les MS périodiques fouillent des MS fréquents un intervalle de MS dont la taille [6] est fixée par l'utilisateur ; les MS flous introduits par (Zadeh 1996) permettent de traiter des données quantitatives en permettant aux sous-ensembles flous d'appartenir partiellement à plusieurs classes. Un type de MS particulier aurait pu être intéressant pour la caractérisation des registres : les MS multidimensionnels introduits par (Pinto et al. 2001). Ils

6. La taille d'une période est le nombre de MS qu'elle contient.

visent à découvrir des MS en considérant à la fois la notion d'ordre et plusieurs dimensions d'analyse. Par exemple, $S = (tid_1, (a_1, ..., a_m), s)$ est un MS multidimensionnel, où a_i sont les différentes dimensions choisies pour décrire le MS et s le MS. Appliqué à notre cas d'usage, un MS multidimensionnel serait *(1(familier),⟨(mot:Mdrrrr, pos:adverbe)⟩)*. Comme nous ne considérons pas d'informations sur la situation d'énonciation, les MS multidimensionnels perdent de leur intérêt, car nous savons déjà que le MS est du registre familier puisqu'il vient de ce sous-corpus. Dès lors, les MS multidimensionnels introduiraient une complexité algorithmique supplémentaire qui ne serait pas contrebalancée par un gain d'information notable pour notre objectif.

4.2.3 Synthèse

Dans cette section, nous avons présenté formellement les MS ainsi que diverses manières de les filtrer en sous-groupes : les MS fréquents, maximaux, clos, émergents, etc. À cette occasion, nous avons pu constater que les MS sont particulièrement adaptés aux données textuelles car ils considèrent l'ordre entre les mots et les décrivent avec un ensemble de facteurs. Cela permet de les décrire avec différents traits linguistiques donnant une description des données textuelles plus riche que celle obtenue avec les ME. Nous avons ensuite donné une vue d'ensemble des divers types d'algorithmes permettant de découvrir des MS intéressants à partir d'une base de données. Notre objectif en dressant ce panorama des algorithmes existants était d'expliquer pourquoi leur complexité algorithmique est exponentielle et comment ces algorithmes mettent en place différentes stratégies pour la limiter. Nous avons vu que ces stratégies ont toutes des avantages et des inconvénients : certaines réduisent le temps d'extraction des MS intéressants, d'autres la mémoire nécessaire au maintient des MS candidats. Aucune approche n'est parfaite, c'est à l'utilisateur de trouver un compromis entre les MS qu'il souhaite extraire et ses contraintes matérielles liées à la machine qu'il utilise.

Pour prouver la robustesse des algorithmes de fouille de MS face à des corpus de données réelles, la section suivante présente plusieurs cas d'usage donnant des exemples d'application des techniques de fouille de MS à partir de données textuelles.

4.3 Application des techniques de fouille de motifs séquentiels à des données textuelles

Dans cette section, nous détaillons différents travaux utilisant la fouille de MS pour extraire des connaissances à partir de données textuelles. Selon le domaine et le type de données séquentielles, les algorithmes sont utilisés différemment et adaptés à la tâche. Notre objectif n'est pas de lister de manière exhaustive tous les domaines et toutes les études utilisant des techniques de fouille de MS à partir de corpus de textes. Nous voulons plutôt montrer :

— la robustesse des algorithmes de fouille de MS face à des données réelles ;

— comment les algorithmes sont adaptés aux études les utilisant ;

— la pertinence des MS pour des études de linguistiques de corpus.

Nous passons tout d'abord en revue des travaux travaillant sur le domaine biomédical, pour lequel les techniques de fouille ont beaucoup été utilisées. Puis, nous présentons d'autres applications de la fouille dans des études linguistiques.

4.3.1 Fouille de motifs séquentiels dans des données biomédicales

Historiquement, les recherches en bio-informatique ont beaucoup utilisé les techniques de fouilles de MS qui s'adaptent à la séquentialité des données biomédicales. Par exemple, la fouille de MS dans les protéines fournit des informations intéressantes comme l'identification des familles de protéines. Pour n'en citer que quelques uns, nous pouvons citer les travaux de (M. Huang et al. 2004) qui proposent de découvrir automatiquement de nouvelles interactions protéine-protéine. Leur méthodologie aligne des phrases afin d'en extraire les parties similaires. Le corpus anglais est composé de 1 200 phrases collectées à partir des 50 premiers résultats retournés par la requête "*protein–protein interaction*". Chaque phrase devait au moins contenir deux noms de protéines différentes. Le corpus obtenu est annoté automatiquement en parties grammaticales avec l'outil GENIA [7]. En amont de l'extraction de motifs séquentiels, une liste des verbes fréquents est établie et les phrases qui ne mentionnent pas au moins un verbe de cette liste ou bien sans verbe sont exclues. L'extraction de motifs séquentiels fréquents à partir du corpus filtré contraint la

7. `https://github.com/spyysalo/genia-pos`

fouille avec un $minsup = 5$. Finalement, 134 relations protéine-protéine sont extraites. Mentionnons aussi (Ho, Lukov et Chawla 2005), qui explore des structures récurrentes de protéines grâce aux MS pour découvrir de nouveaux rôles fonctionnels des protéines présentes dans ces structures. Les auteurs adaptent l'algorithme *SPAM* fouillant les MS fréquents, introduit par (Sohn, Lee et Rim 2009), à leur tâche et proposent *Pex-SPAM*. *Pex-SPAM* ajoute à *SPAM* une contrainte de *gap* et une contrainte d'expression régulière, c'est-à-dire que les MS devront contenir l'expression régulière spécifiée par l'utilisateur. Le corpus [8] utilisé contient 148 protéines et a été constitué par (Möller, Kriventseva et Apweiler 2000). Chaque protéine est représentée par un itemset dont les items indiquent les propriétés de la protéine : $I = \{prot, prop_1, prop_2, ..., prop_n\}$ où $prot$ donne le nom de la protéine décrite par $prop_n$. L'ensemble des MS extraits varie de 0 à 1 200 séquences selon les différentes contraintes appliquées. (Yun et Yangyong 2007) quant à eux présentent *BioPM* : un algorithme d'extraction de MS adoptant une représentation des données par motifs croissants puisque *BioPM* est dérivé de *PrefixSpan*. *BioPM* est adapté à la découverte de nouvelles fonctions des protéines :

1. Afin de s'adapter aux MS protéiques parfois longs, *BioPM* définit une fenêtre w pour contrôler la largeur de la croissance des MS candidats.
2. *BioPM* introduit un score pour établir un degré de correspondance entre deux séquences protéiques $BLOSUM$, afin de décider sur une base de données projetée si c'est MS fréquent ou non.
3. *BioPM* ajoute la contrainte de *gap*.

Les expériences menées utilisent un corpus de 1 400 séquences issues de 10 familles de protéines venant de la base de données PFAM [9]. Toujours pour découvrir de nouvelles fonctions des protéines, (Q. Wang, Davis et Ren 2016) propose l'algorithme *FBSB*. Le corpus considéré est constitué de 2 747 séquences issues de 10 familles de protéines provenant de la base de données PFAM. Avec un *minsup* à 15%, l'ensemble de MS fréquents extrait compte 600 séquences. Les expériences menées ont montré que *FBSB* est plus efficace que *BioPM*. Nous pouvons également mentionner (Cellier, Charnois et Plantevit 2010) qui utilisent les Ms maximaux pour extraire des patrons linguistiques afin de découvrir des relations possibles entre pathologies et contextes biologiques. Dans le but de réduire le nombre de MS extraits, les auteurs proposent deux stratégies : appli-

8. Le jeux de données est accessible via `ftp://ftp.ebi.ac.uk/databases/testsets/transmembrane`.

9. Le jeux de données PFAM est accessible via `http://pfam.xfam.org/`.

quer des contraintes linguistiques et faire une fouille récursive. Le corpus en anglais utilisé regroupe deux ressources : un corpus de 1 806 phrases qui viennent de PubMed [10], et un corpus de 2 995 phrases sélectionnées par un expert. Tous les noms propres qui désignent un gène sont normalisés avec le token *AGENE*. Chaque séquence représente une phrase dont chaque itemset représente un mot avec deux items : la forme fléchie du mot et sa catégorie grammaticale du discours [11] Ainsi, "*CCDC with TMEM in Vitro .*" est transformé en la séquence suivante $\langle(AGENE, np)(with, in)(AGENE, np)(in, in)(Vitro, np)(., sent)\rangle$. Quatre contraintes sont appliquées pour l'extraction des MS : la contrainte de fréquence minimale (*minsup*) ; les MS doivent contenir deux noms propres ; les MS doivent contenir un verbe ou bien un nom commun ; le motif doit être maximal afin de réduire la redondance des MS extraits. L'ensemble des MS extraits compte 65 000 séquences. Pour en réduire le nombre et en garder que les plus pertinentes les auteurs proposent une fouille de MS récursive. Introduite par (Crémilleux et al. 2009), la fouille de MS récursive réduit le nombre de MS extraits en répétant successivement la fouille : à chaque récursion, le résultat précédent est considéré comme le nouveau jeu de données. À la fin du processus d'extraction récursif, l'ensemble ne compte plus que 667 séquences qui sont analysées par un expert afin d'en valider manuellement les plus pertinentes. Enfin, citons (Béchet et al. 2012) qui propose d'utiliser la fouille de MS fréquents pour découvrir des relations entre gènes et maladies rares à partir d'un corpus constitué via la base de données PubMed en anglais. Seulement les phrases contenant au moins une mention d'une maladie rare et d'un gène sont conservées dans le corpus. Chaque phrase conservée est transformée en séquence dans laquelle chaque mot est représenté par un itemset qui contient deux items : le lemme du mot et la catégorie grammaticale du mot [12] L'extraction des MS fréquents est faite sous 6 contraintes différentes : la fréquence minimale (*minsup*) ; la contrainte de gap ; la portée maximale ; la longueur minimale ; la contrainte d'appartenance, c'est-à-dire que les motifs doivent contenir au moins un gène, une maladie rare, et un nom ou un verbe ; la contrainte d'association qui associe un type d'item à un autre type d'item. L'ensemble des motifs extraits est alors analysé par des experts qui valident les séquences pertinentes. La table figure 4.9 détaille le nombre de motifs extraits puis validés par les experts de 3 expériences. Chaque ligne de la table représente une expérience. La première colonne précise le *minsup* considéré, la seconde colonne montre que le gap est toujours de [0,10],

10. `https://pubmed.ncbi.nlm.nih.gov/`
11. Les catégories grammaticales sont obtenue automatiquement grâce à l'outil TreeTagger proposé par (Schmid 1994).
12. Les catégories grammaticales sont obtenues automatiquement avec TreeTagger.

minsup	gap	longueur min	nb. motifs	nb. motifs validés
0,50%	[0,10]	4	22 794	6 310
0,20%	[0,10]	4	126 777	54 429
0,05%	[0,10]	all	1 530 085	416 786

FIGURE 4.9 – Résultats selon différentes contraintes de *minsup*, de *gap* et de longueur minimale extraits de (BÉCHET et al. 2012).

la troisième colonne donne la contrainte de longueur minimale. Enfin, les deux dernières colonnes détaillent le nombre de MS extraits et le nombre de MS validés.

4.3.2 Fouille de motifs séquentiels à partir de textes

Les motifs séquentiels ont également été utilisés pour fouiller des textes afin d'en extraire des connaissances linguistiques. Les motifs séquentiels permettent de respecter l'ordre des mots. La fouille de textes permet d'extraire les MS fréquents dans un large corpus, autrement dit de tirer les motifs langagiers caractéristiques d'un ensemble de textes.

Par exemple, (LUCAS, CRÉMILLEUX et TURMEL 2003) classe des articles académiques en anglais avant qu'ils soient relus pour être édités. Plus précisément, les auteurs cherchent à distinguer les articles « bien écrits » de ceux « mal écrits ». Les articles « mal écrits » sont orientés vers l'outil de correction automatique de l'éditeur. Le but est d'assurer leur « lisibilité » pour avoir « la meilleure compréhension possible » des articles académiques. Leurs travaux conjuguent deux approches :

1. une approche linguistique manuelle, qui analyse les textes pour en tirer des descripteurs linguistiques tels que des terminaisons de mots (*-ed*, *-ing*, *-ly*) ; des connecteurs logiques (*Despite*, *Indeed*, *Because*), adjectifs anaphoriques (*Its*, *Their*, *Such...*) ; marques typographiques (*[]*, *()*, *""*) ;
2. une approche automatique, qui extrait les MS *δ-libres* fréquents à partir de MS décrivant les données textuelles avec des traits linguistiques, afin de valider les descripteurs experts issus de l'analyse linguistique.

Plusieurs expériences ont fait varier la segmentation des textes où les segments peuvent être les parties de l'article, les paragraphes, les phrases ou bien les unités séparées par des virgules (*virgulots*) ; la fréquence minimale (*minsup*) ; la valeur de δ. Selon la segmentation choisie, les textes sont représentés par une liste de booléens qui indique la présence

FIGURE 4.10 – Vue générale de la méthodologie de (QUINIOU et al. 2012)

ou l'absence des descripteurs linguistiques experts. Le dernier booléen de la liste représente toujours la classe du texte (0 pour *bien écrit*, 1 pour *mal écrit*). La méthodologie adopte l'hypothèse linguistique selon laquelle un texte est une *hiérarchie inclusive*. Autrement dit, chaque segment d'un texte *hérite* en quelque sorte des propriétés des segments qui le précèdent. Ainsi, les *virgulots* héritent des descripteurs linguistiques des *parties*. L'ensemble des *motifs δ-libres* extrait varie de 29 séquences (considérant les *parties* de l'article, avec un $minsup = 50$ et $\delta = 0$), à 48 507 séquences (considérant les *virgulots* de l'article, avec un $minsup = 60$ et $\delta = 10$). Ces travaux sont repris par (LUCAS et CRÉMILLEUX 2004) qui utilisent les MSE pour extraire des motifs caractéristiques des classes considérées (articles académiques *bien écrits* vs. *mal écrits*). Tout comme (LUCAS, CRÉMILLEUX et TURMEL 2003) les expériences font varier la segmentation des textes en *partie*, *paragraphes*, *phrases* ou *virgulots*. Un des résultats obtenu avec les motifs séquentiels émergents est l'incapacité des *virgulots* à produire une discrimination entre les classes. Les auteurs en concluent que le *virgulot* ne semble pas pouvoir être caractérisé en soi, tandis que les informations les plus importantes pour la lisibilité sont supportées par les hauts niveaux (tels que les *parties* ou *paragraphes* par exemple).

Pour le français, (QUINIOU et al. 2012) proposent d'utiliser *les motifs séquentiels émergents* afin de caractériser les genres de texte différents : genre poétique, genre épistolaire, genre romanesque. Leurs travaux se situent dans le paradigme linguistique d'approches guidées par les données, c'est-à-dire découvrir de nouvelles connaissances linguistiques à partir du corpus sans poser d'hypothèse *a priori*. Cette approche s'oppose à celle où l'exploration du corpus tend à valider ou bien réfuter une hypothèse formulée *a priori*. La distinction entre ces deux approches est introduite par (TOGNINI-BONELLI 2001). La méthodologie (figure 4.10) propose de sélectionner N corpus dont les motifs séquentiels fréquents seront extraits, avant de calculer les motifs séquentiels émergents. Ces derniers sont ensuite validés par un expert dans le but de les interpréter. Les expériences menées considèrent les genres de textes suivants : la poésie, la correspondance et le roman. Chaque corpus est composé de textes qui couvrent la période 1 800-1 900 issus de la base de don-

nées FranText [13]. Les trois corpus sont étiquetés par Cordial [14] pour obtenir les formes fléchies, les lemmes et les catégories morpho-syntaxiques. Par exemple le syntagme "*les petits chats*" devient la séquence $S = \langle\{les,\ le,\ DET\}\ \{petits,\ petit,\ ADJ\}\ \{chats,\ chat,\ NC\}\rangle$. La fouille des motifs séquentiels est contrainte par une fréquence minimale (*minsup*) et une contrainte de *gap*. Ils utilisent l'algorithme *CloSpan* (Yan, J. Han et Afshar 2003) pour extraire les motifs séquentiels clos avec un $minsup = 0,15\%$. Enfin, le seuil sur le taux de croissance pour filtrer les MSE est fixé à $1,001$. Les résultats montrent que plus la contrainte de gap est stricte, plus le nombre de motifs extraits est réduit. À l'inverse, plus la contrainte est relâchée, plus l'ensemble des motifs fréquents est grand. Prenons par exemple, les résultats pour le genre épistolaire : avec $gap = [1, 1]$ l'ensemble des motifs fréquents compte 16 936 séquences, tandis qu'avec $gap = [1, 5]$ l'ensemble compte 96 549 séquences. En outre, les résultats globaux montrent que c'est le genre épistolaire qui dénombre le plus de motifs langagiers spécifiques avec jusqu'à 50% de ses motifs fréquents qui sont émergents lorsque la contrainte de *gap* est $[1, 4]$. En revanche, le genre romanesque est le moins spécifique en terme de style puisqu'il a, au plus, 6% de ses motifs fréquents qui sont émergents lorsque la contrainte de *gap* est $[1, 1]$. Cette approche est reprise par (Legallois, Charnois et Poibeau 2016). Ces travaux s'intéressent à l'extraction des *clichés* des *romans sentimentaux*. Pour ce faire, ils comparent les motifs de trois corpus : 50 romans policiers [15], 50 romans littéraires [16] et 50 romans sentimentaux [17]. Les auteurs définissent le cliché comme « un stéréotype d'expression devenu banal sous l'effet de la répétition ». Leur méthodologie se base sur l'extraction de MSE. Pour ce faire, ils transforment leur corpus étiqueté par Cordial en ne gardant que les formes des unités dites invariantes, (par exemple : les prépositions, les conjonctions, etc.) ; les lemmes de certains verbes très fréquents (par exemple : les auxiliaires, les verbes aspectuels, les verbes modaux, etc.) ; les catégories morpho-syntaxiques (par exemple : nom commun, nom propre, adverbe, verbe, etc.). Dès lors, le passage extrait de (Legallois, Charnois et Poibeau 2016) :

> Comment peut-on être aussi snob et hautaine ? Pourtant, au fil des jours, il doit bien reconnaître que Marnie sait aussi se montrer enthousiaste et généreuse, et surtout très sexy

13. La base de donnée est accessible via `https://www.frantext.fr/`.
14. Cordial est un outil développé par la société Synapse - `www.synapse-fr.com`.
15. Ce corpus est composé d'œuvres d'auteurs tels que Jonquet, Izzo, Thery, Vargas, Manchette, etc.
16. Ce corpus est composé d'œuvres d'auteurs tels que Le Clézio, Rouaud, Carrère, Modiano, etc.
17. Ce corpus est entièrement composé d'œuvres éditées par Harlequin - `https://www.harlequin.fr/`.

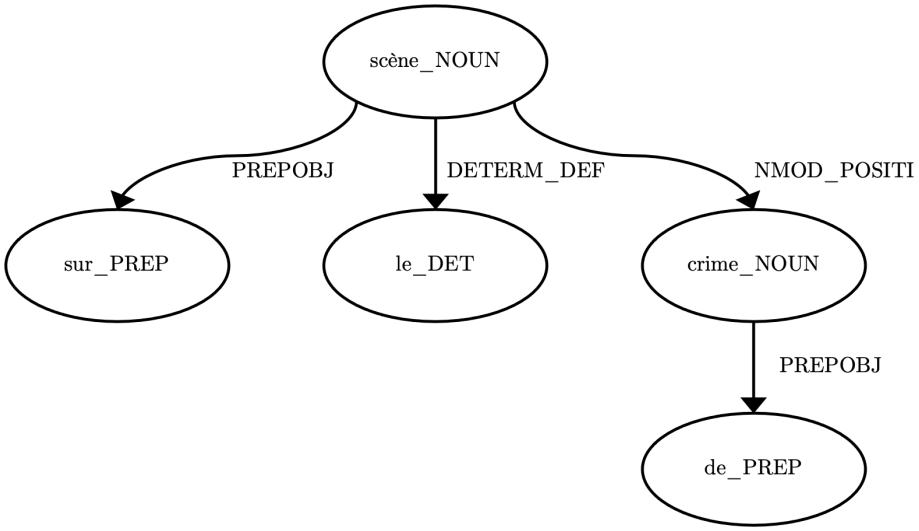

FIGURE 4.11 – Arbre lexico-syntaxique récurrent automatiquement extrait pour le syntagme *sur la scène du crime* (GONON et al. 2016).

devient la séquence extraite de (LEGALLOIS, CHARNOIS et POIBEAU 2016) :

> comment pouvoir on être aussi ADJ et ADJ ? pourtant, à le NC de le jour, il devoir bien INF que NP savoir aussi se INF ADJ et ADJ, et surtout très ADJ

Les auteurs n'utilisent pas la notion d'émergence mais un calcul des spécificités pour obtenir la spécificité d'un motif. Les motifs spécifiques sont ensuite validés par un expert avant d'être analysés par un linguiste. Enfin, citons (GONON et al. 2016) qui se placent eux même dans la lignée de (QUINIOU et al. 2012). Le but de leurs travaux est d'analyser en contexte l'utilisation de l'expression *scène de crime* dans les romans policiers. Pour ce faire, ils cherchent à repérer des motifs fréquents dans un corpus de polars. Toutefois, ils ne fouillent pas de MS mais des *motifs hiérarchiques* qui sont des *arbres lexico-syntaxiques récurrents (ALR)* introduits par (TUTIN et KRAIF 2016). Les ALR tendent à repérer les récurrences de MS émergents dans des arbres syntaxiques (un exemple est donné figure 4.11). Le corpus utilisé pour l'étude est issu d'un ensemble de textes littéraires en français contemporains [18] qui compte 16 millions de mots. Ce corpus a été divisé en plusieurs sous-corpus selon divers sous-genres. Les auteurs ont pris pour genre cible le genre policier, le sous-corpus cible est donc celui du genre policier : il compte environ 7 millions de mots pour 75 œuvres. À partir des ALR extraits de ce sous corpus, les auteurs mènent une analyse linguistique afin de comprendre le comportement du syntagme *scène de crime* en contexte.

18. Ce corpus peut être consulté via `http://emolex.u-grenoble3.fr/emoBase/`.

4.3.3 Synthèse

Nous avons donné des exemples d'application de fouille de MS dans des données textuelles venant des domaines de la recherche en bio-informatique et de la recherche en linguistique de corpus. La fouille de MS permet de traiter efficacement les données biomédicales telles que les chaînes d'ADN ou bien les relations entre protéines. De fait, l'extraction des MS fréquents tire profit d'une modélisation stable des connaissances du domaine. Cette formalisation acceptée et partagée par toute la communauté scientifique permet de contraindre l'extraction avec des règles précises qui introduisent des connaissances expertes lors de la fouille de MS. Quant à la fouille de MS à partir de textes pour l'extraction de connaissances linguistiques, bien qu'elle paraisse adaptée à la tâche, la littérature scientifique est peu dotée d'études sur le sujet. Ce volume ténu d'études pourrait être expliqué par le manque de dialogue entre le domaine informatique et la discipline linguistique. Enfin, le panorama des diverses applications met en avant la forte capacité d'adaptation des techniques de fouille à différentes problématiques. Selon les besoins utilisateurs et les spécificités du domaine d'application, les structures de données ou bien les contraintes d'extractions sont modifiées. Enfin, tous les cas d'usage montrent la conséquence du nombre volumineux des MS extraits. Du fait de leur quantité, la sélection des motifs pertinents passe toujours par une validation humaine.

Globalement, ces exemples d'applications ont confirmé la robustesse des algorithmes face à des données réelles d'une part et leur pertinence pour des études linguistiques d'autre part. Elles ont également mis en exergue la nécessité d'une expertise humaine lors de la découverte des MS intéressants pour adapter les algorithmes avec des contraintes expertes ainsi que lors de l'évaluation pour vérifier manuellement la pertinence des MS. Dans le cadre de notre travail, nous avons justement souhaité minimiser l'intervention d'experts pour découvrir les MSE caractéristiques des registres de langue.

4.4 Conclusion

Ce chapitre avait pour but de trouver quel type d'approche formelle était préférable à utiliser pour découvrir automatiquement des descripteurs caractéristiques des registres de langue. En l'occurrence, comme nous le détaillerons dans le chapitre 5, nous utiliserons les notions de MSE pour la caractérisation discriminante des registres de langue et les MS clos pour compresser les ME fréquents.

Pour comprendre les avantages et inconvénients des MS, nous avons tout d'abord

présenté les ME qui sont les motifs dérivés les MS. La première partie de ce chapitre a introduit les ME ainsi que des notions étendues aux MS comme celles de la fréquence, la clôture et l'émergence. Nous avons montré l'intérêt de la fouille de ME fréquents, tout en montrant la limite de cette formalisation qui ne considère pas l'ordre entre ses objets. En détaillant les algorithmes de fouille de ME fréquents, nous avons souligné la question de la complexité algorithmique exponentielle liée à l'étape cherchant les ME fréquents dans une base de données en générant des ME candidats, ainsi que la nécessité de filtrer l'ensemble de ME fréquents en un sous-ensemble plus interprétable car moins volumineux. Bien que nous ayons écarté les ME pour notre travail, cette section a servi à poser les bases théoriques nécessaires à la compréhension des MS que nous avons choisis d'utiliser, ainsi qu'à souligner les principales limites de ces approches.

En effet, les MS sont particulièrement adaptés à la représentation d'un descripteur linguistique car un MS respecte l'ordre des mots avec des séquences et décrit chaque mot avec plusieurs traits linguistiques via des items stockés dans des itemsets. Comme la complexité algorithmique des techniques fouillant les MS fréquents est encore plus importante que celle de la fouille de ME fréquents, nous avons détaillé dans la seconde section de ce chapitre différents types d'algorithmes associés aux stratégies mises en place pour limiter ce coût algorithmique. Cela nous a permis de mettre en lumière qu'aucun algorithme n'est parfait, car chacun favorise soit le temps d'exécution, soit la mémoire allouée. C'est à l'utilisateur de contraindre la fouille de MS pour réduire l'espace de recherche considéré selon sa tâche d'exploration. Dans notre travail, nous avons choisi de fouiller les MS clos car réduisant la redondance et la volumétrie des MS fréquents sans perte d'information, à partir desquels nous cherchons les MSE pour caractériser les registres de langue.

Enfin, nous avons montré dans la dernière section de ce chapitre que les techniques de fouille de MS fréquents passaient l'échelle face à des données réelles. Nous avons vu que si les algorithmes étaient capables de fouiller de large corpus, c'était notamment grâce aux fortes contraintes posées par des experts humains sachant quels types de MS ils cherchaient. En outre, l'intervention humaine s'est également montrée nécessaire pour évaluer les MS découverts afin de vérifier leur pertinence.

Nous ressortons de cet état de l'art en ayant choisi le type de MS à mettre en œuvre pour la caractérisation des registres de langue à partir d'un corpus de textes. Le panorama des travaux proposés dans la littérature a mis en avant l'utilisation de contraintes souvent expertes pour réduire l'espace de recherche et garantir la découverte de MS intéressants

pour l'utilisateur. Cependant, dans le cadre de notre travail nous avons souhaité obtenir un ensemble de MS intéressants et interprétables sans avoir à poser de contraintes expertes (en l'occurrence linguistiques). Le chapitre suivant expose notre méthodologie qui établit une chaîne de traitement cherchant à contraindre le moins possible la fouille de MS tout en assurant le passage à l'échelle des algorithmes et l'interprétabilité des résultats obtenus.

Chapitre 5

Fouille des ensembles de motifs séquentiels émergents caractéristiques des registres de langue

Sommaire

Dans ce chapitre, nous exposons deux études distinctes. La première est une preuve de concept qui confirme l'usage de la fouille de MSE pour découvrir des motifs [1] intéressants. La seconde vise à obtenir un ensemble de MSE caractéristiques d'un registre de langue cible par rapport à un registre de langue source à partir de données réelles. Bien qu'étant deux études indépendantes, les résultats de la première conditionnent la confiance que nous avons dans l'intérêt des résultats de la seconde.

Avec la première étude, nous avons levé le verrou sur l'incertitude de l'intérêt des MSE découverts. Pour cela, nous avons fouillé des langages artificiels contenant une base de vérité, c'est-à-dire une liste définie par l'utilisateur de motifs à retrouver. En suivant un paradigme de recherche d'informations, nous avons évalué les MSE découverts en faisant un parallèle entre eux et des documents retournés : un motif séquentiel émergent appartenant à la base de vérité a été considéré comme un document pertinent et son

1. À partir de ce chapitre le terme *motifs* renvoient aux *motifs séquentiels*.

taux de croissance comme le rang auquel il a été retourné. Trois mesures de recherche d'informations ont évalué les MSE en regardant si les MSE attendus avaient bien été découverts et si leur rang les plaçait en tête des résultats. Les expériences conduites ont validé la robustesse de la fouille de MSE car la majorité des motifs attendus a été retrouvée et le taux de croissance a correctement hiérarchisé les MSE.

La seconde fouille de MSE, cette fois ci à partir de données réelles, est guidée par deux objectifs : fouiller la totalité du corpus TREMoLo-Tweets pour découvrir des MSE décrivant l'usage réel des registres de langue par les locuteurs ; obtenir des MSE les plus interprétables possibles lorsqu'ils sont considérés un à un. Plus un motif contient de mots et de traits décrivant ces mots, c'est-à-dire plus un MSE contient d'itemsets et d'items, plus il est interprétable car précis. Par exemple, le motif $S_1 = \langle$\{lemme :aimer, sous-mot :_Aim, sous-mot :ez_, pos :verbe, temps :présent de l'indicatif, nombre :pluriel, syntax :racine\} \{lemma :trop, pos :adverbe, syntax :modifieur\} \{lemme :le, sous-mot :_les_, pos :déterminant\}$\rangle$, qui contient 3 itemsets pour 13 items, est plus interprétable que le motif $S_2 = \langle$\{syntax :racine\} \{syntax :modifieur\}$\rangle$, qui contient 2 itemsets pour 2 items. Effectivement, en croisant les informations portées par les items de S_1, nous déduisons que S_1 décrit un motif linguistique placé en début de phrase grâce à la majuscule du sous-mot « _Aim », que le vouvoiement n'est pas un vouvoiement de politesse puisqu'il est précisé que le verbe est au pluriel. Ces mêmes informations ne peuvent pas être déduites de S_2, qui est pourtant une sous-séquence de S_1 pouvant décrire le même motif linguistique. Cette différence vient des items plus génériques de S_2 sans information lexicale. Dès lors, un moyen d'obtenir des MSE interprétables serait d'avoir un grand nombre d'items par itemsets. Cependant, augmenter le nombre d'items par itemsets augmente la complexité algorithmique de la fouille de MSE. Cette solution serait envisageable sur un petit jeu de données. Or, comme TREMoLo-Tweets est un corpus volumineux avec 228 270 tweets pour 6 201 339 mots, la complexité algorithmique deviendrait vite trop importante et la fouille n'aboutirait pas faute de mémoire ou de temps. Nous n'avons donc pas assuré l'interprétabilité des MSE en utilisant un grand nombre d'items, mais en choisissant un nombre restreint d'items (en l'occurrence de traits linguistiques) dont les types permettent à l'utilisateur de comprendre le motif en les croisant entre eux. Enfin, pour conserver des motifs contenant le plus grand nombre d'itemsets sans perte d'information, nous avons cherché les MSE à partir des motifs clos. Cela assure à l'utilisateur d'avoir des motifs contenant plusieurs itemsets et items : seuls les motifs n'ayant aucun sur-ensemble avec la même fréquence sont conservés.

Nous commençons ce chapitre en décrivant notre chaîne de traitement complète pour la fouille de MSE : nous exposons de quelle manière chacune de ses étapes détermine la façon dont les MSE caractérisent les registres de langue. Nous avons appliqué cette chaîne de traitement lors des deux études mentionnées ci-dessus, elles ont validé sa robustesse : la première en confirmant la pertinence des MSE découverts, la seconde en assurant sa capacité à traiter de larges jeux de données. Les deux dernières sections de ce chapitre en détaille les expériences et les résultats.

5.1 Chaîne de traitement complète pour la fouille de motifs séquentiels émergents

La chaîne de traitement présentée dans cette section explore un grand jeu de données pour en découvrir des MSE caractéristiques d'un registre de langue cible. Pour cette tâche, nous avons veillé à réduire le risque de manquer des MSE intéressants. Cette volonté a guidé la manière dont nous avons fixé différentes valeurs de contraintes ou seuils utilisés pour fouiller les MSE. Une vision d'ensemble de notre chaîne de traitement est tout d'abord donnée, nous précisons ensuite les enjeux et problématiques liés à chacune de ses étapes.

5.1.1 Vision globale de la chaîne de traitement

L'objectif de notre chaîne de traitement est de caractériser un registre de langue cible C par rapport à un registre de langue source S à partir de deux corpus les représentant. Par exemple, pour caractériser le registre familier par rapport au registre soutenu, nous avons fouillé les MSE d'un corpus de textes étiquetés du registre familier par rapport à un corpus de textes du registre soutenu. La figure 5.1 illustre la trame générale de notre chaîne de traitement. Elle prend en entrée les deux corpus des registres cible et source, puis retourne l'ensemble de MSE caractéristiques du registre cible. Pour cela, quatre briques de traitement se succèdent :

(1) **la transformation des textes de C et S en séquences** stockées dans deux bases de données, respectivement, D_c et D_s ;

(2) **la fouille de motifs clos** à partir de D_c faite sous deux contraintes : un seuil de fréquence minimale noté $minsup_c$; une contrainte de gaps minimal et maximal notée gap_c ;

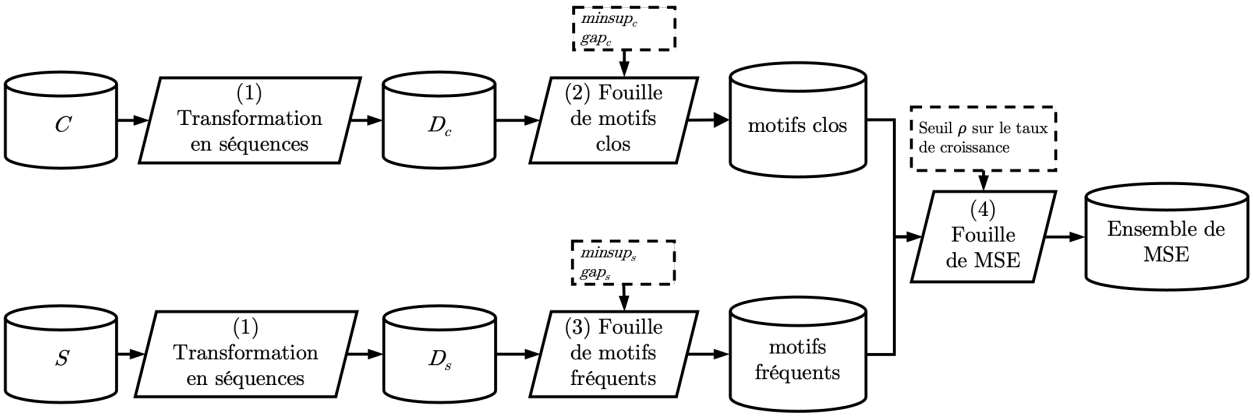

FIGURE 5.1 – Chaîne de traitement complète pour la fouille de MSE à partir de motifs clos et fréquents.

(3) **la fouille de motifs fréquents** à partir de D_s effectuée sous deux contraintes similaires que pour la fouille de motifs clos, notées $minsup_s$ et gap_s, mais dont les valeurs peuvent différer ;

(4) **la fouille de MSE** à partir des motifs clos et fréquents cherchant tous les motifs dont le taux de croissance est supérieur à un seuil ρ.

Avant de détailler les enjeux de ces étapes dans la section suivante, nous précisons ici pourquoi nous avons fouillé deux types de motifs différents dans les étapes (2) et (3) : les motifs clos et fréquents. La recherche de MSE à partir de deux ensembles de motifs clos risque de fausser le calcul du taux d'émergence d'un motif en manquant des motifs inclus dans d'autres. En effet, pour calculer le taux de croissance d'un motif clos de D_c noté S, nous devons comparer la fréquence de S dans D_c avec sa fréquence dans D_s. Or, si nous avions fouillé les motifs clos de D_s, S aurait pu être inclus dans un motif clos de D_s. Pour être certains de ne pas manquer S dans les motifs clos de D_s, il aurait fallu recalculer tous les sous-ensembles contenus dans les motifs clos de D_s. Autrement dit, retrouver l'ensemble de motifs fréquents dans D_s. C'est pourquoi, nous avons directement fouillé les motifs fréquents de D_s et non ses motifs clos. Par exemple, pour calculer le taux de croissance du motifs clos $S_1 = \langle\{\text{a, b}\}\rangle$ de D_c, nous devons chercher sa fréquence dans D_c et dans D_s. Si nous avions également fouillé les motifs clos de D_s et considérant $S_2 = \langle\{\text{a, b}\}\{\text{c}\}\rangle$ et S_1 avec la même fréquence dans D_s, alors S_1 aurait été inclus dans S_2 et n'aurait pas été retrouvé dans D_s. En revanche, en fouillant les motifs fréquents de D_s nous comparons S_1 avec tous les motifs inclus dans S_2, soit : $\langle\{\text{a}\}\rangle$, $\langle\{\text{b}\}\rangle$, $\langle\{\text{c}\}\rangle$, $\langle\{\text{a, b}\}\rangle$, $\langle\{\text{a}\}\{\text{c}\}\rangle$, $\langle\{\text{b}\}\{\text{c}\}\rangle$ et $\langle\{\text{a, b}\}\{\text{c}\}\rangle$. Dès lors, S_1 est retrouvé dans D_s et nous pouvons

calculer son taux de croissance en comparant ses fréquences dans D_c et D_s.

5.1.2 Enjeux et problématiques des étapes de la chaîne de traitement

Chaque brique de traitement détermine la manière dont les MSE caractérisent les registres de langue. Cette section détaille justement comment ces briques déterminent linguistiquement les résultats de notre étude. Nous les présentons ci-dessous en suivant l'ordre dans lequel elles se succèdent (comme illustré par la figure 5.1).

Transformation des textes en séquences Trois types d'objets doivent être instanciés lorsque nous transformons des textes en séquences : la séquence, l'itemset et l'item. Dans le cadre de notre travail, cela revient à se demander : quel segment textuel est représenté par une séquence ? quelle unité textuelle est décrite par un itemset ? et quels traits linguistiques sont utilisés pour la décrire via des items ?

Séquence Tout d'abord, la segmentation d'un texte en séquences rend ses segments indépendants les uns des autres. De plus, les séquences obtenues deviennent les objets à partir desquels nous calculons le support d'un motif. En premier lieu, l'indépendance des segments entre eux a pour principale conséquence de perdre des informations linguistiques en supprimant du contexte. Par exemple, considérant le texte t « J'aime la mer, j'y vais souvent. Elle est apaisante, je m'y ressource. », le segmenter à l'échelle de la phrase coupe le lien entre le nom commun « mer » de la première phrase et le pronom personnel « Elle » commençant la seconde. En second lieu, la définition du segment textuel correspondant à une séquence détermine la granularité à laquelle la fréquence d'un objet est calculée. En effet, la fréquence ne renvoie pas au nombre de fois qu'un motif M apparaît dans D, mais au nombre de fois que M est supporté par une séquence de D. Autrement dit, si M apparaît dans deux séquences stockées dans D, deux fois dans la première et quatre fois dans la seconde, la fréquence de M n'est pas calculée en comptabilisant six apparitions de M mais deux, car M est supporté par deux séquences. Ainsi, plus la séquence représente un grand segment textuel, plus la fréquence d'un motif donne une image large de sa présence dans D ; à l'inverse plus la séquence est petite, plus la fréquence donne une image détaillée de sa présence. Par exemple, considérant le texte t, si nous décidons qu'un virgulot [2]

2. Un virgulot est un segment textuel dont les limites sont les signes de ponctuation suivants : des points, des virgules ou bien des points virgules.

ID	Séquences
1	⟨{lemme : je}, {lemme : '}, {lemme : aimer}, {lemme : la}, {lemme : mer}, {lemme : ,}⟩
2	⟨{lemme : je}, {lemme : '}, {lemme : y}, {lemme : aller}, {lemme : souvent}, {lemme : .}⟩
3	⟨{lemme : elle}, {lemme : être}, {lemme : apaisant}, {lemme : ,}⟩
4	⟨{lemme : je}, {lemme : me}, {lemme : '}, {lemme : y}, {lemme : ressourcer}, {lemme : .}⟩

TABLE 5.1 – Exemple de base de données D_v contenant quatre séquences à partir d'un texte segmenté en quatre virgulots : chaque mot est décrit par un itemset contenant le lemme du mot.

représente une séquence et que chaque mot est représenté par un itemset composé d'un item donnant le lemme du mot, nous obtenons la base de données D_v illustrée table 5.1. Elle contient les quatre segments textuels, « J'aime la mer, » « j'y vais souvent. » « Elle est apaisante, » « je m'y ressource. », sous la forme de quatre séquences. La fréquence du motif $M = \langle\{\text{lemme : je}\}\rangle$ dans D_v est de 75%, car M est supporté par trois séquences sur quatre. En revanche, si nous segmentons le texte au niveau de la phrase, une séquence représente alors une phrase où chaque itemset représente un mot avec son lemme. Nous obtenons deux séquences stockées dans une base de données D_p. La fréquence de M dans D_p est plus élevée que dans D_v puisqu'elle est de 100% : M est supporté par toutes les séquences de D_p. Ces exemples montrent comment le choix du segment textuel représenté par une séquence détermine les motifs contenus dans l'ensemble de MSE : il modifie les valeurs de la fréquence d'un motif, dont dépend son taux de croissance, à partir duquel un motif est jugé émergent ou non.

Itemset Lorsque nous choisissons l'objet représenté par un itemset, nous choisissons quelle unité textuelle minimale nous voulons décrire via un ensemble de facteurs. Dans le cas de notre étude, l'unité textuelle peut être une lettre, une syllabe, un mot ou encore une suite de mots [3]. Plus l'unité textuelle est petite, plus les unités apparaissant fréquemment ensembles seront nombreuses. À l'inverse, plus elle est grande, plus les unités apparaissant fréquemment ensembles seront rares. Par exemple, si nous considérons une base de données D_l dérivée de D_v, où les itemsets décrivent chaque lettre et non plus chaque mot, en précisant s'il s'agit d'une voyelle ou d'une consomme, alors le motif $M = \langle\{\text{voyelle}\}\{\text{voyelle}\}\{\text{consomme}\}\rangle$ de trois itemsets a une fréquence de 100% dans D_l, car M est supporté par les quatre séquences de D_l. En revanche, aucun motif de trois itemsets n'a une fréquence de 100% dans D_v, car aucune suite de trois lemmes n'est com-

3. Un mot est défini comme un "*Son ou groupe de sons articulés ou figurés graphiquement, constituant une unité porteuse de signification à laquelle est liée, dans une langue donnée, une représentation d'un être, d'un objet, d'un concept, etc*" - `https://www.cnrtl.fr/definition/mot`.

mune aux quatre séquences de D_v. Dans le cadre de nos deux études présentées dans ce chapitre, nous avons choisi le mot et la ponctuation comme unités textuelles décrites par un itemset. Nous avons fait ce choix car :

- il limite le nombre de MSE découverts ;
- nous nous intéressons principalement aux relations des mots entre eux pour caractériser les registres de langue ;
- les outils automatiques nous permettant d'obtenir des traits linguistiques, tels que les fonctions syntaxiques ou grammaticales, sont entraînés à étiqueter des mots.

Item Enfin, les items contenus dans les itemsets illustrent les facteurs décrivant l'unité textuelle représentée par l'itemset. En d'autres termes, les items portent les traits linguistiques avec lesquels nous voulons décrire l'unité textuelle minimale considérée, en l'occurrence les mots et la ponctuation. Le choix des traits linguistiques est une question centrale dans nos travaux, car ils constituent les informations à partir desquelles les MSE seront découverts. Selon le cas d'usage et l'objectif de l'étude le choix des valeurs des items diffère. Dans notre cas, nous avons représenté chaque unité linguistique avec des traits linguistiques appartenant à divers niveaux d'analyse de la langue car nous supposons que le croisement de ces niveaux apporte une information plus robuste pour caractériser les registres de langue que s'ils étaient considérés séparément. Ces traits linguistiques sont les mêmes pour toutes les unités décrites.

Fouille de motifs séquentiels clos Les motifs clos sont fouillés sous deux contraintes : $minsup_c$ et gap_c. $minsup_c$ est un seuil de fréquence minimale : tous les motifs dont la fréquence est inférieure à $minsup_c$ sont élagués de l'espace de recherche. Une valeur élevée de $minsup_c$ a l'avantage de faire diminuer le nombre de motifs clos découverts, mais elle a le désavantage de nous faire manquer des motifs clos intéressants avec une fréquence plus faible. Concrètement, lorsque nous fixons la valeur de $minsup_c$, nous décidons à partir de quelle fréquence un motif peut caractériser un registre cible : doit-il être présent dans au moins 1% de la base de données ? ou bien doit-il être présent dans au moins 25% des séquences ? Dans nos travaux, nous avons donné une faible valeur à $minsup_c$ car nous avons préféré avoir une grande quantité de motifs clos découverts que de prendre le risque de manquer des motifs intéressants. La seconde contrainte est celle de gap, notée gap_c, pour laquelle l'utilisateur doit fixer deux valeurs : le nombre minimal et le nombre maximal d'itemsets inconnus entre deux itemsets connus. Nous avons utilisé gap_c pour ne chercher

que des motifs dont les itemsets sont contiguës, c'est-à-dire des descripteurs linguistiques dont les mots sont tous consécutifs. Cela contrebalance la quantité de motifs clos découverts en concentrant la fouille sur des motifs facilement interprétables car composés de séquences ininterrompues de mots.

Fouille de motifs séquentiels fréquents La fouille de motifs fréquents est également faite sous deux contraintes : $minsup_s$ et gap_s. Elles sont des mêmes types que $minsup_c$ et gap_c, mais leurs valeurs peuvent différer, notamment parce que la valeur de $minsup_s$ peut assurer que les motifs fréquents auxquels seront comparés les motifs clos sont moins présents dans le registre source que dans le registre cible. Quant à gap_c, c'est une contrainte qui garantie la découverte de motifs fréquents dont les itemsets sont contiguës afin de pouvoir les comparer avec les motifs clos.

Fouille de motifs séquentiels émergents Enfin, nous fouillons les MSE en considérant un seuil sur le taux de croissance des motifs noté ρ. Pour qu'un motif soit considéré comme émergent il faut que son taux de croissance soit supérieur à ρ. Plus la valeur de ρ est élevée, plus les motifs considérés comme émergents sont très caractéristiques du registre cible et peu nombreux. Comme pour les valeurs de minsup des étapes précédentes, nous avons fixé la plus petite valeur possible à ρ, c'est-à-dire 1, en préférant réduire le risque de manquer des MSE intéressants avec un faible taux de croissance que de limiter la quantité de MSE.

Nous avons vu dans cette section les conséquences des briques de traitement sur les MSE découverts en fin de chaîne. Lors de nos expériences, notre priorité a été de ne pas manquer de MSE intéressants, dès lors tous les seuils lors des fouilles de motifs clos, fréquents et émergents ont des valeurs basses. Cela a pour conséquence de découvrir un grand nombre de MSE. Pour avoir confiance en l'intérêt de tous ces MSE découverts, nous avons mené une preuve de concept présentée dans la section suivante.

5.2 Preuve de concept à partir de langages artificiels

Ce qui rend difficile la vérification de l'intérêt des MSE découverts c'est l'absence de base de vérité avec laquelle les comparer, c'est-à-dire une liste de motifs que l'utilisateur sait intéressants pour son cas d'usage. Or, les utilisateurs emploient généralement la fouille de MSE pour dégager des connaissances d'un jeux de données en découvrant des

MSE inattendus. Par conséquent, ils n'ont pas de connaissances *a priori* sur les MSE à découvrir et donc de base de vérité permettant d'en vérifier la pertinence. Pour répondre à cette problématique, nous avons créé un cadre artificiel dans lequel nous connaissions les motifs à retrouver. Avec ce cadre, nous avons testé la capacité de la fouille de MSE à retrouver des motifs que nous savions présents dans les textes explorés. Nous présentons dans cette section les travaux menés pour construire une base de vérité, puis nous décrivons le protocole expérimental mis en place pour fouiller les MSE. Ces derniers sont ensuite comparés à la base de vérité dans la dernière partie de cette section.

5.2.1 Construction de la base de vérité et du corpus de langages artificiels

Nous voulons prouver que la fouille de motifs découvre des motifs pertinents pour caractériser un registre de langue. Pour cela nous avons construit une base de vérité composée de descripteurs linguistiques que nous voulions retrouver dans les motifs. Mais, comment choisir des descripteurs constituant la base de vérité pour qu'ils prouvent la capacité des motifs à caractériser un registre à divers niveaux d'analyse de la langue ? Et comment être certains que les textes fouillés contiendraient les motifs de la base de vérité à retrouver ? Tout d'abord, nous expliquons comment nous avons choisi ces motifs à retrouver. Puis, nous détaillons de quelle manière nous avons généré des langages artificiels les contenant.

Choix des motifs Nous voulons vérifier que les MSE contiennent uniquement des motifs caractéristiques du registre cible sans motif caractéristique du registre source et qu'ils sont capables de traiter divers niveaux d'analyse de la langue. Pour cela, nous avons constitué deux ensembles de motifs pour la base de vérité (1) un ensemble M_c contenant les motifs du registre cible et (2) un ensemble M_s contenant ceux du registre source. Tous les motifs de M_c et M_s doivent contenir des formes lexicales, des formes morphologiques, des constructions morphosyntaxiques et syntaxiques particulières pour vérifier la capacité des MSE à traiter ces niveaux d'analyse de la langue. Nous avons choisi des descripteurs linguistiques validés dans (Mekki, Battistelli, Lecorvé et al. 2018 comme distinguant les registres entre eux et appartenant à ces niveaux d'analyse. Parmi eux, citons : la contraction de *cela* en *ça* identifiée par (Gadet 1997), l'utilisation de *ça* en position sujet (Mekki, Battistelli, Lecorvé et al. 2018), le sujet *nous* transposé en *on*

ID	M_c Registre familier	M_s Registre soutenu
1	(*on* | *ça*) + VERBE	(*il* | *elle*) + VERBE
2	∅ ... *pas*	*ne* ... *pas*
3	SUJET + VERBE + ?	VERBE + SUJET + ?
4	(*radoter* | *chanter*)	(*chanter* | *répéter*)
5	(*chanson* | *chansonnette* | *musique*)	(*romance* | *ballade*)

TABLE 5.2 – Exemples de descripteurs linguistiques introduits dans les langages artificiels pour les registres familier et soutenu.

(BILGER et CAPPEAU 2004), la contraction de la négation complète ou partielle (BILGER et CAPPEAU 2004) et l'inversion des places du sujet et du verbe opérée ou non opérée dans une phrase interrogative (GADET 2003). À ces descripteurs, nous avons ajouté un lexique dont chaque mot est associé à un registre de langue. La table 5.2 donne des exemples pour les cinq descripteurs énumérés ci-dessus, la table montre leur répartition dans M_c et M_s. Le symbole « | » signifie « ou ». Chaque ligne correspond à un descripteur linguistique, la première colonne donne son identifiant, la deuxième sa forme pour le registre familier et la troisième sa forme pour le registre soutenu. Cela veut dire que nous nous attendons à ce que les MSE découvrent des MS décrivant les mêmes descripteurs que ceux de la table 5.2. Par exemple, pour valider la fouille de MSE caractéristiques du registre cible familier, nous devons retrouver dans les résultats tous les MS correspondant aux motifs de M_c sans aucun MS correspondant aux motifs de M_s. Afin d'être certains que les textes explorés possèdent ces motifs, nous avons généré des langages artificiels les contenant.

Génération des langages artificiels Nous avons utilisé des grammaires hors-contexte probabilistes implémentées en python pour générer des langages artificiels comprenant des descripteurs linguistiques choisis en amont (M_c et M_s) plus ou moins présents selon les registres de langue. Par exemple le descripteur « *ça* + verbe » sera plus souvent présent dans les textes du registre familier que ceux du registre soutenu. Dans nos expériences nous avons considéré les registres familier et soutenu. Pour chacun d'eux, nous avons implémenté une grammaire hors-contexte probabiliste [4] dont les règles génèrent 1 000 phrases les unes après les autres. Les trois phrases suivantes sont des exemples du langage artificiel du registre familier : « ça répète cette chanson ? », « on chante la chansonette ? » et « il radote une fable » (d'autres exemples sont donnés en annexes IV). Chaque mot a été éti-

4. Les trois grammaires utilisées sont données en annexes III.

queté, avec son lemme, sa catégorie morphosyntaxique, et sa fonction syntaxique [5]. Nous avons choisi la phrase comme segment textuel constituant une séquence, car les données ont été générées phrase par phrase. La section suivante expose comment les MSE ont été fouillés à partir de ce corpus de langages artificiels.

5.2.2 Fouille des motifs séquentiels émergents

Nous avons vérifié que les MSE attendus se retrouvaient bien dans les résultats. Pour cela nous avons suivi la chaîne de traitement présentée section 5.1. Nous présentons ici le protocole expérimental mis en place pour fouiller les MSE, puis les résultats obtenus.

Protocole expérimental pour la fouille Considérant un registre de langue cible R_c et un registre de langue source R_s, les paramètres fixés pour les extractions des motifs clos et fréquents sont les suivants :

- le $minsup_c$ pour l'extraction des motifs fréquents et clos de R_c est de 5% ;
- le $minsup_s$ pour l'extraction des motifs fréquents et clos de R_s est de 2,5% ;
- le seuil ρ est fixé à 1 ;
- les contraintes gap_c et gap_s sont toutes les deux de $P[1, 1]$.

Ici, $minsup_s$ représente la moitié de $minsup_c$, afin d'assurer que les motifs comparés étaient au minimum deux fois moins présents dans le registre R_s. Tous les motifs dont le taux de croissance est strictement supérieur à ρ est considéré comme un MSE. Au total deux fouilles de MSE ont été faites : la première a pour R_c le registre familier et R_s le registre soutenu, la seconde inverse les registres et a pour R_c le registre soutenu et R_s le registre familier. Pour fouiller les motifs clos et fréquents, nous avons utilisé l'algorithme *CloSpec* de (Béchet et al. 2015) qui est détaillé en section 4.2.2.

Protocole d'évaluation Nous avons évalué les MSE en adoptant un paradigme de recherche d'informations où les MSE sont jugés *pertinents* lorsqu'ils appartiennent à M_c et *non-pertinents* lorsqu'ils appartiennent à M_s, leur taux de croissance est considéré comme le rang auquel ils ont été retournés. Nous avons utilisé trois mesures d'évaluation issues de la recherche d'informations : AUROC (Narkhede 2018), AP (Kishida 2005) et NDCG (McSherry et Najork 2008). Ces trois mesures vérifient que les éléments pertinents

5. L'outil automatique utilisé est TreeTagger : `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`.

R_c	R_s	AUROC	AP	NDGC
Familier	Soutenu	0.95	0.91	0.99
Soutenu	Familier	0.99	0.96	0.99

TABLE 5.3 – Résultats des extractions de MSE (familier par rapport à soutenu) extraits de (MEKKI, BÉCHET et al. 2020)

ont bien été tous retrouvés et qu'ils sont retournés avant les éléments non-pertinents. Plus leurs valeurs sont hautes, meilleurs sont les résultats.

Évaluation des résultats La table 5.3 présente les résultats, obtenus lors des deux fouilles de MSE, évalués avec les trois métriques AUROC, AP et NDGC. Leurs valeurs hautes montrent que :

- la grande majorité des motifs de M_c est retrouvée dans les MSE ;
- presque aucun motif de M_s n'est présent dans les MSE ;
- le taux de croissance fait bien émerger les motifs de M_c en tête des résultats avec peu de motifs de M_c ayant un taux de croissance inférieur ou égal à ρ.

Ces résultats valident notre hypothèse selon laquelle les MSE sont pertinents pour caractériser un registre de langue.

Détails des résultats Pour avoir une idée de la complexité algorithmique causée par la recherche des motifs fréquents et clos, nous avons regardé lors des expériences la quantité de mémoire vive utilisée, le poids des fichiers de résultats, le temps nécessaire à l'algorithme pour fouiller les données et le nombre de motifs extraits. La table 5.4 donne les détails des expériences. Chaque ligne correspond à une fouille de motifs distincte et les colonnes précisent le registre de langue, le type de motifs, la quantité de mémoire utilisée, le poids du fichier de résultats, la durée de la fouille de MSE et le nombre de motifs découverts. Nous rappelons que la base de données explorée contient 1 000 séquences pour en moyenne cinq itemsets possédant tous quatre items. Les résultats présentés par la table 5.4 montrent déjà une quantité importante de motifs extraits, et une quantité de mémoire vive utilisée notable. Notons également, qu'il semble y avoir une forte redondance dans les motifs car il y a beaucoup moins de motifs clos que de motifs fréquents. Cela nous informe qu'il y a beaucoup de motifs inclus dans un autre motif ayant la même fréquence. Dans cette section nous avons montré que notre approche employant la fouille de MSE était viable pour la caractérisation des registres de langue. La section suivante présente

Registre	Type de motifs	Mémoire vive utilisée	Poids du fichier de résultats	Durée de la fouille de motifs	Nombre de motifs extraits
Familier	fréquent	0 GO	1 GO	2 mn	63 066
	clos	4 GO	50 KO	2,3 mn	1 122
Soutenu	fréquent	0 GO	8 GO	13 mn	113 746 190
	clos	18 GO	78 KO	8 mn	1 545

TABLE 5.4 – Détails des fouilles de motifs fréquents et clos à partir des langages artificiels des registres familier et soutenu.

l'application de cette approche à un corpus de données réelles, le corpus TREMoLo-Tweets.

5.3 Fouille de motifs séquentiels émergents à partir du corpus TREMoLo-Tweets

L'objectif de cette seconde fouille de MSE à partir de données réelles est de découvrir des descripteurs linguistiques caractéristiques des registres de langue illustrant l'usage réel des registres. En outre, les tweets n'ayant pas été analysés dans la littérature sous le prisme des registres, nous espérons trouver des descripteurs propres à ce genre de textes. Le principal enjeu de ces travaux est de contraindre le moins possible la fouille de MSE pour ne pas manquer de MSE intéressants, tout en explorant un large jeux de données puisque TREMoLo-Tweets compte 228 270 tweets. Pour cela, nous avons cherché à restreindre le coût combinatoire en choisissant un nombre limité d'items fixes pour décrire les unités textuelles, mais en donnant le plus d'informations possible sur ces unités avec les choix de traits linguistiques utilisés. Nos choix pour transformer les tweets en séquences sont introduits dans la première partie de cette section, la suivante expose les expériences découvrant l'ensemble de MSE. Enfin, les résultats obtenus sont détaillés en fin de section.

5.3.1 Transformation des tweets en séquences

Nous détaillons ici les données réelles utilisées et leur transformation en séquences. Les choix faits se sont adaptés aux particularités des tweets comme son format court ou bien ses technomorphèmes, tout en prenant en compte de nouveaux usages linguistiques comme la disparition partielle des signes de ponctuations classiques.

Sous-corpus	Nombre de tweets
Familier	69 797
Courant	134 249
Soutenu	24 224
Total	228 270

TABLE 5.5 – Le nombre de tweets par sous-corpus de registre de langue.

Corpus de tweets Nous avons fouillé les MSE à partir du corpus TREMoLo-Tweets présenté plus en détails section 3.4. Une de nos contributions lors de sa constitution est d'avoir intégré à notre annotation manuelle des marqueurs linguistiques propres à ce type de textes, tels que l'utilisation des pictogrammes comme du lexique ou l'intégration syntaxique des hashtags. Les annotateurs les ont associés aux registres qu'ils attribuaient à un tweet. Nous supposons que le classifieur de type CamemBERT (MARTIN, Benjamin MULLER, ORTIZ SUÁREZ et al. 2020), qui a généralisé l'annotation manuelle à l'ensemble des tweets, a également associé ces marqueurs à des registres de langue. C'est pourquoi nous pensons les retrouver dans les MSE. Le corpus étiqueté automatiquement résultant de ces travaux compte 228 270 tweets pour un total de 6 millions de mots. La table 5.5 détaille le nombre de tweets par registre de langue [6]. Les sous-corpus sont déséquilibrés puisque le registre courant est représenté par un sous-corpus contenant 5,5 fois plus de tweets que celui du registre soutenu et 2 fois plus de tweets que celui du registre familier. Or, comme la fréquence d'un motif est calculée en divisant son support par le nombre de séquences de la base de données dans laquelle il est stocké : moins il y a de séquences dans la base de données, plus il y a de motifs fréquents ; à l'inverse plus il y a de séquences dans la base de données, moins il y a de motifs fréquents. Aussi, nous aurons plus de motifs fréquents et clos pour le registre soutenu, puis pour le registre familier et enfin pour le registre courant. Les paragraphes suivants exposent comment nous avons transformé ces données textuelles en séquences.

Séquence Nous avons choisi de représenter un tweet entier par une séquence car l'absence de ponctuation classique dans certains tweets rend l'unité de la phrase trop instable.

(64) #MonPireDate j'ai pleuré parce que je pensais à mon ex 😅 Désolé si tu vois ça 😂

(65) jamais tranquille c un truc de malade même contre brest

6. Un tweet est considéré comme appartenant à un registre lorsque sa probabilité d'appartenance au registre est supérieure à 0,50.

L'exemple (64) montre que l'utilisateur a utilisé un pictogramme pour terminer sa première phrase. Quant à l'exemple (65), il donne un tweet où il n'y a aucun signe de ponctuation. De plus, les tailles des tweets sont homogénéisées grâce à la limite imposée par Twitter autorisant un maximum de 240 caractères par tweet. Les séquences ont donc également un nombre d'itemsets homogène.

Itemset Chaque itemset représente soit un mot, soit un hashtag, soit un pictogramme, soit un signe de ponctuation. Nous avons motivé notre choix du mot comme unité textuelle dans la section 5.1.2 présentant notre chaîne de traitement complète.

Items Pour limiter la complexité algorithmique lors de la fouille des motifs fréquents et clos, nous devons choisir un nombre restreint d'items par itemset. En choisissant les types de traits linguistiques portés par ces items, nous avons cherché à donner une description complète du mot avec un minimum de traits. Pour cela, nous avons sélectionné des traits de niveaux d'analyse de la langue différents car leur croisement permet d'apporter des informations sur le contexte du mot à l'utilisateur. Si la fonction syntaxique d'un mot se terminant par « z » indique qu'il est la racine d'une phrase, alors ce mot a plus de probabilité d'être un verbe utilisé avec le vouvoiement et donc d'appartenir au registre soutenu. En revanche, si sa catégorie grammaticale indique qu'il est un adverbe, alors le mot peut être un néologisme appartenant au registre familier comme « pépouz », « zonz » ou encore « zouz ». Au total nous avons choisi cinq traits linguistiques fixes pour chaque mot, c'est-à-dire cinq items pour chaque itemset. Pour le niveau morphologique, nous avons considéré deux traits linguistiques :

1. les unités syllabiques composant le mot appelées *sous-mots* ou *word pieces* en anglais, lorsque le sous-mot est précédé de « _ » c'est qu'il commence le mot, lorsqu'il est suivi de « _ » c'est qu'il termine le mot ;
   ex : _*chant* et *ons*_
2. les caractéristiques morphologiques d'un mot données par son genre, son nombre, son temps verbal.
   ex : présent de l'indicatif au pluriel masculin.

Pour le niveau lexical, le lemme du mot est utilisé, c'est-à-dire sa forme non fléchie (par exemple, le lemme de *chantons* est *chanter*). Ensuite, le niveau morphosyntaxique est décrit via la catégorie grammaticale du mot précisant s'il est un pronom personnel, un déterminant, un verbe, un adjectif, un adverbe, etc. Enfin, le dernier niveau considéré

est le niveau syntaxique qui décrit la fonction syntaxique du mot précisant s'il est sujet, racine, objet, etc. Tous les traits, à l'exception des sous-mots, ont été obtenus par l'outil d'annotation automatique Talismane [7] (Urieli et Tanguy 2013). Les sous-mots quant à eux ont été obtenus grâce à l'outil de segmentation proposé par CamemBERT [8]. En se basant sur ces choix de traits linguistiques, une phrase comme :

"*Les fillettes dorment.*"

a été transposée en une séquence de 4 itemsets :

⟨ (*lemme:les*, *pos:det*, *morpho:pluriel*, *syntax:det*, *sous-mot:_le*, *sous-mot:s_*),
(*lemme:fillette*, *pos:nom-commun*, *morpho:féminin*, *syntax:sujet*, *sous-mot:_fille*, *sous-mot:ttes_*),
(*lemme:dormir*, *pos:verbe*, *morpho:present-ind pluriel*, *syntax:racine*, *sous-mot:_dorm*, *sous-mot:_ent*),
(*lemme:.*, *pos:ponctuation*, *syntax:ponctuation*, *sous-mot:_._*) ⟩.

Dans cet exemple, le symbole *lemme* précède le lemme du mot, *pos* sa catégorie grammaticale [9], *morpho* ses caractéristiques morphologiques, *syntax* sa fonction syntaxique et *sous-mots* ses sous-mots. Chaque tweet est transformé en séquence de cette manière avant d'être stocké dans une base de données, comme il y a 228 270 tweets la base de données contient 228 270 séquences qui seront explorées pour découvrir l'ensemble complet de MSE.

5.3.2 Fouille des motifs séquentiels émergents

Le protocole expérimental mis en place pour la fouille de MSE est introduit, avant de détailler quantitativement les résultats obtenus. Nous concluons cette section en explorant linguistiquement les MSE découverts. Les expériences conduites valident la capacité des MSE à caractériser un registre de langue cible.

5.3.2.1 Protocole expérimental

Comme pour la première expérience présentée section 5.2, la fouille de MSE caractérise un registre cible R_c en partant d'un registre source R_s. Au total, six fouilles de MSE ont été faites car nous avons considéré six paires de R_c et R_s différentes. La table 5.6 donne

7. La présentation de l'outil est disponible via ce lien `https://github.com/joliciel-informatique/talismane/wiki`.
8. L'outil est accessible via ce lien `https://huggingface.co/camembert-base`.
9. pour Part Of Speech en anglais

ID	R_c	R_s
1	Familier	Courant
2		Soutenu
3	Courant	Familier
4		Soutenu
5	Soutenu	Familier
6		Courant

TABLE 5.6 – Listes des six couples de registres considérés dans nos expériences.

Registre	**Durée de la fouille de MS**		**Nombre de MS découverts**	
	MS fréquents	**MS clos**	**MS fréquents**	**MS clos**
Familier	0h 38mn	5h 33mn	10 006 362	2 341 661
Courant	0h 54mn	10h	9 478 176	2 735 775
Soutenu	2h 00mn	34h	135 061 353	8 895 962

TABLE 5.7 – Durée de la fouille de motifs fréquents et clos pour les trois registres de langue avec le nombre de motifs découverts pour chacun d'eux. Les trois fouilles sont faites avec un *minsup* à 1% et une contrainte de gap à $P[1, 1]$.

les six paires en précisant quels registres de langue sont R_c et R_s. Nous avons utilisé l'algorithme CloSpec (BÉCHET et al. 2015) pour toutes nos expériences. Notre priorité étant de ne pas manquer de motifs intéressants, nous n'avons pas fortement contraint les fouilles de motifs fréquents et clos durant les expériences :

- le $minsup_c$ et $minsup_s$ ont une valeur de 1% pour conserver des motifs fréquents et clos avec de petites fréquences ;
- le seuil sur le taux de croissance ρ est fixé à 1 afin d'obtenir l'ensemble des MSE même faiblement émergents ;
- la contrainte de *gap* est fixée à $P[1, 1]$ pour obtenir des motifs dont les itemsets sont contiguës.

Le fait que les six expériences aient abouti est une première validation de notre stratégie consistant à choisir un nombre restreint d'items pour contrebalancer de faibles contraintes sur la fouille de motifs fréquents et clos.

5.3.2.2 Détails quantitatifs des résultats

Puisque nous avons privilégié le fait de pas rater de motifs intéressants pour caractériser les registres de langue, nous avons obtenu une grande quantité de résultats. Cette

R_c	R_s	Nombre de MSE	Nombre de motifs dont le $TC = +\infty$	Pourcentage de motifs dont le $TC = +\infty$
Familier	Courant	326 552	11 542,	3,53 %
	Soutenu	226 938	69 374	30,57 %
Courant	Familier	416 554	36 516	8,77 %
	Soutenu	61 121	5 491	8,98 %
Soutenu	Familier	2 330 679	1 668 317	71,58 %
	Courant	2 356 624	1 091 392	46,31 %

TABLE 5.8 – Proportions des MS dont le taux de croissance TC est égal à ∞ pour les six paires de registres.

section donne une idée d'ensemble de ces résultats d'un point de vue quantitatif. La table 5.7 détaille les temps des extractions des ensembles de motifs fréquents et clos, pour les trois registres, ainsi que le nombre de motifs extraits. Les différences d'échelle entre les motifs fréquents et les motifs clos, déjà présentes dans les expériences exposées section 5.2.2, sont confirmées avec les fouilles à partir du corpus TREMoLo-Tweets. En effet, la fouille des motifs fréquents pour les trois registres a systématiquement retourné un ensemble de séquences plus important, que celui des motifs clos. À l'inverse, les durées d'extraction ont toujours été plus importantes pour les motifs clos que pour les motifs fréquents. Dès lors, la fouille des motifs fréquents a eu pour principal enjeu la réduction du nombre de motifs découverts tandis que la fouille des motifs clos celui d'en réduire les coûts matériels. En outre, pour le registre soutenu, les ensembles des motifs fréquents et clos sont significativement plus importants que ceux des registres familier ou courant. De fait, l'ensemble des motifs fréquents pour le registre soutenu est en moyenne 13 fois plus important que pour les deux autres registres. Toujours pour le registre soutenu, l'ensemble des motifs clos est en moyenne trois fois plus important que pour les registres familier et courant. Ces écarts d'échelle peuvent être expliqués par l'effet statistique provoqué par l'écart de taille entre les bases de données séquentielles à partir desquelles sont découverts les motifs fréquents et clos. En effet, le jeu de données séquentielles pour le registre soutenu est trois et cinq fois plus petit que, respectivement, celui du registre familier et courant. Ainsi, du fait d'un ensemble de séquence plus petit, les motifs pour le registre soutenu fréquents ou clos ont été plus nombreux.

La table 5.8 donne les proportions de MSE dont le taux de croissance est infini. Chaque ligne de la table représente une paire de registres différente. La différence de quantité entre les ensembles de motifs fréquents ou clos, pour le registre soutenu et les autres

registres, s'est retrouvée dans les différences de volume des MSE extraits. Effectivement, un écart significatif s'est creusé entre le nombre de MSE extraits pour les paires du registre soutenu et ceux des autres paires de registres. Les MSE pour le registre soutenu sont en moyenne neuf fois plus nombreux que les MSE des autres paires de registres. Afin d'expliquer cette différence, nous avons supposé qu'il existe un contraste plus important entre les formes linguistiques du registre soutenu et celles des registres familier et courant. À l'inverse, les séquences des registres familier et courant ont plus de similitudes entre elles. Les pourcentages de séquences dont le taux de croissance est infini ont confirmé cette hypothèse. Nous rappelons qu'un MSE caractéristique de R_c avec un taux de croissance infini est un motif fréquent dans R_c et non fréquent dans R_s. Les importantes proportions de séquences avec un taux de croissance infini révèlent un large ensemble de descripteurs linguistiques présents dans les textes du registre soutenu, mais presque absents des textes du registre familier ou courant, c'est-à-dire dont la fréquence est inférieure à 1% dans les jeux de données de ces deux registres. Ainsi, l'ensemble des MSE contenant le plus de MSE avec un taux de croissance infini est celui du registre cible soutenu par rapport au registre source familier.

5.3.2.3 Détails qualitatifs des résultats

Dans cette section nous explorons manuellement les ensembles de MSE découverts pour en montrer la qualité. Au regard de la quantité de MSE, cette exploration manuelle a représenté un travail important et chronophage soulignant la nécessité d'un traitement automatique pour obtenir un ensemble de MSE plus facilement exploitable. Nous retrouvons dans les MSE découverts des descripteurs linguistiques propres aux tweets utilisés lors de la tâche d'annotation manuelle du sous-corpus de TREMoLo-Tweets. Cela nous informe que la fouille de MSE a retrouvé les descripteurs linguistiques utilisés par les annotateurs experts pour différencier les registres de langue. Aussi, bien que cette exploration manuelle des résultats soit partielle, elle fortifie notre confiance dans la pertinence des MSE pour distinguer les registres entre eux.

Deux paires de registres sont présentées : le R_c familier par rapport au R_s courant et le R_c familier par rapport au R_s soutenu. Nous avons choisi ces couples de registres car le premier explore des MSE issus de registres proches, tandis que le second regarde des MSE issus de registres éloignés.

ID	MSE	*TC*
1	⟨ (*syntax:modifieur*) , (*sous-mot:_j*) ⟩	$+\infty$
2	⟨ (*sous-mot:_t*), (*pos:verbe*) ⟩	$+\infty$
3	⟨ (*sous-mot:_toi_*) ⟩	$+\infty$
4	⟨ (*pos:adjectif, lemme:gros*) ⟩	$+\infty$
5	⟨ (*pos:nom-propre*) , (*pos:clitique-sujet, morpho:3eme personne*), (*pos:verbe, morpho:présent*) ⟩	$+\infty$
6	⟨ (*sous-mot:_c_*) ⟩	$+\infty$
7	⟨ (*lemme:rajouter*) ⟩	$+\infty$
8	⟨ (*sous-mot:sh_*) ⟩	$+\infty$
9	⟨ (*sous-mot:rrr_*) ⟩	$+\infty$
10	⟨ (*syntax:modifieur*), (*lemme:pictogramme*) (*lemme:pictogramme*) ⟩	1,52

TABLE 5.9 – Exemples de MSE caractéristiques du familier vs. courant.

Familier vs. courant Si l'on trie les MSE par ordre décroissant de taux de croissance, tous les MSE dont les taux de croissance sont $+\infty$ sont *ex aequo* à la première place. La table 5.9 présente 10 exemples de ces MSE qui caractérisent le registre familier par rapport au registre courant. Pour chacun des motifs de la table 5.9, des exemples de tweets issus du corpus TREMoLo-Tweets sont présentés ci-dessous. Les motifs 1, 2 et 3 mettent en exergue l'utilisation du premier et second pronoms personnels singuliers. Ils semblent renforcer la tendance déjà observée lors de l'exploration linguistique du corpus TREMoLo-Tweets où le registre familier apparaît être utilisé dans des tweets dont le but est conversationnel. Les tweets (66) à (68) constituent des exemples du motif 1 : ⟨ (*syntax:modifieur*), (*sous-mot:_j*) ⟩.

(66) @X super entraineur laurent blanc gros j'espère il ira jamais chez vous

(67) @X Ouais je suis un peu déçue mais bon si y'a un désistement je suis prioritaire et sinon bah en terminale j'irai en stmg

(68) @X : Mdrrr j'avais encore jamais vu un seul GP fais pas genre j'suis un anti Gasly tu sais très bien ce que je pense 😂

Les tweets (69) à (71) constituent des exemples du motif 2 : ⟨ (*sous-mot:_t*), (*pos:verbe*) ⟩.

(69) @X t as dit je suis choqué par suarez ... il etait top 3 avc cr7 et messi pdt 5 ans

(70) @X Pierre t inquiète les gens sont méchant reste comme tu es

(71) @X Bah ouais, jsais pas ce que tu cherche mdrr t es chelou. Jvais pas dire que Arsenal et liverpool sont au même niveau. Donc logique que en termes de performance Arsenal soit une « petite équipe » par rapport à liverpool

Les tweets (72) à (74) constituent des exemples du motif 3 : ⟨ (*sous-mot:_toi_*) ⟩.

(72) @X @X @X Il éduque sanson t'es un fou toi mdr

(73) Westbrook il est en train de faire une rondo 😔 😔 réveil toi bro

(74) @X Les arbitres avaient miser bucks comme toi bg

Le motif 4 renvoie à des unités discursives appelées « *marqueurs discursifs (MD)* » par (Dostie et Pusch 2007). Ce sont des éléments de langages qui ponctuent des échanges habituellement oraux. Plus précisément, ils sont définis par (Dostie et Pusch 2007) comme suit :

> « Ils [les MD] ne contribuent pas au contenu propositionnel des énoncés et c'est pourquoi leur présence ou leur absence ne modifie pas la valeur de vérité des énoncés auxquels ils sont joints. Ils ont tendance à constituer des unités prosodiques indépendantes, si bien qu'ils sont en général extérieurs à la structure de la phrase. Ils sont optionnels sur le plan syntaxique, c'est-à-dire que, dans les cas où ils sont joints à un énoncé, leur absence n'entraîne pas une agrammaticalité. »

Les tweets (75) à (77) constituent des exemples du motif 4 : ⟨ (*pos:adjectif, lemme:gros*) ⟩

(75) @X @X @X en gros hier sur fall guys coro a dit que la K corp perderait aujourd hui et qu'il avait lancer une malédiction mdr il a continuer le troll même pendant les games de KCorp voila

(76) @X @X Maroua doit des dettes à Kihou che pas quoi et Maroua paye pas son loyer et traître Kihou voilà en gros

(77) @X en gros c'est les fandom qui ont décidé de s'allier pour faire perdre bts en faisant gagner astro j'crois ? ? et ils ont aussi fait planter le site pour que les army puissent plus voter

Les motifs 5 et 6, quant à eux, confirment la pertinence des descripteurs linguistiques proposés dans le guide d'annotation [10] (Mekki, Battistelli, Lecorvé et al. 2021)

10. Le guide d'annotation est disponible sur HAL : `https://hal.archives-ouvertes.fr/hal-03218217`.

puisqu'ils renvoient aux descripteurs décrits page 18 et page 24 de ce guide : *Élément doublé* et *Écriture électronique* (respectivement). Les tweets (78) à (80) constituent des exemples du motif 5 : (⟨ (*pos:nom-propre*) , (*pos:clitique-sujet*, *morpho:3*^{eme} *personne*), (*pos:verbe*, *morpho:présent*) ⟩).

(78) Westbrook il est comme sa lebron le poster il le block mais vreument url_path

(79) Matt Houston il me donne envie d'être amoureuse

(80) RT @X : Ptdr non tootatis elle abuse du bail là

Les tweets (81) à (83) constituent des exemples du motif 6 : ⟨ (*sous-mot:_c_*) ⟩.

(81) jamais tranquille c un truc de malade même contre brest

(82) Putain lakers rockets la ils sont en mode precision 3 pts max c est une dinguerie..

(83) @X @X @X Je l'utilise comme leme c fou comme sa a vite tourner 😭 😭

Le motif 7, quant à lui, peut renvoyer à de nouveaux usages numériques liés à l'ajout d'utilisateurs à un cercle d'amis en ligne ou bien à des groupes de conversation. Les tweets (84) à (86) constituent des exemples du motif 7 : ⟨ (*lemme:rajouter*) ⟩.

(84) Rt si tu veux que j'te rajoute url_path

(85) @X Mdrr desac pas stv jte rajoute ds un grp le sang

(86) heyyy coucou mes mutus ! ! donnez moi vos insta que je vous rajoute il me faut des kpop stan la 🥺

Les motifs 8 et 9 illustrent l'intérêt d'utiliser les *sous-mots* comme traits linguistiques pour décrire chaque *mot*. Ces MSE extraits montrent que certaines terminaisons morphologiques sont spécifiques du registre familier. Les tweets (87) à (89) constituent des exemples du motif 8 : ⟨ (*sous-mot:sh_*) ⟩.

(87) P T D R. Vous suez face à Brest wesh sois réaliste url_path

(88) RT @X : Les gens compare wejdene à Riri wsh réveillez vous non y'a aucune comparaison possible

(89) @X @X Ahah j'en étais sûr, non mais la le pire c'est la commu dbz qui full trash Naruto en oplus de la commu op, he crois actuellement, silhouette est l'op le plus détesté

Les tweets (90) à (92) constituent des exemples du motif 9 : ⟨ (*sous-mot:rrr_*) ⟩.

(90) dans 1 semaine chuis en Suède mdrrrr rien n'est prêt c'est la panik

(91) je suis mort Djoko disqualifié parce qu'il a mis une tête à un juge ptdrrrrrrrrr

(92) @X @X Xptdrrrr il est taré après les lakers vont gagner 100% je n'ai aucun doute

Enfin, le motif 10 rejoint les analyses linguistique présentées section 3.4.4 : les pictogrammes sont utilisés comme des ponctuations qui peuvent être répétées pour marquer l'intensité de la modalisation du locuteur sur son propos. Les tweets (93) à (95) constituent des exemples du motif 10 : ⟨ (*syntax:modifieur*), (*lemme:pictogramme*), (*lemme:pictogramme*) ⟩.

(93) RT @X : Franchement les 2 sont magnifiques mais Silhouette» ❤️ ❤️ ❤️

(94) Kaaris x Bosh ils ont tout plié en Deux Deux 🔥 🔥 🔥 url_path

(95) Axelle dit moi stp 🥺 🥺 🥺 🥺 🥺 🥺 🥺

Ces quelques exemples de MSE extraits ont montré la robustesse de l'extraction, sachant que les registres familier et courant sont deux registres proches dont les délimitations peuvent être difficiles à tracer. Nous avons pu mesurer la qualité des MSE en nous appuyant sur :

- le fait que certains MSE ont renvoyé directement à des descripteurs linguistiques issus, soit de la littérature sur le sujet, soit de l'exploration linguistique du corpus, dont certains ont été compris dans notre guide d'annotation ;
- le fait que les MSE extraits ont montré des motifs linguistiques caractéristiques du familier à l'échelle de terminaisons discriminantes.

Soutenu vs. familier A l'inverse de la paire de registre familier vs. courant, la caractérisation du soutenu par rapport au familier repose sur 2 registres fortement contrastés. Des traits linguistiques attendus, caractéristiques du registre soutenu, ont été retrouvés parmi les MSE extraits. Mentionnons, par exemple, le motif 1 qui renvoie au vouvoiement : ⟨ (*sous-mot:vous_*) ⟩. Les tweets (96) à (98) constituent des exemples du motif 1.

ID	Motif	GR
1	⟨ (*sous-mot:vous_*) ⟩	$+\infty$
2	⟨ (*sous-mot:_Pour_*) ⟩	$+\infty$
3	⟨ (*lemme:alors*, *syntax:modifieur*, *pos:adverbe*) ⟩	$+\infty$
4	⟨ (*pos:nom commun*, *sous-mot:_#*), (*pos:ponctuation*) ⟩	$+\infty$
5	⟨ (*lemme:de*), (*sous-mot:_#*) ⟩	$+\infty$
6	⟨ (*sous-mot:_entre*)(*sous-mot:_#*) ⟩	$+\infty$
7	⟨ (*syntax:sujet*)(*lemme:pictogramme*) ⟩	$+\infty$
8	⟨ (*sous-mot:lance_*) ⟩	$+\infty$
9	⟨ (*sous-mot:hui_*) ⟩	$+\infty$
10	⟨ (*sous-mot:ez_*) ⟩	1,76

TABLE 5.10 – Exemples de motifs séquentiels émergents caractéristiques du soutenu vs. familier.

(96) @X Si vous permettez pour une diffusion élargie, chez Plenel, ""Tiers État"" signifie le Peuple par opposition à la noblesse (soit disant supprimée) et au clergé (religieux soit disant exclu du champ sociétal, mais omniprésent ces 20 dernières années).

(97) #LeSaviezvous ? Le criquet pèlerin est le ravageur migrateur le plus destructeur au monde. En savoir plus 👉 url_path url_path

(98) @X @X @X René Bousquet était préfet de la République. Je crois que vous n'avez pas compris le sens du tweet. Dire "ministre de la République", c'est pas un totem d'immunité. On n'est pas à chat perché. @X

Aussi, les motifs 2 et 3 correspondent pour l'un à une préposition (pour), pour l'autre à un adverbe (alors) qui contribuent à structurer le texte. Les tweets (99) à (101) constituent des exemples du motif 2 qui permet d'identifier la position en début de phrase de la préposition grâce au sous-mot qui conserve la majuscule : ⟨ (*sous-mot:_Pour_*) ⟩.

(99) | #FranceRelance Les jeunes ont souvent été les premières victimes de la crise économique que nous vivons. Pour qu'ils puissent s'insérer rapidement et durablement sur le marché du travail, le @X s'engage : ⬇️ url_path

(100) Pour @X (@X) en France "le problème le plus important n'est pas la sécession (...) mais les racistes et les néo-fascistes" #IslamoCollabo #Indigénisme #Communautarisme #Immigration url_path

(101) @X @X @X @X Pour vous, le masque arrête-t-il le virus ?

L'adverbe *alors* contenu dans le motif 3, ⟨ (*lemme:alors*, *syntax:modifieur*, *pos:adverbe*) ⟩, est employé afin de structurer les tweets (102) à (104).

(102) @X Si moi, petit écrivaillon de banlieue, je savais dès 1986 à quoi m'en tenir sur les "amours" de #Matzneff, la défense à géométrie variable de Girard ne tient pas 1 seconde "je ne savais pas, j'étais aux USA", alors même que les écrits de GB éclairent son rôle de factotum de Berg

(103) #PopCultureFact : #Nintendo renomma sa console #Famicom #NES (Nintendo Entertainment System) pour la sortir aux USA en 1985. Elle relança un marché alors moribond et en qq mois, l'expression « jouer à la Nintendo » remplaça « jouer aux #jeuxvidéo » : in #HighScore sur @X url_path

(104) Si l'on réfléchit, alors on ne peut cautionner ce que fait Plenel : c'est haineux, ignoble et indéfendable 😖 url_path

Si les MSE 1, 2 et 3 sont des traits linguistiques classiques, les MSE 4, 5 et 6 intègrent des *technomorphèmes*, c'est-à-dire des éléments linguistiques propres aux écrits numériques. Ces motifs montrent que les hashtags sont intégrés à la norme grammaticale. Les tweets (105) à (107) constituent des exemples du motif 4 : ⟨ (*pos:nom commun*, *sous-mot:_#*), (*pos:ponctuation*) ⟩.

(105) Le module russe Zarya est le premier élément de l'#iss. Il est lancé le 20 novembre 1998 avant d'être rejoint 2 semaines plus tard par le module américain Unity. Aujourd'hui il sert principalement d'espace de stockage. CREDIT : NASA #espace #astronomie url_path

(106) Le pourcentage de DSI qui envisage de migrer vers le #cloud est passé de 54 % à 89 % après la #COVID19. Le #cloudcomputing est de plus en plus considéré comme un facilitateur de #remoteworking url_path

(107) @X Pour l'instant, #CharlieHebdo respecte scrupuleusement les limites définies par les tabous et la #censure . Quand @X voudra vraiment tester les limites de la tolérance en France, ce genre de caricature sera publié : url_path

Les tweets (108) à (110) constituent des exemples du motif 5 qui illustre également l'intégration des hashtags à la norme linguistique : ⟨ (*lemme:de*), (*sous-mot:_#*) ⟩.

(108) Tranquillement, l'équipe de #Trump ment et manipule des propos de #Biden. Twitter a marqué la publication comme "manipulée". C'est de la désinformation pure et simple. url_path

(109) Nous sommes scandalisés par les rejets volontaires du groupe #Lafarge dans la #Seine d'eaux usées contenant un mélange de particules de #ciment, de liquides de traitement et des microfibres de #plastique. Le groupe doit être sévèrement condamné ! #Paris url_path

(110) « Laval Agglomération est connectée et ouverte sur le monde... Nous ne pouvons pas rester insensibles en matière de #solidarité... » #Liban | url_path via @X en #Mayenne url_path

Enfin, le MSE 6 met aussi en lumière l'utilisation des hashtags comme des mots classiques : ⟨ (*sous-mot:_entre*)(*sous-mot:_#*) ⟩. Les tweets (111) à (113) constituent des exemples du motif 6.

(111) Notre note politique sur l'#AssembléedeBretagne a mis en évidence l'absurdité actuelle de la distribution des #compétences entre #collterr 🇫🇷. L'avenir est à l'intégration des politiques publiques ➡️ #ODD #agenda2030. Il nous faut des #Régions et #Villes au pouvoir plus intégré. url_path url_path

(112) #Darmanin se dit à « 100.000 lieues » de faire « le lien entre #immigration et #insécurité » et invoque ses origines familiales url_path url_path

(113) Psychodrame entre #Hidalgo et les écolos à Paris. Hidalgo est prisonnière de ses alliés, qui eux-mêmes sont otages de leurs extrêmes. Quand on voit qu'une manif de 20 hystériques obtient la tête d'1 adjoint qui serait l'ami de #Matzneff on se dit que le fascisme n'est pas loin

Un autre type de *technomorphème* est intégré à la norme grammaticale : les pictogrammes. Le MSE 7 montre que ces derniers peuvent être employés comme du lexique traditionnel syntaxiquement intégré : ⟨ (*syntax:sujet*)(*lemme:pictogramme*) ⟩.

(114) La destruction de l'#Amazonie empire. La 🇫🇷 est complice de ce désastre en raison de ses importations, @X l'a reconnu l'été dernier. Depuis ? Uniquement des paroles, 0 acte concret. url_path #JeudiPhoto #Marseille 🔥 url_path

(115) NON MERCI #RACHIDA, on n'a pas besoin de ta sensiblerie ! 🙄 Et encore moins dans une présidentielle à un moment où la 🇫🇷 est clairement à la croisée des chemins. #RachidaDati #Floyd #AdamaTraoré #AdamaViolleur #GangTraoré #LaRacailleTue #LR #LaDroiteLaPlus... url_path url_path

(116) #FranceRelance 🇫🇷 c'est aussi #EuropeRelance : l'🇪🇺 sera présente dans chacun des projets de ce plan, il ne faut pas avoir l'🇪🇺 honteuse, ni l'🇪🇺 invisible ⬇️ @X @X @X url_path'

Il est intéressant de noter que les pictogrammes indiquent une communication plutôt politique et/ou institutionnelle. Enfin, les derniers motifs que sont les MSE 8, 9 et 10 mettent en exergue l'apport des sous-mots pour la caractérisation des registres de langue avec 3 terminaisons caractéristiques du registre soutenu. Le premier, le MSE 8 (⟨ (*sous-mot:lance_*) ⟩) est illustré par les tweets (117) à **??**.

(117) 🔥 🔥 🔥 Comme l'école de la confiance et la bienveillance n'est pas l'école de la transparence il est indispensable de partager et relayer le travail des #stylosrouges qui recensent les cas #Covid_19 dans les établissements scolaires!!! 🔥 🔥 🔥

(118) #FranceRelance « Nous irons, c'est l'aspect non-financier, vers davantage de simplification afin de faciliter son appropriation rapide par toutes les entreprises » @X #PlanDeRelance url_path

(119) Écologie : Réorganiser les moyens de surveillance de l'état environnemental de la France ainsi que les principaux organismes de gestion des sols, des forêts et des eaux. #UPR #FRANÇOISASSELINEAU #ecologie url_path

Le MSE 9, ⟨ (*sous-mot:hui_*) ⟩, quant à lui renvoie à la terminaison du mot *aujourd'hui* et permet, à l'instar du MSE 3 (*alors*), de structurer (ici temporellement) le tweet.

(120) Donc le maire de Stains @X se bat pour le maintien de la fresque d'un violeur. #AdamaViolEur Monsieur le maire, maintenant que les faits sont aujourd'hui avérés, allez vous continuer votre combat pour défendre un violeur ? url_path

(121) @X @X Vous en êtes les responsables. Vous avez porté cette idéologie qui aujourd'hui commence à vous dévorer.

(122) Aujourd'hui c'est Soral... Demain c'est Dieudonné. "Quand les bandits sont au pouvoir, la place d'un honnête homme est en prison" (Michel Chartrand) url_path

Enfin, le MSE 10 rejoint le motif 1 car il indique la présence du vouvoiement : ⟨ (*sous-mot:ez_*) ⟩.

(123) De toute façon, ce n'est pas vous qui décidez du contenu des débats. Il y a pire que le point Godwin, c'est le point Shoah. On y revient TOUJOURS. "Goy, tu ne peux pas défendre les Blancs, parce que Hitler, bla bla bla". Adopter une posture anti-nazie par stratégie est idiot.

(124) @X Bonjour. Le placement automatique "un siège sur deux" n'est plus appliqué à bord des trains depuis début juin. Toutes les places peuvent désormais être occupées. Il est donc possible que vous soyez assis à côté d'une personne que vous ne connaissez pas. 1/2

(125) @X Vous avez la réponse à votre question : le VPCE n'est pas président de la section du contentieux.

Les différents MSE caractéristiques du soutenu par rapport au familier ont pu être considérés de qualité car :

- ils comprennent des traits saillants caractéristiques du soutenu traditionnellement associé dans la littérature scientifique (le vouvoiement, les éléments de discours qui le structurent de manière logique et temporel) ;
- ils correspondent à des descripteurs linguistiques mentionnés dans le guide d'annotation (intégration syntaxique des techomorphèmes) ;
- ils ont confirmé l'apport des sous-mots en mettant en exergue des terminaisons caractéristiques du soutenu.

Les MSE ont confirmé le phénomène d'intégration à la norme grammaticale d'éléments propres aux discours numériques : les hashtags ou les pictogrammes ne sont plus réservés au registre familier, mais au contraire sont utilisés pour des communications institutionnelles.

5.4 Conclusion

Ce chapitre a présenté la mise en œuvre du processus d'extraction de MSE pour caractériser un registre de langue par rapport à un autre. La première partie de cette section a présenté une de nos contributions qui réside dans la proposition d'une méthodologie qui a permis d'évaluer la robustesse des techniques de fouilles de motifs en utilisant des textes artificiels. En effet, une difficulté de ces techniques réside dans l'évaluation des motifs retournés qui, de fait, sont nouveaux et ne peuvent être comparés à une base de référence. Les textes artificiels permettent d'intégrer des traits linguistiques connus, afin de

pouvoir vérifier si ces derniers sont retrouvés lors de l'extraction de MSE. Les expériences présentées ont validé la pertinence de cette technique de fouille de motifs.

La seconde partie de cette section a présenté l'extraction de MSE à partir du corpus considéré dans le cadre de notre travail comme représentatif des registres de langue, le corpus TREMoLo-Tweets. Lors de la transformation du jeu de données textuelles en base de données séquentielles, nous avons eu recours à un trait linguistique qui à notre connaissance n'avait jamais été utilisé dans des travaux de fouille de données textuelles : les *sous-mots*. Nous rappelons que les sous-mots sont des unités syllabiques composant un mot, nous les avons obtenus grâce à l'outil de segmentation automatique de CamemBERT. L'exploration des MSE extraits a permis d'illustrer l'apport de ce trait pour une tâche de caractérisation de registres, avec notamment l'extraction de terminaisons de mots discriminantes par exemple. En outre, elle a de nouveau confirmé la capacité des MSE à distinguer un registre cible d'un registre source en retrouvant des descripteurs linguistiques utilisés par les annotateurs experts (par exemple, l'utilisation de pictogramme aidant à la structuration du texte) pour distinguer les registres.

Cependant, comme précisé en introduction de cette section, cette exploration manuelle des résultats est partielle et chronophage. Le chapitre suivant développe les travaux conduits pour tirer un sous-ensemble de MSE plus exploitable.

Chapitre 6

Constitution d'un sous-ensemble interprétable de motifs séquentiels émergents

Sommaire

L'avantage de notre méthodologie repose sur le fait de ne pas exclure de MSE potentiellement pertinents durant la fouille et d'accepter que les résultats soient nombreux et peu exploitables. Ce sont les traitements suivant leur découverte qui les analysent pour en tirer un sous-ensemble plus interprétable, c'est-à-dire moins volumineux et moins redondant. La table 6.1 donne 20 exemples de MSE caractéristiques du R_c familier par rapport au R_s courant. Ces exemples illustrent la redondance des MSE perceptible à travers la proportion d'items différents par rapport au nombre total d'items des 20 MSE : avec 12 items différents pour un total de 59 items, chaque item est répété au moins quatre fois. Comment ne pas écarter de MSE intéressants en les réduisant en un sous-ensemble moins redondant ? Pour cela, nous avons mis en place deux tâches se succédant. Tout d'abord, nous avons regroupé les MSE selon leurs similarités pour réduire la redondance. Puis, nous avons constitué un sous-ensemble de MSE composé de *MSE représentants*, chacun issu de différents groupes, pour réduire leur nombre. La figure 6.1 illustre l'approche globale pour constituer un sous-ensemble de MSE dans laquelle (1) nous divisons l'ensemble complet de MSE en n groupes de MSE similaires (2) desquels nous en tirons n MSE représentants. La première étape est plus simple que la seconde car les groupes de MSE sont suffisamment différents entre eux pour les répartir dans des groupes selon leur similarité,

ID	MSE
1	⟨{sous-mot :_tu_} {pos :V}⟩
2	⟨{sous-mot :_tu_} {pos :V, morpho :singulier}⟩
3	⟨{sous-mot :_tu_} {pos :V, morpho :singulier, morpho :présent}⟩
4	⟨{sous-mot :_tu_} {morpho :2e p}⟩
5	⟨{sous-mot :_tu_} {morpho :singulier}⟩
6	⟨{sous-mot :_tu_} {morpho :singulier} {morpho :singulier}⟩
7	⟨{sous-mot :_tu_} {morpho :singulier} {syntax :mod}⟩
8	⟨{sous-mot :_tu_} {morpho :singulier, morpho :2e p}⟩
9	⟨{sous-mot :_c} {pos :V}⟩
10	⟨{sous-mot :_C} {pos :V}⟩
11	⟨{sous-mot :_c} {pos :V} {pos :det}⟩
12	⟨{sous-mot :_c} {pos :V} {pos :det} {morpho :singulier}⟩
13	⟨{sous-mot :_c} {pos :V} {pos :det} {morpho :masculin}⟩
14	⟨{sous-mot :_c} {pos :V} {pos :det, morpho :singulier}⟩
15	⟨{lemme :faire} {pos :ADV}⟩
16	⟨{lemme :faire} {pos :ADV, syntax :mod}⟩
17	⟨{lemme :faire} {morpho :singulier}⟩
18	⟨{lemme :faire} {morpho :singulier} {pos :NC}⟩
19	⟨{lemme :faire} {morpho :singulier} {pos :NC, morpho :singulier}⟩
20	⟨{lemme :faire} {morpho :singulier} {pos :NC, syntax :objet}⟩

TABLE 6.1 – Exemples de MSE caractéristiques le R_c familier par rapport au R_s courant.

en revanche les MSE au sein de chaque groupe sont tellement similaires entre eux qu'il est difficile d'établir des critères objectifs pour en sélectionner un représentant. Par exemple, en observant les MSE de la table 6.1, nous distinguons facilement trois groupes de MSE : le premier groupe incluant les MSE du 1er MSE de la table au 8ème MSE ; le deuxième groupe incluant ceux du 9ème au 14ème MSE ; et le troisième incluant ceux du 15ème au 20ème MSE. En revanche, sur quels éléments baser la sélection de trois MSE représentants chacun issu d'un des trois groupes ? Tous les MSE des trois groupes sont très similaires entre eux. Par exemple, en quoi le MSE 11 serait un meilleur représentant que le MSE 13 pour le deuxième groupe ? La tâche est impossible sans description en amont des critères qu'un MSE doit remplir pour représenter son groupe. Ne voulant toujours pas poser d'*a priori* linguistique sur ce que devrait être un MSE représentant, nous avons proposé plusieurs critères basés sur les éléments liés à la formalisation des MSE plutôt que sur des valeurs linguistiques de ces éléments. Par exemple, nous avons regardé la fréquence des itemsets au sein d'un groupe de MSE plutôt que de pondérer un niveau d'analyse de la langue particulier en regardant la valeur des items. La première section de ce chapitre est

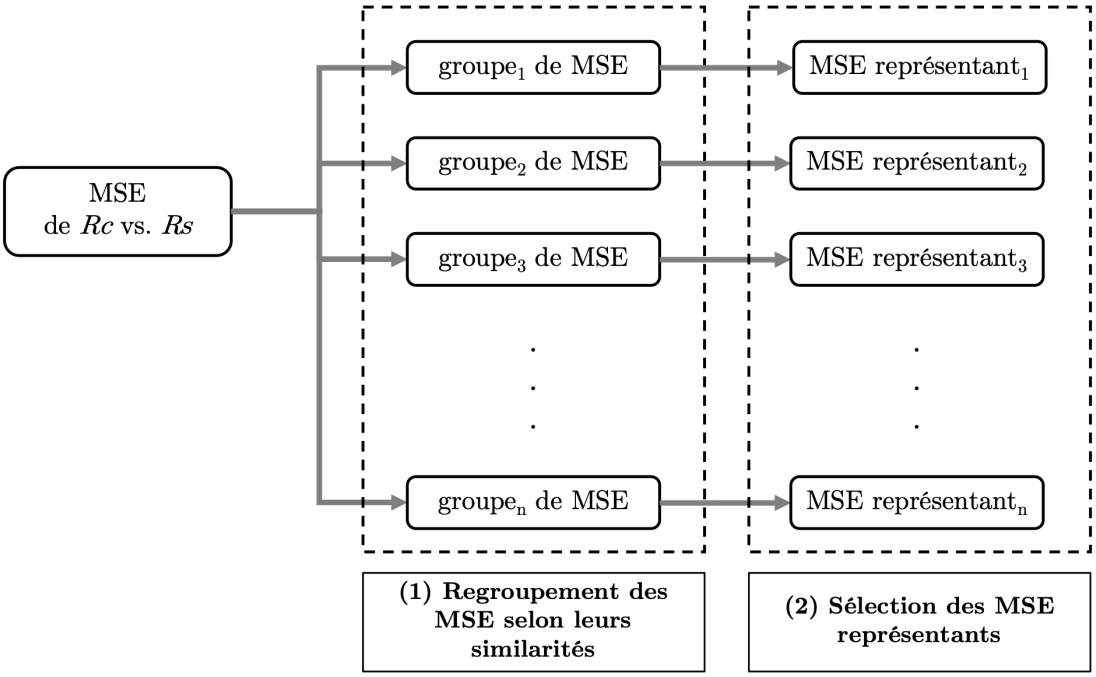

FIGURE 6.1 – Approche globale pour réduire les MSE.

dédiée aux travaux réduisant la redondance des MSE en les regroupant selon leur similarité. Dans la section suivante, nous exposons les travaux réduisant la quantité de MSE à un sous-ensemble de MSE représentant. La troisième et dernière section de ce chapitre expose les travaux évaluant ce sous-ensemble.

6.1 Réduction de la redondance des motifs séquentiels émergents

Nous avons fait face à deux enjeux lors de cette tâche : limiter la complexité algorithmique causée par la comparaison des MSE entre eux et déterminer les critères à partir desquels la similarité entre deux MSE est calculée. Pour répondre au premier, nous avons proposé l'algorithme RGMSE [1]. Pour répondre au second, nous avons exploré la littérature scientifique sur le sujet afin d'en sélectionner une mesure adaptée aux MSE et correspondant à nos critères : S^2MP introduite par (SANEIFAR et al. 2008). Nous détaillons dans cette section les expériences utilisant RGMSE et S^2MP pour regrouper les MSE selon

1. pour **R**e**G**roupement de **M**otifs **S**équentiels **É**mergents

leur similarité, après avoir dressé un panorama des mesures de similarités existantes pour les MSE.

6.1.1 Contexte et motivations

Notre but est de regrouper les MSE selon leurs ressemblances, autrement dit d'obtenir des groupes de MSE dont les individus sont similaires entre eux et différents des individus des autres groupes. Pour assurer la qualité d'un regroupement de MSE, nous avons choisi une mesure de similarité répondant à nos contraintes pour comparer les MSE parmi les mesures proposées dans la littérature détaillées dans cette section.

6.1.1.1 Pourquoi choisir une mesure de similarité spécifique aux motifs séquentiels émergents ?

Avant de lister les cinq mesures de similarité proposées, nous expliquons ici quelles sont les contraintes liées à nos travaux pour choisir une d'entre elles. Tout d'abord, nous avons pour objectif de caractériser un phénomène linguistique. Nos critères pour comparer les MSE doivent donc veiller à avoir du sens linguistiquement. Ensuite, les techniques de regroupement ont toutes des verrous algorithmiques liés au coût de calcul engendré par le nombre de comparaisons entre les individus à regrouper. Ce coût de calcul est augmenté avec la comparaison des MSE si nous ne choisissons pas une mesure de similarité adaptée. Ces deux points sont discutés dans les paragraphes suivants.

Nous rappelons que lorsque nous comparons deux séquences, nous comparons deux segments textuels où chaque mot est représenté par un itemset ; lorsque nous comparons deux itemsets entre eux, nous comparons deux mots via deux ensembles de traits linguistiques les décrivant. Pour que la mesure de similarité ait du sens linguistiquement, il faut qu'elle tienne compte :

- de l'ordre des itemsets des MSE, pour rendre compte l'importance de l'ordre des mots entre eux ;
- du nombre d'items communs entre deux itemsets à des positions similaires entre deux MSE, pour rendre compte de l'importance des traits linguistiques communs aux deux mots comparés.

Les MSE étant des séquences d'ensembles, une approche générique considère les ensembles comme des listes et multiplie les comparaisons en faisant varier l'ordre des items pour chaque MSE. Cette multiplication des comparaisons augmente considérablement

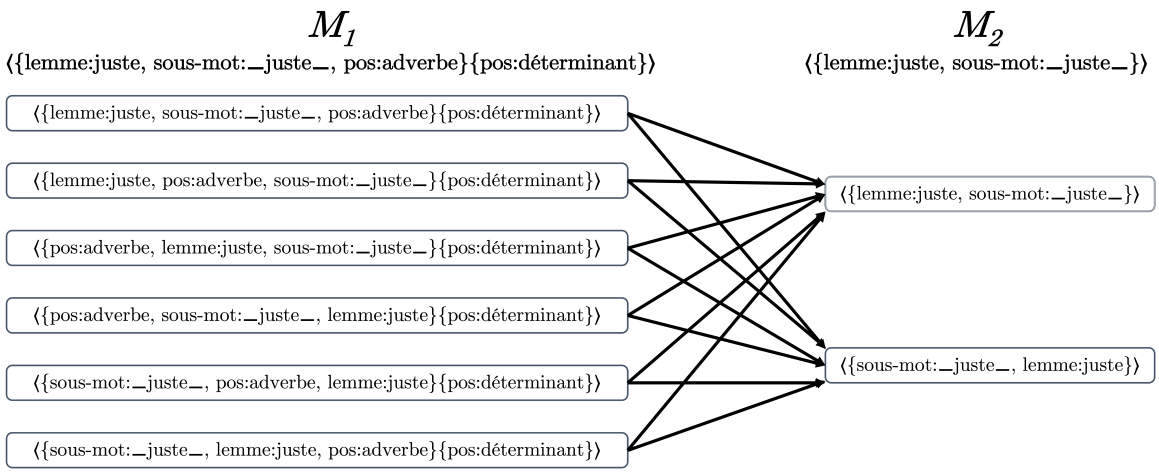

FIGURE 6.2 – Exemples des comparaisons à faire pour estimer la similitude entre deux motifs.

le coût algorithmique nécessaire au regroupement de données. Par exemple, considérons deux motifs M_1 et M_2 dans la figure 6.2 :

1. $M_1 = \langle$ (*sous-mot:_juste_*, *lemme:juste*, *pos:adverbe*), (*pos:déterminant*) $\rangle$,
2. $M_2 = \langle$ (*sous-mot:_juste_*, *lemme:juste*) $\rangle$.

La figure 6.2, donne des exemples des variations de configuration de M_1 et M_2 pour les comparer entre eux. À gauche de la figure, les six variations de M_1 sont illustrées faisant face aux deux variations de M_2 à droite de la figure. Au total, il y a donc 6×2 comparaisons à faire. Dès lors, la mesure de similarité doit être adaptée à la comparaison d'ensembles pour éviter la génération de toutes les combinaisons possibles provoquant l'augmentation du coût algorithmique.

Ainsi, le choix d'une mesure de similarité spécifique aux MSE est motivé par notre objectif linguistique de caractérisation des registres ainsi que le besoin de limiter le coût algorithmique en évitant la multiplication des calculs liés aux itemsets. La section suivante donne une vision d'ensemble des différentes mesures pour les motifs proposées dans la littérature.

6.1.1.2 Quelle mesure de similarité pour nos travaux ?

Cette section présente cinq mesures pour estimer la similarité entre deux MSE. Elle expose leurs avantages et limites, afin de motiver notre choix d'utiliser la mesure S^2MP

proposée par (Saneifar et al. 2008) dans nos travaux.

La distance d'édition La première mesure est la distance d'édition proposée dans les travaux de (Capelle, Masson et Boulicaut 2002). Ils utilisent la distance d'édition (Levenshtein et al. 1966) pour extraire des motifs séquentiels sous des contraintes de similarité. Les auteurs définissent un motif séquentiel comme une liste ordonnée de symboles appartenant à un ensemble fini d'alphabets. Dans ce travail, un motif séquentiel est considéré comme une séquence d'événements, où chaque itemset est un événement, dont les items sont les attributs. Une limite de ce modèle, est la possibilité de manquer des séquences similaires.

La mesure LCS La deuxième mesure, la mesure LCS, est introduite par (Sequeira et M. Zaki 2002) afin de comparer des motifs séquentiels. *LCS* donne la taille de la plus longue sous-séquence commune de deux séquences. Ce travail présente 3 limites : *LCS* ne prend pas la position des itemsets en compte ; elle ne considère pas la taille de la sous-séquence qui n'est pas commune ; la valeur de *LCS* n'est pas affectée par le nombre d'items différents dans les itemsets de la sous-séquence commune.

Similarité sur les items communs La troisième mesure se base sur le nombre les items communs entre 2 séquences, elle est proposée dans les travaux de (Guralnik et Karypis 2001). Après avoir aligné les séquences entre elles (pour maximiser leurs similarités), la mesure compte le nombre d'items en commun entre 2 itemsets, puis met à l'échelle le comptage afin que la similitude soit toujours un nombre compris entre 0 et 1. Étant donné deux séquences, S'_1 et S'_2, la similarité sim_1 sur leurs items communs est définie par l'équation 6.1.

$$\mathrm{sim}_1\left(S'_1(i), S'_2(i)\right) = \frac{|S'_1(i) \bigcap S'_2(i)|}{\frac{\left|S'_1(i)\right|+\left|S'_2(i)\right|}{2}} \tag{6.1}$$

La principale limite de cette mesure est de favoriser le contenu en commun des itemsets, au détriment de l'ordre des itemsets entre eux. L'alignement des motifs entre eux permet de découvrir des motifs séquentiels qui ont le même ordre d'itemsets, mais il ne permet pas de pondérer négativement la mesure de similarité lorsque les itemsets sont ordonnés différemment.

Similarité dans un espace vectoriel La quatrième mesure représente les motifs séquentiels dans un espace vectoriel, elle est également introduite par (Guralnik et Karypis 2001). Chaque itemset est représenté par un vecteur I dans l'espace des items : $I = (i_1, i_2, ..., i_n)$, où i_j est un booléen indiquant la présence du j^{eme} item dans l'itemset (1 pour sa présence, 0 pour son absence). Étant donné cette représentation, la distance cosinus est utilisée pour calculer la similarité entre deux séquences. La principale limite de cette mesure de similarité est de donner plus d'importance à l'ordre des itemsets, qu'aux contenus communs aux 2 itemsets. La représentation des motifs transforme les ensembles des itemsets en listes ordonnées. Cette transformation peut écarter des motifs pourtant similaires, par exemple l'itemset $I_1 = (a, b, c)$ sera jugé différent de l'itemset $I_2 = (c, b, a)$, alors qu'ils sont similaires.

La mesure S^2MP Nous avons retenu pour nos travaux la mesure S^2MP (pour **S**imilarity **M**easure for **S**equential **P**atterns) introduite par (Saneifar et al. 2008), car elle permet :

- de comparer le contenu des itemsets sans les transformer en liste ;
- de rendre compte de l'ordre des itemsets entre eux en considérant, non seulement l'ordre, mais l'écart entre deux itemsets similaires.

S^2MP est basée sur deux scores pour calculer la similarité de deux MS : le score de correspondance et le score d'ordre. Le score de correspondance mesure la ressemblance de deux séquences sur la base des items partagés ; tandis que le score d'ordre mesure la ressemblance de deux séquences en fonction de l'ordre et des positions des itemsets dans les séquences. Ils sont tous les deux présentés ci-dessous.

Score de correspondance L'objectif du score de correspondance est de mesurer la similarité entre une séquences S_1 et une séquence S_2, en comparant le contenu de leurs itemsets. Ce score de correspondance, appelé $moyenneCoresp$, est calculé à partir de la liste $corspList$. Chaque élément de cette liste correspond à la position de l'itemset I' avec lequel l'itemset I est le plus proche : la première place de $corspList$ correspond au premier itemset I_1 de S_1 et la valeur de l'entier stocké à cette place correspond à la position de l'itemset de S_2 avec lequel I_1 est le plus proche ; la deuxième correspond à l'itemset I_2 de S_1 et la valeur de l'entier stocké en deuxième correspond à la position de l'itemset de S_2 avec lequel I_2 est le plus proche ; et ainsi de suite. À partir de $corspList$, nous calculons $moyenneCoresp$ en faisant la moyenne des scores $scoreCorsp$ entre chaque paire d'itemsets I et I' qu'elle contient. L'itemset I' est l'itemset avec lequel I est le plus

proche en terme de contenu. Pour découvrir quel est l'itemset I' de S_2 avec lequel I est le plus similaire, nous calculons *scoreCorsp* entre I et tous les itemsets de S_2 : I' est l'itemset avec lequel I a obtenu le score *scoreCorsp* le plus élevé.

$$scoreCorsp(I, I') = \frac{|I \cap I'|}{(|I| + |I'|)/2} \tag{6.2}$$

scoreCorsp introduit par l'équation 6.2 regarde la proportion d'items communs à I et I' par rapport au nombre moyen d'items contenus dans I et I'. Par exemple, considérant $S_1 = \langle\{c, d\}, \{c\}\rangle$ et $S_2 = \langle\{a\}, \{c, d\}, \{a\}, \{c\}\rangle$, pour trouver l'itemset le plus proche de $I_1 = \{c, d\}$ nous calculons ses *scoreCorsp* avec tous les itemsets de S_2, soit :

— $scoreCorsp(I_1, I'_1) = \frac{0}{(2+1)/2} = 0$,

— $scoreCorsp(I_1, I'_2) = \frac{2}{(2+2)/2} = 1$,

— $scoreCorsp(I_1, I'_3) = \frac{0}{(2+1)/2} = 0$,

— $scoreCorsp(I_1, I'_4) = \frac{1}{(2+1)/2} = 0,7$.

Dès lors, l'itemset I_1 est mis en correspondance avec l'itemset I'_2 car c'est avec lui qu'il obtient le *scoreCorsp* le plus élevé.

Score d'ordre Le score d'ordre, noté *scoreOrdre*, compare la similarité de deux séquences S_1 et S_2 en regardant les positions de leurs itemsets via la liste *corspList*. Pour cela, *orderScore* est obtenu à partir de deux scores : le premier, noté *ordreTotal*, mesure le nombre d'itemsets de S_1 dont les positions ne croisent pas celles des itemsets de S_2 ; le second, noté *ordrePosition*, mesure la distance entre deux itemsets similaires. La figure 6.3 donne deux exemples de listes de correspondances $corspList_1$ et $corspList_2$. Chacune d'elle est présentée avec les deux séquences S_1 et S_2 qu'elle compare.

Considérant $corspList_1$ et $corspList_2$, *scoreTotal* regarde si les positions des itemsets de S_1 et S_2 mis en correspondance se croisent ou non. En l'occurrence, les itemsets de $corspList_1$ ne se croisent pas, mais ceux de $corspList_2$ se croisent. Une suite d'itemsets contenant des itemsets dont les positions ne se croisent pas est appelée une sous-séquence croissance notée *croissI*, en l'occurrence $croissI_1 = [1, 3]$. Nous donnons la définition de *scoreTotal* avec l'équation 6.3, *scoreTotal* fait le rapport du nombre total d'itemsets ne se croisant pas noté *croissItemsets* sur la moyenne du nombre de tous les itemsets dans les

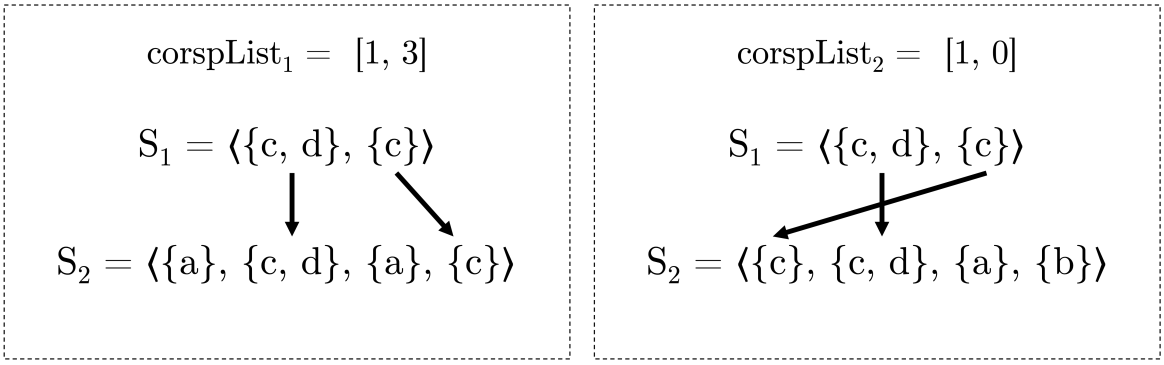

FIGURE 6.3 – Exemples de deux listes de correspondances, $corspList_1$ et $corspList_2$ associées aux exemples de séquences S_1 et S_2 qu'elles comparent.

deux séquences noté $moyenNbItemsets$. Dès lors, le $scoreTotal(corspList_1) = \frac{2}{3} = 0.7$

$$scoreTotal = \frac{croissItemsets}{moyenNbItemsets} \tag{6.3}$$

Le score *ordrePosition* (équation 6.4), quant à lui, mesure l'écart entre les positions des itemsets mis en correspondance à partir de *corspList*.

$$ordrePosition = \sum_{i=1}^{|croissI|} \frac{|valCL(i) - valCL(i-1)| - |\ posCL(valCL(i)) - posCL(valCL(i-1))\ |}{moyenNbItemsets} \tag{6.4}$$

Où :

- $valCL(i)$ = la valeur de la i^{eme} position mémorisée dans la *corspList*
- $posCL(x)$ = la position de l'itemset x dans la *corspList*

Considérant $croissI_1$ nous obtenons $ordrePosition(croissI_1) = \frac{(3-1)-(2-1)}{3} = 0.3$. Enfin, le score d'ordre *ordreScore* est obtenu en calculant le produit des scores *ordreTotal* et *ordrePosition*, de toutes les *croissI* de la *corspList* : seul le score le plus élevé est conservé et représente l'*ordreScore* (équation 6.5).

$$\begin{aligned} ordreScore &= \max\{ordreTotal(sub) \times (1 - ordrePosition(sub))\} \\ sub &\in \{\ \textit{la sous-séquence croissI contenant le plus grand nombre d'itemsets}\ \} \end{aligned} \tag{6.5}$$

Par exemple, considérant $croissI_1$ comme étant *sub*, alors $orderScore = 0.7 \times (1 - 0.3) = 0.49$.

Score de similarité Nous calculons le degré de similarité final, noté $SimDegre$ (équation 6.6), comme une agrégation entre l'*ordreScore* et le poids moyen de la correspondance des itemsets représenté par l'*scoreCorsp*. L'agrégation peut être une moyenne pondérée en définissant le coefficient de chaque score : l'ordre peut être considéré comme plus (ou moins, ou autant) important que le contenu en fonction du contexte de l'application.

$$SimDegre = \frac{(ordreScore \times Co_1) + (scoreCorsp \times Co_2)}{Co_1 + Co_2} \tag{6.6}$$

La valeur de $SimDegre$ varie de 0 à 1 : plus sa valeur est proche de 0 plus les deux MS comparés sont différents, à l'inverse plus sa valeur est proche de 1 plus les MS sont similaires. Nous donnons six exemples de comparaisons dont les valeurs de $SimDegre$ sont de plus en plus fortes :

1. comparaison 1 : $SimDegre = 0$
   (a) ⟨ (*sous-mot:_vous_*) ⟩
   (b) ⟨ (*lemme:ne, sous-mot:_n*), (*pos:verbe, sous-mot:ez_*) ⟩
2. comparaison 2 : $SimDegre = 0$
   (a) ⟨ (*sous-mot:_vous_*) (*sous-mot:ez_*) ⟩
   (b) ⟨ (*sous-mot:ez_*) ⟩
3. comparaison 3 : $SimDegre = 0.5$
   (a) ⟨ (*sous-mot:ez_*) (*sous-mot:_pas_*) ⟩
   (b) ⟨ (*lemme:ne*) (*sous-mot:ez_*) ⟩
4. comparaison 4 : $SimDegre = 0.5$
   (a) ⟨ (*lemme:ne, sous-mot:_n*) ⟩
   (b) ⟨ (*lemme:ne, sous-mot:_n*), (*pos:verbe, sous-mot:ez_*) ⟩
5. comparaison 5 : $SimDegre = 0.8$
   (a) ⟨ (*lemme:ne, sous-mot:_n*), (*pos:verbe, sous-mot:ez_*) ⟩
   (b) ⟨ (*lemme:ne, sous-mot:_n*), (*sous-mot:ez_*) ⟩
6. comparaison 6 : $SimDegre = 1$
   (a) ⟨ (*lemme:ne, sous-mot:_n*), (*pos:verbe, sous-mot:ez_*) ⟩
   (b) ⟨ (*lemme:ne, sous-mot:_n*), (*pos:verbe, sous-mot:ez_*) ⟩

6.1.2 Partitionnement de l'ensemble des motifs séquentiels émergents

Cette section présente le protocole expérimental mis en place pour le partitionnement automatique des MSE (section 6.1.2.1) les regroupant selon leur similarité mesurée avec S^2MP, ainsi que les résultats de ce partitionnement (section 6.1.2.2).

6.1.2.1 Protocole expérimental

Dans cette section, nous présentons notre algorithme de regroupement de motifs avant d'exposer le protocole expérimental l'utilisant. Nous évaluons ensuite les groupes obtenus.

Notre algorithme de regroupement des motifs : RGMSE Pour regrouper les MSE selon leur similarité, nous avons proposé l'algorithme RGMSE dont l'approche est similaire à celle de la classification ascendante hiérarchique (CAH). Cependant, RGMSE diffère sur deux points pour accélérer le traitement d'une grande quantité de données. Le premier est un tirage sans remise des individus de l'ensemble de motifs, le second est un procédé de division des groupes volumineux en plusieurs sous-groupes pour accélérer le calcul des médoïdes [2]. L'utilisateur doit préciser à RGMSE trois paramètres : $minSim$ un seuil de similarité minimale entre deux individus ; $maxLen$ un seuil de taille maximale d'un cluster ; $nbrIter$ fixant le nombre d'itérations qui répète les étapes 2 et 3. RGMSE prend en entrée une liste de motifs M et retourne un ensemble de groupes de motifs G, ce regroupement se déroule en quatre étapes principales :

(1) **Le regroupement des objets par similitude selon *minSim*** : RGMSE s'initialise en considérant chaque motif m de la liste M comme un groupe g à lui tout seul. En partant du premier motif m_1 correspondant donc au premier groupe de motif, RGMSE tire les autres motifs de M dans leur ordre d'apparition. Si la similarité entre m_1 et le motif tiré est supérieure ou égale à $minSim$, alors le motif est ajouté au groupe g_1 et retiré de M. Lorsque RGMSE a terminé de parcourir M, il passe au motif suivant considéré comme le groupe suivant, et ainsi de suite.

(2) **La recherche des médoïdes pour chaque cluster** : Pour chaque groupe de G, RGMSE cherche son médoïde selon deux approches :

- Si un groupe g a un nombre de motifs $|g|$ supérieur ou égal à $maxLen$; alors le groupe est divisé en $\frac{|g|}{maxLen}$ sous-groupes. Pour chacun de ces sous-groupes,

2. Le médoïde est le motif le proche de la position moyenne de tous les motifs du groupe.

RGMSE calcule leur médoïde. L'ensemble des médoïdes obtenus à partir des sous-groupes est considéré comme l'ensemble des motifs de g à partir duquel est calculé le médoïde final.

- Sinon, RGMSE cherche directement le médoïde final à partir des objets de g.

(3) **La redistribution des objets parmi les clusters selon la similarité maximale entre eux et tous les médoïdes** : RGMSE parcourt M à partir de m_1 pour lequel il calcule sa similarité avec tous les médoïdes des groupes de G. m_1 rejoint le groupe du médoïde avec lequel sa similarité est maximale. RGMSE passe ensuite à m_2, puis m_3, etc.

(4) **La répétition de l'étape 2 et 3 *nbrIter* fois** : soit l'utilisateur a fixé la valeur de *nbrIter* ; soit RGMSE répète les étapes 2 et 3 jusqu'à ce qu'il converge, c'est à dire jusqu'à ce que la répartition des objets dans les clusters ne bouge plus.

RGMSE réduit son temps de recherche grâce à l'étape (1). Les étapes (2) et (3) redistribuent les motifs dans les groupes avec lesquels ils sont les plus similaires. Cette redistribution contrebalance la répartition de l'étape (1) dépendante de l'ordre dans lequel les motifs sont tirés. Nous présentons dans le paragraphe suivant les paramètres expérimentaux utilisés lors de nos expériences.

Paramètres expérimentaux Dans le cadre des expériences présentées dans ce chapitre, nous avons fixé :

1. *minSim* à 0.50, afin d'avoir un seuil de similarité minimale ni trop stricte ni trop souple ;
2. *maxLen* à 500, pour obtenir des sous-groupes suffisamment grands afin de ne pas avoir un nombre de médoïdes à calculer trop important, et suffisamment petits pour réduire le temps de calcul des médoïdes ;
3. *nbrIter* à 2, afin d'assurer la qualité de la répartition des individus, sans alourdir le temps de calcul provoqué par un nombre d'itération supérieur.

Les paramètres de ce protocole expérimental ont cherché à trouver un compromis entre le besoin de réduire la complexité algorithmique et la qualité de la partition des MSE.

Mesures pour l'évaluation des partitions La partition obtenue peut être validée selon différents procédés. Les travaux présentés dans (Theodoridis et Koutroumbas 2006) proposent de classer ces procédés de validation en trois types :

1. **une validation externe** qui compare la partition obtenue avec une partition de référence externe, c'est-à-dire une partition connue de l'utilisateur que doit recrée l'algorithme ;
2. **une validation interne** qui utilise des informations internes des groupes, pour évaluer la qualité de la partition, sans comparaison avec une partition de référence externe ;
3. **une validation comparative** qui compare 2 partitions, en faisant varier différentes valeurs de paramètres, pour le même algorithme (par exemple, la valeur de k avec l'algorithme du *k-means* pour faire varier le nombre de groupes).

Dans le cadre de notre travail, les validations des types externe et comparative ont été écartées, car aucune partition de référence externe n'était disponible ; et nous souhaitions éviter la multiplication d'expériences nécessaires à la validation. C'est pourquoi, nous avons exploré les mesures qui permettaient une **validation interne** de la partition, c'est à dire qui se basaient sur les informations internes aux groupes de MSE. Ces mesures ont vérifié que les groupes obtenus :

- maximisent la similarité entre les points d'un même groupe, appelée la *cohésion*, c'est à dire qu'il y a bien une distance minimale *intra-groupe* basée sur la distance moyenne entre les individus de C_i ;
- minimisent la similitude entre les points de C_i et ceux d'un second groupe C_j, appelée la *séparation*, c'est à dire qu'il y a bien une distance maximale *inter-groupes* basée sur la distance moyenne entre les individus de C_i et C_j).

La majorité des mesures d'évaluation proposées dans la littérature fait le rapport entre la cohésion CH et la séparation SP. Ce qui diffèrent se sont les manières de calculer CH et SP, ainsi que le domaine de variation de leurs résultats. Nous en présentons quatre dans cette section à partir desquelles nous en avons choisies deux pour nos travaux : le coefficient de silhouette et l'indice de Daves-Bouldin.

Coefficient de silhouette La première des mesures est le coefficient de silhouette introduit par (Rousseeuw 1987). Pour un individu i d'un groupe C_i, son coefficient de silhouette est défini par l'équation 6.7 :

$$Sil(i) = \frac{SP(i) - CH(i)}{\max(CH(i), SP(i))} \tag{6.7}$$

Où :

- $CH(i)$ calcule la cohésion de C_i en faisant la moyenne des distances entre l'individu i et tous les autres individus de C_i ;
- $SP(i)$ calcule la séparation de C_i avec son plus proche voisin C_j en faisant la moyenne des distances entre l'individu i et tous les individus de C_j.

Le coefficient de silhouette d'un groupe, proprement dit, est la moyenne des coefficients de silhouette pour tous ses individus. Son domaine de variation va de -1 (pour la pire partition) à 1 (pour la meilleure partition). La limite de ce coefficient réside principalement dans la multiplication des calculs à faire. En revanche, son domaine de variation permet d'avoir des résultats facilement interprétables puisque bornés entre -1 et 1.

Indice de Dunn La deuxième mesure est l'indice de Dunn présenté dans (Dunn 1973). L'indice de Dunn représente CH en regardant la distance maximale séparant deux individus d'un même groupe, et SP avec la distance minimale séparant deux individus répartis dans deux groupes différents. L'indice de Dunn varie entre 0 (pour son pire score) et $+\infty$ (pour son meilleur score). Le fait qu'il repose uniquement sur deux variables (une distance maximale, et une distance minimale) rend l'indice de Dunn peu robuste. En outre, son domaine de variation le rend difficilement interprétable.

Indice de Calinski-Harabasz La troisième mesure est l'indice de Calinski-Harabasz introduit par (Caliński et Harabasz 1974). CH est obtenu en faisant la moyenne des distances entre tous les individus du groupe et son centre, SP représente la distance entre le centre du groupe et un centre global (point moyen de toute la partition). L'indice de Calinski-Harabasz varie entre 0 (pour son pire score) et $+\infty$ (pour son meilleur score). Tout comme l'indice de Dunn, ce domaine de variation rend son interprétation difficile. Une seconde limite réside dans le coût algorithmique engendré par le calcul du centre global.

Indice de Davies-Bouldin Enfin, la quatrième mesure est l'indice de Daves-Bouldin présenté dans (Davies et Bouldin 1979). Pour calculer CH, il faut regarder la moyenne du rapport maximal entre la distance d'un individu au centre de son groupe. SP représente la distance entre 2 centres de groupes différents (celui de son groupe, celui du groupe voisin le plus proche). L'indice de Davies-Bouldin, noté S_{DB}, est défini par l'équation 6.8

et varie entre 0 (pour son meilleur score) et $+\infty$ (pour son pire score) :

$$S_{DB} = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left\{\frac{CH(C_i)+CH(C_j)}{SP(C_i,C_j)}\right\} \tag{6.8}$$

Où :

- $SP(C_i, C_j)$ représente la distance entre le centre de C_i et le centre du groupe le plus proche C_j ;
- $CH(C_i)$ représente la distance moyenne entre tous les individus du groupe C_i et son centre.

Bien que proche de l'indice de Dunn, l'indice de Davies-Bouldin est plus robuste car il considère la distance moyenne entre tous les individus d'un groupe et son centre.

Comme précisé en début de section, nous avons seulement retenu le coefficient de silhouette et l'indice de Davies-Bouldin pour valider les partitions de données. L'indice de Dunn a été écarté car insuffisamment robuste et peu interprétable. L'indice de Calinski-Harabasz, quant à lui, a été exclu car peu interprétable et coûteux algorithmiquement. Pour toutes nos expériences, la mesure de similarité utilisée pour calculer CH et SP est la mesure S^2MP présentée en fin de section 6.1.1.2.

6.1.2.2 Résultats expérimentaux

Au total six expériences ont été menées, puisque nous considérons six ensembles de MSE à partitionner correspondant aux six paires de registres. Selon la paire de registres de langue, la taille de l'ensemble de MSE a significativement varié : allant de 61 121 individus pour le plus petit ensemble, à 2 356 624 individus pour le plus grand (figure 5.8). Dans cette section, ce que nous appelons :

1. La *partition 1* est la partition (autrement dit, les différents groupes de MSE) obtenue après la première étape, c'est à dire le regroupement des motifs selon un seuil minimal de similarité ;
2. la *partition 2* est la partition obtenue après la première itération durant laquelle chaque motif de la *partition 1* a été de nouveau réparti pour rejoindre le groupe du médoïde avec lequel sa similarité est maximale ;
3. la *partition 3* est la partition, dite finale, obtenue après la seconde itération qui, à nouveau, a réparti chaque motif de la *partition 2* dans le groupe du médoïde avec lequel il est le plus similaire.

ID	R_c	R_s	**Coefficient de silhouette**	**Indice de Davies-Bouldin**
1	Familier	Courant	0,31	0,42
2		Soutenu	0,30	0,40
3	Courant	Familier	0,28	0,45
4		Soutenu	0,31	0,40
5	Soutenu	Familier	0,23	0,52
6		Courant	0,27	0,53

TABLE 6.2 – Résultats pour le coefficient de silhouette et l'indice de Davies-Bouldin, pour les partitions 3, des six paires de registres.

Avant de présenter les partitions obtenues, nous validons leurs qualités en utilisant les deux métriques présentées dans la section précédente.

Validation des partitions La table 6.2 montre les résultats pour les *partitions 3* des données des six paires de registres. Les résultats du coefficient de silhouette ont validé la qualité des partitions avec des valeurs plus proches de 1 que de -1 signifiant que la cohésion et la séparation des groupes sont bonnes. Ces résultats ont été confirmés par ceux de l'indice de Davies-Bouldin qui ont tous des valeurs proches de 0. Une légère différence de qualité, entre les partitions des registres familier et courant (ID 1, 2, 3, et 4) par rapport à celles du registre soutenu (ID 5 et 6), est marquée à la fois par les valeurs du coefficient de silhouette, et celles de l'indice de Davies-Bouldin. Elle indique que les regroupements de MSE ont été de moins bonne qualité pour les paires de registres soutenu vs. familier et soutenu vs. courant. De fait, les formes linguistiques spécifiques du registre soutenu respectent toutes la norme linguistique en suivant à la fois la norme grammaticale et la norme d'usage. L'absence de formes linguistiques non standards a homogénéisé l'ensemble des MSE. Cette uniformité des MSE peut expliquer pourquoi leur partitionnement automatique a été plus difficile que pour les registres familier et courant comportant des formes non standardes. La table 6.3 présente le détail des coefficients de silhouette pour les 3 partitions différentes des paires de registres familier et courant. Chaque ligne correspond à une paire de registres de langue et les trois dernières colonnes donnent les scores de silhouette pour les partitions 1, 2 et 3. La table montre que la partition 2 a gagné significativement en qualité par rapport à la partition 1 avec un écart des résultats plus important qu'entre les partitions 2 et 3. Ces validations nous apprennent trois faits principaux :

R_c	R_s	**Score de silhouette**		
		Partition 1	**Partition 2**	**Partition 3**
Familier	Courant	0.19	0.33	0.31
	Soutenu	0.20	0.35	0.30
Courant	Familier	0.17	0.31	0.28
	Soutenu	0.20	0.34	0.31

TABLE 6.3 – Évolution du coefficient de silhouette à travers les 3 partitions des MSE pour les paires de registres familier et courant.

ID	R_c	R_s	**Nbr de motifs dans le plus grand groupe**		
			Partition 1	**Partition 2**	**Partition 3**
1	Familier	Courant	27 630	9 684	5 341
2		Soutenu	17 073	10 894	6 474
3	Courant	Familier	23 856	8 647	6 182
4		Soutenu	3 811	1 877	1 620
5	Soutenu	Familier	200 886	39 993	38 976
6		Courant	131 783	41 641	30 072

TABLE 6.4 – Taille maximale des groupes pour les trois partitions des six paires de registres.

1. les groupes de MSE sont de bonne qualité avec des MSE très similaires au sein d'un même groupe et bien différents des MSE d'autres groupes ;
2. l'homogénéité des formes linguistiques décrites par les MSE impactent la qualité des partitions avec des partitions de moins bonne qualité pour le registre soutenu ;
3. l'organisation en trois temps de RGMSE semble pertinente avec une qualité de la seconde partition meilleure que la première.

Détails quantitatifs sur les groupes obtenus pour les différentes partitions La première étape du clustering, qui a permis de former des groupes de motifs similaires selon un seuil de similarité minimale ($minSim$), a eu l'avantage de réduire rapidement l'espace de recherche grâce à un tirage sans remise des motifs. Toutefois, ce tirage sans remise a montré deux limites principales : la première est de rendre la *partition 1* très dépendante de l'ordre dans lequel les motifs ont été tirés ; la seconde est d'obtenir des groupes de MSE volumineux lorsque la valeur de $minSim$ n'était pas élevée. Les deux itérations suivantes, qui ont redistribué les motifs, ont répondu à ces deux limites. Elles ont permis, indépendamment de l'ordre du tirage, à un MSE de rejoindre le groupe avec lequel il était le plus similaire en calculant sa similarité avec son médoïde. Cette redistribution

ID	R_c	R_s	**Nombres de groupes de motifs par partition**		
			Partition 1	**Partition 2**	**Partition 3**
1	Familier	Courant	1 803	1 781	1 735
2		Soutenu	1 381	1 369	1 338
3	Courant	Familier	1 820	1 800	1 753
4		Soutenu	771	760	740
5	Soutenu	Familier	3 437	3 390	3 290
6		Courant	3 552	3 516	3 475

TABLE 6.5 – Nombre de groupes obtenus lors des partitions 1, 2 et 3 selon les couples de registres.

des MSE a réduit les groupes de motifs volumineux en rééquilibrant la partition. La table 6.4 montre, pour les six paires de registres et les trois partitions, le nombre de motifs représentants contenu dans le plus grand groupe de chaque partition. Chaque ligne correspond à une paire de registres et les trois dernières colonnes donnent la taille du groupe pour les partitions 1, 2 et 3. De manière générale, au fur et à mesure des itérations, une diminution de la taille maximale des groupes peut être constatée entre chaque partition. De plus, la table montre que la première étape, générant la *partition 1*, a bien favorisé la création de grands groupes : la table indique une réduction, de la taille maximale entre la *partition 1* et la *partition 2*, nette et plus importante que celle entre la *partition 2* et *partition 3*.

Le nombre de groupes, contenant une grande quantité de MSE, a représenté une proportion minoritaire des partitions. Pour la *partition 3*, de l'ensemble de motifs du familier vs. courant, si l'on trie les groupes selon leurs tailles, le premier tiers (c'est à dire les groupes les plus petits) a représenté 72% des groupes de la partition totale. À l'inverse, le dernier tiers (c'est à dire les groupes les plus grands) a représenté seulement 11% des groupes. Cette tendance s'est retrouvée pour toutes les autres paires de registres. Par exemple, la paire des registres soutenus vs. courant, où les plus petits groupes ont représenté 66% de la partition, tandis que les plus grands groupes en ont représenté 14%. Prenons également, la paire des registres courant vs. soutenu, où les plus petits groupes ont constitué 73% de la partition, contre 10% pour les plus grands. Cette répartition des groupes, selon leurs tailles, a indiqué que la mesure de similarité utilisée a bien permis de rassembler des MSE selon leurs similitudes de manière fine avec une majorité de petits groupes.

La table 6.5 détaille le nombre de groupes obtenus lors des *partitions 1*, *2* et *3*, selon

les couples de registres. Une tendance similaire pour les six paires de registres peut être constatée : le nombre de groupes a constamment diminué au fur et à mesure des itérations. Cela indique que la redistribution des motifs a été cohérente. Elle aurait été incohérente si le nombre de groupes augmentait puis diminuait : cela aurait signifié que des motifs jugés similaires lors de la *partition 2*, aurait ensuite été jugés différents lors de la *partition 3*. Ces résultats quantitatifs des trois partitions ont montré que les limites, liées au tirage sans remise de la première itération de RGMSE, ont été déjouées par les itérations suivantes puisque : la répartition des motifs, entre les différents groupes, a été équilibrée ce qui a été illustré par un nombre de groupes qui a diminué de manière constante entre la *partition 1* et la *partition 3* ; la taille maximale des groupes a été significativement réduite dès la *partition 2*.

Détails qualitatifs sur les groupes obtenus par les différentes partitions Nous venons de présenter les résultats d'un point de vue quantitatif, cette section présente les résultats expérimentaux d'un point de vue qualitatif. Si les groupes contenant peu de motifs représentent la majorité des groupes, alors cela veut-il dire qu'ils contiennent des MSE très spécifiques ? Afin de faire une première vérification de la qualité de la partition des données, cette section expose l'exploration manuelle de ces groupes qui contiennent peu de MSE ; avant au contraire de regarder les MSE des groupes qui en contiennent beaucoup. Tout d'abord, nous supposons que si ces MSE n'ont pas été regroupés avec d'autres groupes c'est qu'ils sont très différents du reste des MSE. Pour vérifier cette hypothèse, nous avons examiné la taille des séquences, c'est à dire le nombre d'itemsets qu'elles contiennent. Plus les séquences sont longues, plus les descripteurs linguistiques qu'elles représentent contiennent de mots et de traits linguistiques. Étant donné que les MSE sont issus de motifs clos, plus les motifs sont longs, plus ils incluent de sous-séquences ayant la même fréquence et plus elles réduisent la redondance des MSE. De longs MSE ont alors eu une plus grande probabilité d'être différents d'autres MSE. Par exemple, considérons la table 6.6 qui donne six motifs fréquents non clos. Chaque ligne présente un motif associé à son *tid* et à sa fréquence. En choisissant le motif 5 contenant le plus grand nombre d'items nous ne réduisons pas la probabilité d'avoir un motif similaire. Effectivement, le motif 4 n'a qu'un item qui diffère du motif 5. En revanche, si nous conservons que les motifs clos, alors seulement les motifs 5 et 6 sont conservés puisque les motifs 1, 2, 3 et 4 sont inclus dans le motif 5. C'est pourquoi, plus un motif clos est long, moins il y a de chance d'avoir un autre motif clos similaire. Toutefois, lorsque l'on explore

tid	séquence	fréquence
1	$\langle\{a\}\rangle$	5%
2	$\langle\{c\}\rangle$	5%
3	$\langle\{a,b\}\rangle$	5%
4	$\langle\{a,b\},\{d\}\rangle$	5%
5	$\langle\{a,b,c\},\{d\}\rangle$	5%
6	$\langle\{f\},\{b,c\}\rangle$	5%

TABLE 6.6 – Exemples de six MS fréquents ayant tous la même fréquence

ID	**Registre 1**	**Registre 2**	**% de séquences avec soit un lemme, soit un sous-mot dans l'ensemble des séquences uniques**
1	Familier	Courant	82 %
2		Soutenu	79 %
3	Courant	Familier	71 %
4		Soutenu	79 %
5	Soutenu	Familier	83 %
6		Courant	82 %

TABLE 6.7 – Pourcentage de séquences, qui contiennent soit un lemme, soit un sous-mots, dans l'ensemble des séquences uniques de la *partition 3* pour les 6 paires de registres.

les groupes contenant un seul MSE, ces derniers semblent avoir une majorité de motifs très courts (entre 1 et 2 itemsets). Par exemple, pour la paire de registres familier vs. courant, les séquences contenant 1 à 2 itemsets représentent 84% des séquences uniques des groupes de MSE[3] ; pour la paire de registres courant vs. familier elles représentent 73% ; et pour la paire de registres soutenu vs. courant 86%. Ce constat va à l'encontre de notre intuition qui supposait que les séquences uniques sont mises à l'écart, c'est à dire contenues dans des groupes dont elles sont l'unique motif, à cause de leur grande taille (longueur). Alors, pourquoi les séquences uniques sont-elles majoritairement des séquences courtes, c'est à dire constituées de peu d'itemsets (1 à 2) ? L'explication se trouve dans les types de traits linguistiques contenus dans ces séquences. La table 6.7 détaille les pourcentage de séquences qui contiennent, soit un lemme, soit un sous-mot, dans l'ensemble des séquences uniques de la *partition 3* pour les six paires de registres. La table montre que la majorité des séquences uniques contient des informations de type lexicale (les lemmes), ou bien de type morphologique avec des formes fléchies (les sous-

3. Ce que nous appelons une *séquence unique* est un MSE qui est contenu dans un groupe dont il est le seul individu.

mots). Ces types de traits linguistiques possèdent une dimension informative importante grâce aux formes lexicalisées qu'ils comportent. Pour illustrer cette importante dimension informative, des exemples de séquences uniques, qui contiennent 1 ou 2 itemsets pour quatre paires de registres de la *partition 3*, sont donnés ci-dessous avec des exemples de tweets (issus du corpus *TREMoLo-Tweet*) :

(1) Familier vs. Courant :

1. ⟨ (*sous-mot:dr_*) ⟩ : ptdr vendredi 38 degrés si c'est pas pour mourir c'est pq srx
2. ⟨ (*sous-mot:sh_*) ⟩ : @X elle traine trop avec des trash léna, c'est pour ça que j'ai encore du mal
3. ⟨ (*sous-mot:_Mais_*) ⟩ : @X Mais vas-y toi moi j'appelle ça yeti fin du débat 😭
4. ⟨ (*syntaxe:sujet*), (*sous-mot:_va_*) ⟩ : @X Jte sort des équations de 1ère S et toi en tant que T stmg tu va voir flou c facile a dire

Les exemples 1 et 2 montrent que les terminaisons spécifiques (portées par les sous-mots) du registre familier par rapport au registre courant ont bien été isolées des autres MSE et n'ont pas été noyées dans de grands groupes de motifs. De même, les sous-mots ont permis de faire ressortir des formes fléchies dans leurs intégralités comme l'illustrent les exemples 3 et 4, ou bien l'exemple 5.

(2) Familier vs. Soutenu :

5. ⟨ (*sous-mot:_top_*) ⟩ : les memes sur adele sont trop top putain http ://path.com
6. ⟨ (*lemme:avoir*),(*lemme:faire*) ⟩ : @X je lui ai dit de parler astro... c un homme de 39 ans, jsp pk j'ai fait ça
7. ⟨ (*pos:verbe*),(*lemme:tout*) ⟩ : Euh l'album de kaaris il est énerver j'aime tout les sons
8. ⟨ (*morpho:singulier*), (*lemme:savoir*) ⟩ : Parcontre landorly il sait pas que leris c'est son daron ! ! ?

Les exemples 6, 7 et 8 montrent que les motifs courts avec des lemmes sont facilement interprétables tout en conservant une certaine capacité à être généralisables.

(3) Soutenu vs. Familier :

9. ⟨ (*sous-mot:ation_*) ⟩ : RT @X : [OPINION] « Quelle est la clé pour une intégration en français réussie ? » #polqc #Immigration #JDQ http ://path.com
10. ⟨ (*sous-mot:isme_*) ⟩ : Entre l'exécration de soi et l'égotisme accointé au narcissisme, il y a, je pense, un juste milieu ! #PlusDe70kgEtSereine http ://path.com

11. ⟨ (*sous-mot:_Nous_*) ⟩ : @X @X @X @X @X Pour l'instant, le roi n'a exprimé que ses regrets. Nous attendons avec impatience les excuses qui ouvriront certainement les voies aux réparations. Ceci étant dit, j'apprécie votre combat à la RDC souveraine face surtout aux ingérences de @X . Vive la convention de Vienne

12. ⟨ (*lemme:après*),(*pos:déterminant*) ⟩ : Julian Alaphilippe va nous faire une épopée en jaune c'est une certitude. Le conservera t-il après le col de la Loze ? #TDF2020

Contrairement aux exemples 1 et 2 (caractéristiques du registre familier par rapport au registre courant), les terminaisons illustrées par les exemples 9 et 10 sont standardes et intégrées à la norme linguistique. Leur intégration à la norme est mise en exergue par leur reconnaissance par des institutions officielles. Par exemple, la première terminaison - le suffixe *-ation* - est définie dans le TLFI [4] comme le « *Suffixe issu du latin "-tionem", entrant dans la construction de nombreux substituts féminins qui expriment une action ou le résultat de cette action.* ». La seconde terminaison - le suffixe *-isme* - est décrite par l'Académie Française comme très productive : « *Le suffixe "-isme" est très productif. Il entre dans la composition de mots désignant des courants de pensée philosophiques ou politiques.* » [5]. L'exemple 11 (tout comme l'exemple 3) montre que les sous-mots ont permis de capturer des informations telle que la place du mot dans la phrase avec la majuscule. Enfin, les exemples 12 et 13 illustrent des MSE spatio-temporels qui structurent le texte.

(4) Soutenu vs. Courant :

13. ⟨ (*lemme:avant*) ⟩ : La « colère du Grand Liban » embrase le centre-ville - Le rassemblement place des Martyrs était pacifique avant de dégénérer en affrontements avec l'armée et les forces de l'ordre. http ://path.com via @X

14. ⟨ (*lemme:certain*) ⟩ : Heureusement que l'on n'a pas imposé le #MasqueObligatoire partout comme certains le prônaient. Cela permet d'avoir une réponse graduée face au « rebond », d'éviter des mesures coercitives trop délirantes et de pousser au port des #masques à un moment où cela paraît vraiment utile.

15. ⟨ (*lemme:porter*) ⟩ : Cessez de porter atteinte à la souveraineté de mon pays d'origine, l'Algérie ! Avez-vous oublié la convention des droits de l'homme. La Belgique, ma terre d'accueil, ma seconde patrie ne deviendra jamais une zone de non-droit.

4. TLFI pour **T**résor de la **L**angue **F**rançaise **I**nformatisé, la définition est accessible via ce lien : `https://www.cnrtl.fr/definition/-tion`.

5. L'article complet sur la construction en *-isme* est accessible via ce lien : `https://www.academie-francaise.fr/construction-en-isme`.

16. ⟨ (*sous-mot:_@*), (*syntaxe:racine*) ⟩ : RT @X : #DEBATGG - Face aux faits de violence, @X parle d'« incivilités »... Hors sol ? #GGRMC

Les exemples 14 et 15 montrent des MSE qui ne comportent qu'un trait linguistique (les lemmes). Ces MSE facilement interprétables permettent, par exemple, de construire un lexique spécifique du registre soutenu par rapport au registre courant. L'exemple 16 quant à lui est particulièrement intéressant car il renvoie à un trait linguistique intégré à l'ensemble de motifs linguistiques proposé pour l'analyse des CMO telle que nous l'avons proposée dans notre guide d'annotation p. 15 [6] : *Mention de l'identifiant de l'utilisateur dans un syntagme*. Dès lors, le fait que la *partition 3* a isolé ce MSE dans un groupe, dont il était le seul individu, a confirmé sa saillance pressentie lors de l'analyse linguistique du corpus.

À l'inverse, si l'on regarde les plus grands groupes (c'est à dire les groupes qui ont fait partie du dernier tiers des groupes triés par ordre croissant de taille), la proportion des séquences, qui comportaient soit un lemme soit un sous-mot, était minoritaire. Autrement dit, lorsque les groupes comportaient beaucoup de MSE, ces derniers portaient majoritairement des traits linguistiques plus génériques (tels que des informations syntaxiques ou sur les catégories grammaticales). Par exemple, pour la *partition 3* de la paire de registres familier vs. soutenu, pour un groupe $G_{F|S}$ contenant 939 MSE, seulement 12% d'entre eux comportaient soit des lemmes soit des sous-mots. De même, pour la paire de registres soutenu vs. familier, un groupe $G_{S|F}$ de 1 453 MSE en contenait uniquement 19% : trois exemples de MSE contenus dans $G_{S|F}$:

1. ⟨ (*morpho:singulier*), (*morpho:singulier, syntaxe:objet*), (*syntaxe:dépendant*), (*morpho:masculin, syntaxe:préposition*) ⟩
2. ⟨ (*morpho:singulier*), (*morpho:singulier, syntaxe:objet*), (*syntaxe:dépendant, pos:nom-commun*), (*morpho:masculin, syntaxe:préposition*) ⟩
3. ⟨ (*morpho:singulier*), (*morpho:feminin, syntaxe:objet*), (*syntaxe:dépendant, pos:nom-commun*), (*morpho:singulier, syntaxe:préposition*) ⟩

6.1.3 Synthèse

Dans cette section nous avons présenté les travaux réduisant la redondance des MSE en utilisant notre algorithme RGMSE. Les résultats obtenus ont montré la robustesse de RGMSE face à un large ensemble de motifs à regrouper car il a pu traiter la totalité des

6. `https://hal.archives-ouvertes.fr/hal-03218217/`

MSE. De plus, les évaluations faites ont indiqué que les partitions sont de qualité avec des groupes de MSE très similaires et très différents des autres groupes. En outre, nous avons vu que les itérations de RGMSE augment la qualité de la partition avec une nette amélioration entre la première et la deuxième itération. Ensuite, l'exploration qualitative des groupes obtenus a mis au jour l'influence des descripteurs linguistiques sur les groupes de MSE ainsi que la capacité de RGMSE à isoler des MSE intéressants mais peu nombreux.

La section suivante détaille comment nous avons réduit la quantité des MSE à partir des groupes obtenus.

6.2 Réduction du nombre de motifs séquentiels émergents

Après avoir réduit la redondance des MSE en les répartissant dans des groupes de MSE similaires, nous proposons d'en réduire le nombre en tirant un motif représentant pour chacun des groupes. À la fin de ce processus, nous obtenons un ensemble de motifs représentants plus exploitable que l'ensemble complet des MSE car résultant de réductions de leur redondance et de leur quantité. Comme illustré par la figure 6.4, la sélection de motifs représentants a résulté de la sélection, pour chaque groupe de MSE, d'un seul MSE considéré comme le représentant de son groupe. Dès lors, afin de sélectionner un motif représentant, il a fallu poser les critères faisant d'un MSE un bon représentant pour son groupe. Autrement dit, quelles étaient les propriétés que devait posséder un MSE pour pouvoir représenter l'ensemble des MSE de son groupe ?

6.2.1 Présentations des méthodes de sélections de motifs représentants

La question posée par le choix d'un MSE représentant est la suivante : quels sont les éléments nécessaires à un MSE pour qu'il représente au mieux son groupe ? N'ayant pas tranché cette question, nous avons proposé différentes manières de sélectionner un MSE représentant. La première sélectionne le médoïde comme étant le MSE représentant. Elle est considérée comme la métrique de référence basique à laquelle comparer les autres. Sa qualification de *basique* tient au fait qu'elle ne soit pas adaptée au format particulier des MSE.

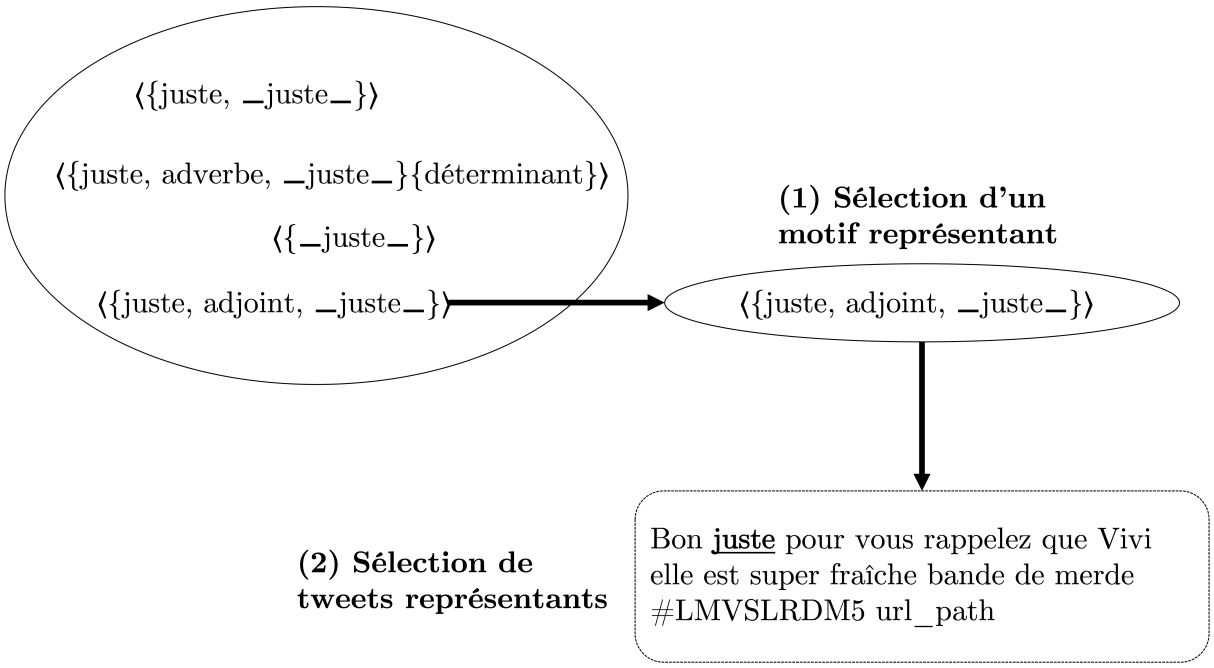

FIGURE 6.4 – Exemples des étapes pour obtenir un motif représentant et un tweet représentant à partir d'un groupe de motifs.

La seconde méthode de sélection du MSE représentant se base sur la fréquence des itemset au sein du groupe. Elle suppose que les motifs ayant des itemsets très fréquents au sein du groupe sont de bons représentants de ce dernier. Étant donné un groupe G et un motif $S = \langle I_1, ...I_n \rangle$, la métrique $ItemsetFreq(S, G)$ (détaillée équation 6.9) rend compte de la somme des fréquences relatives, notée $\frac{freq_G(I_i)}{|G|}$, de chaque itemset I contenu dans S. Le MSE représentant du groupe G est celui qui maximise $ItemsetFreq$.

$$ItemsetFreq(S, G) = \sum_{k=1}^{|S|} \frac{freq_G(I_k)}{|G|} \tag{6.9}$$

$$ItemFreq(S, G) = \sum_{i=1}^{|S|} \frac{freq_G(i_k)}{|G|} \tag{6.10}$$

La troisième métrique reprend la métrique $ItemsetFreq(S, G)$, mais considère la fréquence relative des *items* dans un groupe G et non plus celle des *itemsets* (voir l'équation 6.10). Dès lors, $ItemFreq(S, G)$ rend compte de la somme des fréquences relatives, notée $\frac{freq_G(i_k)}{|G|}$, de chaque item i_k contenu dans le motif S. À nouveau, le MSE représentant de G est celui qui maximise $ItemFreq$. Pour obtenir un MSE représentant rendant compte

ID	Registre 1	Registre 2	Ensemble de MSE	Ensemble de MR	Réduction de %
1	Familier	Courant	326 552	1 734	99,47 %
2		Soutenu	226 938	1 338	99,41 %
3	Courant	Familier	416 554	1 753	99,58 %
4		Soutenu	61 121	740	98,79 %
5	Soutenu	Familier	2 330 679	3 290	99,86 %
6		Courant	2 356 624	3 475	99,85 %

TABLE 6.8 – Détails quantitatifs sur la réduction de l'ensemble complet des MSE en un sous-ensemble composé de motifs représentants notés MR.

de sa capacité plus ou moins importante à caractériser le registre cible, nous pouvons pondérer $ItemsetFreq$ et $ItemFreq$ par le taux de croissance du MSE.

$$ItemsetFreqGR(S, G) = ItemsetFreq(S, G) \times TauxCroissance(S) \tag{6.11}$$

$$ItemFreqGR(S, G) = ItemFreq(S, G) \times TauxCroissance(S) \tag{6.12}$$

Enfin, nous proposons une seconde manière de pondérer les scores $ItemsetFreq$ et $ItemFreq$ d'un motif S qui repose sur le *taux de recouvrement* de S. Étant donné un corpus D, le taux de recouvrement d'un motif S, noté $Cov_D(S)$, indique le nombre de textes dans lequel S apparaît.

$$ItemsetFreqCov(S, C, D) = ItemsetFreq(S, C) \times Cov_D(S) \tag{6.13}$$

$$ItemFreqCov(S, C, D) = ItemFreq(S, C) \times Cov_D(S) \tag{6.14}$$

$ItemsetFreqCov$ et $ItemFreqCov$ (équation 6.13 et 6.14) reposent sur l'hypothèse qu'un bon MSE représentant est un MSE contenu dans le plus grand nombre de textes du corpus D. Les expériences détaillées à la suite de ce paragraphe montrent l'évaluation automatique de ces méthodes de sélection et permettent d'en discerner la plus pertinente. Elles montrent notamment qu'une méthode adaptée au MSE est plus pertinente qu'une méthode généraliste.

6.2.2 Résultats des sélections de motifs représentants

Tout d'abord, notre démarche a fortement réduit le nombre de résultats. La table 6.8 détaille le nombre de MSE, pour l'ensemble complet de MSE et l'ensemble de motifs repré-

Méthode	*Courant vs. Familier*
Med(C)	⟨ (lemme:*devoir*) ⟩
ItemsetFreq	⟨ (sous-mot:*rrr_*, pos:*nom-propre*) ⟩
ItemFreq	⟨ (sous-mot:*rrr_*) ⟩
ItemsetFreqGR	⟨ (sous-mot:*rrr_*, pos:*nom-propre*) ⟩
ItemFreqGR	⟨ (sous-mot:*rrr_*, pos:*nom-propre*) ⟩
ItemsetFreqCov	⟨ (sous-mot:*rrr_*) ⟩
ItemFreqCov	⟨ (sous-mot:*rrr_*, pos:*nom-propre*) ⟩

Méthode	***Soutenu vs. Courant***
Med(C)	⟨ (sous-mot:*_entre_*) ⟩
ItemsetFreq	⟨ (sous-mot:*_entre_*, lemme:*entre*, pos:*préposition*) ⟩
ItemFreq	⟨ (sous-mot:*_entre_*, lemme:*entre*,pos:*préposition*) ⟩
ItemsetFreqGR	⟨ (sous-mot:*_entre_*, syntaxe:*modifieur*) ⟩
ItemFreqGR	⟨ (sous-mot:*_entre_*, syntaxe:*modifieur*) ⟩
ItemsetFreqCov	⟨ (sous-mot:*_entre_*, lemme:*entre*, pos:*préposition*) ⟩
ItemFreqCov	⟨ (sous-mot:*_entre_*, lemme:*entre*, pos:*préposition*) ⟩

TABLE 6.9 – Exemples de motifs représentants sélectionnés selon les sept méthodes pour différentes paires de registres.

sentants : une réduction significative pour les six paires de registres peut être constatée. Cette réduction a permis d'écarter environ 99% des MSE pour les six paires de registres. L'exploration manuelle des motifs représentants selon les méthodes de sélection utilisées a confirmé la qualité de la partition, avec des motifs représentants pertinents en termes de caractérisation de registres de langue ; et a montré la difficulté à trancher quant à une méthode de sélection meilleure qu'une autre. Les sept méthodes de sélection proposées ont permis d'extraire des représentants qui s'adaptent aux besoins de l'utilisateur. Dans le cadre de ce travail, les sept méthodes ont été utilisées pour sélectionner des motifs représentants. La table 6.9 donne des exemples de motifs représentants, pour les paires de registres familier vs. soutenu, et soutenu vs. courant. Les premiers renvoient à des formes linguistiques telles que celles soulignées dans les exemples (126) et (127) ; tandis que les seconds décrivent celles soulignées dans les exemples (128) et (129).

(126) @X @X @X Des grand Denjiro malina Marié, tout sa pour procréer un énergumènes tel que jctr<u>rrr</u> eeeesh foorce a vous deux hein on est pas ensemble sur le coup 😭 😭 😭 😭

(127) je suis mort Djoko disqualifié parce qu'il a mis une tête à un juge ptdrrrrrr<u>rrr</u>

ID	Descripteur	Motif représentant	R_c	TC
		Niveau lexical		
1	Element ponctuant	⟨{lemme :tout, sous-mot :_tout_}⟩	F	1.3
2	Onomatopées	⟨{sous-mot :_ou}⟩	F	$+\infty$
		⟨{sous-mot :h_}⟩	F	2.2
3	"là" ponctuants	⟨{syntaxe :modifieur, sous-mot :_là_}⟩	F	2.4
5	Planificateurs du discours	⟨{lemme :après, pos :préposition} {pos :déterminant, syntaxe :déterminant}⟩	S	1.7
		Niveau morphosyntaxique		
6	Contraction de "cela" en "ça"	⟨{lemme :ça, sous-mot :_ça_, pos :pronom, nombre :singulier} {personne :troisième}⟩	F	$+\infty$
7	Négation sans "ne"	⟨{syntaxe :sujet} {pos :verbe, nombre :singulier} {lemme :pas, sous-mot :_pas_, pos :adverbe}⟩	F	$+\infty$
8	Sujet "on" transposé en "nous"	⟨{lemme :nous, syntaxe :sujet, nombre :pluriel, personne :1e}⟩	S	$+\infty$
10	Terminaison en "-ouze"	⟨{sous-mot :ze_}⟩	F	$+\infty$
11	Terminaison en "-o"	⟨{sous-mot :o_}⟩	F	1.3
12	Verbe "être" au singulier devant un syntagme nominal singulier	⟨{lemme :être, nombre :singulier} {pos :déterminant, nombre :singulier, genre :féminin} {syntaxe :objet, nombre :singulier, pos :nom commun}⟩	F	1.3
13	"ça" + verbe	⟨{lemme :ça, sous-mot :_ça_, pos :pronom, nombre :singulier} {pos :verbe, nombre :singulier}⟩	F	$+\infty$
		Niveau syntaxique		
28	Effacement du "il" impersonnel	⟨lemma :y, pos :sujet pos :ponctuation, syntaxe :ponctuation, lemme :avoir⟩	F	$+\infty$

TABLE 6.10 – Descripteurs issus de (MEKKI, BATTISTELLI, LECORVÉ et al. 2018) et présentés table 2.2 retrouvés parmi les motifs représentants.

(128) Honorer Dieu à l'aide de la religion naturelle, sans égard pour les religions patriarcales. Voici une petite histoire qui conte le dialogue entre un moine et un homme sans réligions. #Voltaire #Lumières #Religion #Laicité url_path

(129) !! Risques d'injonctions contradictoires entre la prévention #canicule et la prévention #Covid_19 les explications de @X 👉 url_path url_path

Ces exemples mettent en avant la similarité des résultats obtenus avec les sept types de motifs représentants. La ressemblance des motifs représentants entre eux rend la sélection d'un meilleur motif pour représenter son groupe impossible pour un expert humain.

Enfin, une première manière de regarder la qualité des motifs représentants obtenus est

de les comparer avec la liste de descripteurs linguistiques issus de la littérature scientifique et regroupés dans (Mekki, Battistelli, Lecorvé et al. 2018). Cette liste représente la seule base de vérité que nous pourrions utiliser pour évaluer la pertinence des motifs représentants. La table 6.10 donne les descripteurs issus de (Mekki, Battistelli, Lecorvé et al. 2018) et présentés table 2.2 qui ont été retrouvés parmi les motifs représentants. Chaque ligne donne un descripteur avec : le même ID qu'il a dans la table 2.2, le motif représentant qui le représente, le R_c caractérisé et son taux de croissance. La moitié des descripteurs a été retrouvée. Cependant, le niveau syntaxique et phonologique ne sont pas du tout représentés car aucun de leurs descripteurs n'a été retrouvé parmi les motifs représentants. Cette comparaison est insuffisante pour valider ou invalider la pertinence des motifs représentants, c'est pourquoi nous avons conduit d'autres travaux pour les évaluer. Ces travaux sont présentés dans la section suivante.

6.3 Évaluation des résultats expérimentaux

Nous présentons dans cette section une évaluation automatique et une évaluation perceptuelle des ensembles de motifs représentants. Nous en avons proposées deux, car elles sont complémentaires. La première adopte un angle d'analyse macro des résultats en regardant les grandes tendances s'en dégageant, tandis que la seconde adopte un prisme d'analyse micro des résultats en les examinant un par un.

L'évaluation automatique a l'avantage d'examiner les sept sous-ensembles de motifs représentants et la totalité des motifs qu'ils contiennent. Son objectif est double : valider la capacité des MSE à caractériser un registre cible A en le distinguant d'un registre source B ; voir si un type de motifs représentants est meilleur qu'un autre pour caractériser A par rapport à B.

Quant à l'évaluation perceptuelle, elle a l'intérêt d'examiner les motifs représentants un par un. Cependant, comme cet examen prend du temps, seulement une petite partie d'un seul type de motifs représentants est évaluée (en l'occurrence ceux sélectionnés avec $itemFreqGR$). Son premier but est de vérifier si la présence d'un motif représentant caractérisant le registre A dans un texte permet à un évaluateur de percevoir A. Son second but est de vérifier que le taux de croissance est un indicateur pertinent pour notre tâche.

Nous présentons l'évaluation automatique suivie de l'évaluation perceptuelle.

6.3.1 Évaluation automatique

Pour évaluer la pertinence d'un sous-ensemble de motifs représentants, nous avons supposé qu'il devait permettre à un classifieur de distinguer des textes d'un registre A d'un registre B, c'est-à-dire à étiqueter correctement un texte donné en lui attribuant soit le registre A soit le registre B. Pour découvrir si une méthode de sélection de motifs représentants était meilleure, nous avons supposé qu'elle amènerait à des prédictions de meilleures qualités. Nous détaillons le protocole expérimental, puis nous présentons les résultats obtenus.

6.3.1.1 Protocole expérimental

Modèle d'apprentissage automatique Le but était d'évaluer la capacité du sous-ensemble de motifs représentants à représenter l'ensemble complet des *MSE*, et à bien caractériser un registre A par rapport à un registre B. Dès lors, le classifieur mis en place devait être un *classifieur binaire* : pour un texte donné, le classifieur devait prédire s'il appartenait à la classe A ou à la classe B.

Choix du classifieur L'algorithme de *forêt d'arbres décisionnels* (*random forest classifier* en anglais) introduit par (Breiman 2001) a été choisi comme classifieur. Une *forêt d'arbres décisionnels* est un *méta-estimateur*, qui utilise un ensemble d'arbres de décision indépendants. La prédiction finale fait la moyenne de ces prédictions indépendantes. Le choix de cet algorithme a reposé sur trois critères : sa capacité à faire une prédiction binaire ; son implémentation disponible proposée par *scikit-learn*[7] ; la possibilité d'analyser les descripteurs d'apprentissage les plus importants pour le classifieur.

Descripteurs d'apprentissage Le classifieur binaire a du prédire si un texte appartenait au registre A ou bien au registre B. L'ensemble des descripteurs d'apprentissage, noté MR'_{AB}, a été constitué à partir de deux ensembles de motifs représentants : l'ensemble caractérisant le registre A par rapport au registre B ; l'ensemble caractérisant le registre B par rapport au registre A. Étant donné l'ensemble MR'_{AB} de n descripteurs d'apprentissage, et un ensemble de m textes à prédire, la matrice T de taille (n,m) a été construite. Le classifieur prend cette matrice T en entrée.

7. L'algorithme *Random Forest* est disponible via ce lien : `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

Hyperparamètres et sous-corpus Les hyperparamètres ont été définis grâce à l'outil SearchGrid [8] qui a calculé leurs valeurs optimales. Le corpus TREMolo-Tweets (228 505 tweets) a été divisé en trois sous-corpus : un sous-corpus d'entraînement (80% du corpus complet), un sous-corpus de validation (10%) et un sous-corpus de test (10%).

Métriques d'évaluation Pour évaluer les prédictions faites par le classifieur trois mesures classiques en TAL ont été utilisées : la *précision*, le *rappel* et la *f-mesure*. La *précision* (équation 6.15) a représenté la proportion de documents pertinents correctement prédits, comme appartenant à la classe A, par rapport au nombre de documents total de documents véritablement de la classe A. Autrement dit, la *précision* a mesuré le nombre de documents pertinents correctement prédits.

$$\text{précision}_A = \frac{\text{nombre de documents correctement attribués à la classe } A}{\text{nombre de documents attribués à la classe } A} \tag{6.15}$$

Le *rappel* (équation 6.16) a représenté la proportion de documents pertinents prédits par rapport au nombre total de documents pertinents présents dans le corpus. Autrement dit, le *rappel* a mesuré le nombre de documents pertinents que le classifieur a manqué.

$$\text{rappel}_A = \frac{\text{nombre de documents correctement attribués à la classe } A}{\text{nombre de documents appartenant à la classe } A} \tag{6.16}$$

La *f-mesure* (équation 6.17) a représenté un compromis en faisant la moyenne harmonique de la précision et du rappel.

$$F = 2 \cdot \frac{(\text{ précision } \cdot \text{ rappel })}{(\text{ précision } + \text{ rappel })} \tag{6.17}$$

6.3.1.2 Résultats expérimentaux

Au total, trois classifieurs binaires ont été implémentés, pour les trois couples de registres suivants : familier vs. courant, courant vs. soutenu, et soutenu vs. familier. Pour chacun de ces trois classifieurs, sept ensembles de descripteurs d'apprentissages ont été testés afin de voir comment les méthodes de sélection des motifs représentants impactaient la performance des classifieurs.

8. SearchGrid est un outil proposé par *scikit-learn* disponible via ce lien : `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`.

ID	Méthode	Précision	Rappel	F-mesure
1	*Med(C)*	0,797	0,800	0,792
2	*ItemsetFreq*	0,782	0,784	0,774
3	*ItemFreq*	0,795	0,798	0,790
4	*ItemsetFreqGR*	0,808	0,811	0,805
5	*ItemFreqGR*	0,811	0,814	0,809
6	*ItemsetFreqCov*	0,792	0,794	0,785
7	*ItemFreqCov*	0,796	0,798	0,790

TABLE 6.11 – Résultats du classifieur binaire Random Forest pour les registres familier vs. courant pour les 7 méthodes de sélections des motifs représentants.

ID	Méthode	Précision	Rappel	F-mesure
1	*Med(C)*	0,891	0,895	0,879
2	*ItemsetFreq*	0,899	0,901	0,866
3	*ItemFreq*	0,899	0,903	0,890
4	*ItemsetFreqGR*	0,909	0,913	0,904
5	*ItemFreqGR*	0,912	0,917	0,910
6	*ItemsetFreqCov*	0,895	0,899	0,885
7	*ItemFreqCov*	0,899	0,902	0,877

TABLE 6.12 – Résultats du classifieur binaire Random Forest pour les registres soutenu vs. courant pour les sept méthodes.

Détails des résultats Les tables 6.11, 6.12 et 6.13 détaillent, pour les couples de registres familier vs. courant et soutenu vs. courant, les mesures de Précision, Rappel et F-mesure. Elles montrent que les résultats ont été globalement bons, et qu'aucune méthode de sélection des motifs représentants ne s'est démarquée. Cependant, nous pouvons noter que les ensembles de descripteurs d'apprentissage sélectionnés grâce à *ItemFreqGR* ont permis d'obtenir des résultats légèrement supérieurs aux autres pour les trois classifieurs binaires. À l'inverse, les ensembles de descripteurs d'apprentissage filtrés grâce à *ItemsetFreq* ont obtenu de moins bons résultats pour les trois classifieurs. Ainsi, ces premiers résultats ont montré la pertinence des motifs représentants sélectionnés selon les sept méthodes proposées avec une légère supériorité pour les motifs représentants obtenus avec *ItemFreqGR*, et une légère infériorité pour ceux obtenus avec *ItemsetFreq*.

Exploration des descripteurs d'apprentissage Une seconde manière d'examiner la pertinence des descripteurs d'apprentissage (composés de motifs représentants) est d'explorer les descripteurs les plus importants pour le classifieur lors de ses prédictions. Pour

ID	Méthode	Précision	Rappel	F-mesure
1	*Med(C)*	0,914	0,941	0,941
2	*ItemsetFreq*	0,935	0,935	0,935
3	*ItemFreq*	0,945	0,947	0,945
4	*ItemsetFreqGR*	0,944	0,944	0,944
5	*ItemFreqGR*	0,945	0,947	0,945
6	*ItemsetFreqCov*	0,941	0,941	0,941
7	*ItemFreqCov*	0,938	0,939	0,938

TABLE 6.13 – Résultats du classifieur binaire Random Forest pour les registres soutenu vs. familier pour les sept méthodes de sélection des MR.

ID	***Importance features***	***Permutation feature importance***
1	⟨ (*sous-mot:rrr_*) ⟩	⟨ (*sous-mot:rrr_*) ⟩
2	⟨ (*sous-mot:_j*, *syntax:modifieur*) ⟩	⟨ (*sous-mot:_md*, *sous-mot:r_*) ⟩
3	⟨ (*sous-mot:_md*, *sous-mot:r_*) ⟩	⟨ (*sous-mot:dr_*) ⟩
4	⟨ (*sous-mot:dr_*) ⟩	⟨ (*sous-mot:_j*, *syntax:modifieur*) ⟩
5	⟨ (*sous-mot:_c_*) ⟩	⟨ (*sous-mot:_me*) ⟩

TABLE 6.14 – Top 5 des descripteurs d'apprentissage les plus importants selon les modules *Importance features* et *Permutation feature importance* (pour l'ensemble des descripteurs filtrés avec *ItemFreqGR* du couple de registres familier vs. courant).

cela, nous avons utilisé deux méthodes :

1. la première a regardé pour chaque descripteur d'apprentissage combien de fois ce descripteur a permis de séparer un arbre de décision : plus un descripteur était important, plus le nombre de fois où il a divisé un arbre était élevé ;
2. la seconde a retiré un à un les descripteurs d'apprentissage, pour mesurer son importance selon le taux d'erreur du classifieur : plus un descripteur était important, plus le taux d'erreur augmentait.

Pour ces deux méthodes, nous nous sommes servis de deux modules de Scikit-learn, respectivement, le module d'*importance features*[9] et le module de *permutation feature importance*[10].

La table 6.14 présente le top 5 des descripteurs d'apprentissage les plus importants obtenus avec les modules *Importance features* et *Permutation feature importance*, pour

9. Le module importance features est disponible sur scikit-learn : `http://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html`.
10. Le module permutation feature importance est disponible sur scikit-learn : `http://scikit-learn.org/stable/modules/permutation_importance.html`.

ID	***Importance features***
1	⟨ (*sous-mot:_#*, *pos:nom commun*), (*pos:ponctuation*) ⟩
2	⟨ (*sous-mot:_url_*), (*syntax:modifieur*) ⟩
3	⟨ (*pos:ponctuation*, *syntax:ponctuation*), (*sous-mot:_"_*, *lemme:"*, *syntax:ponctuation*) ⟩
4	⟨ (*lemme:.*, *sous-mot:_._*, *pos:ponctuation*, *syntax:ponctuation*), (*nombre:singulier*, *pos:déterminant*, *syntax:déterminant*) ⟩
5	⟨ (*lemme:de*), (*sous-mot:_#*, *syntax:préposition*) ⟩

TABLE 6.15 – Top 5 des descripteurs d'apprentissage les plus importants selon le module *Importance features* pour l'ensemble des descripteurs filtrés avec *ItemsetFreq* du couple de registres soutenu vs. familier.

les descripteurs filtrés avec *ItemFreqGR*, du couple de registres familier vs. courant. Elle montre que 4/5 descripteurs trouvés sont communs aux deux modules. Ces descripteurs sont courts avec un seul itemset par séquence. Tous les itemsets contiennent des traits linguistiques morphologiques avec des sous-mots précisant la forme des préfixes ou des suffixes des mots. Aussi, pour prédire si un texte est du registre familier ou courant, les débuts et fins de mots ont été des éléments importants. La table 6.15 présente le top 5, des descripteurs d'apprentissage les plus importants, pour prédire si un texte est du registre soutenu ou familier. Ces descripteurs ont été obtenus avec le module *Importance features*. Ils indiquent que les éléments de ponctuations et des éléments propres aux tweets, tels que les hashtags ou bien les URLs, ont été des traits linguistiques pertinents pour distinguer les registres soutenu et familier. Cela a à nouveau confirmé l'apport des *technomorphèmes* (PAVEAU 2013), pour l'analyse linguistique des registres de langue à partir d'un corpus de tweets.

Ainsi, les prédictions automatiques ont permis de valider la pertinence des motifs représentants. Toutefois, est-ce qu'un motif linguistique permettant de correctement prédire un registre, est un motif qui caractérise bien ce registre ?

6.3.2 Évaluation perceptuelle

L'évaluation automatique des motifs représentants, a permis de mettre au jour l'apport des motifs pour des prédictions de qualité. Afin de confirmer cette première évaluation, une seconde évaluation reposant sur le jugement humain a été proposée. Cette évaluation perceptuelle, a pour but de vérifier si les motifs représentants permettent effectivement de caractériser un registre de langue. Pour cela, nous avons demandé à un examinateur de sé-

lectionner, à partir de tweets, celui appartenant le plus au registre cible. Seulement un des deux tweets comportait un motif représentant caractéristique du registre de langue cible. Le tweet sélectionné n'avait pas vocation à être considéré comme familier, courant ou soutenu dans l'absolu, mais à l'être par rapport à l'autre tweet. La section suivante introduit le protocole expérimental dont les résultats, validant la qualité des motifs représentants, sont exposés en fin de chapitre.

6.3.2.1 Protocole expérimental

Tirages des motifs représentants et des tweets à évaluer Le protocole expérimental a visé à mettre en place une tâche d'évaluation où l'examinateur devait sélectionner entre deux tweets le tweet le plus familier, le plus courant, et le plus soutenu. Pour cela, trois couples de registres ont été considérés : le registre cible courant face au registre source familier ; le registre cible familier face au registre source soutenu ; et le registre cible soutenu face au registre source courant. Pour chacun de ces couples de registres :

- l'ensemble de motifs représentants, a été obtenu avec la mesure *ItemFreqGR* ;
- l'ensemble de motifs représentants, a été trié selon le taux de croissance des motifs dans l'ordre décroissant ;
- l'ensemble de motifs représentants, a été divisé en $N = 20$ sous-ensembles ;
- pour chacun des sous-ensembles, $X = 5$ motifs ont été tirés aléatoirement ;
- pour chacun des motifs, $Y = 10$ tweets représentants ont été tirés aléatoirement.

Dès lors, pour chaque couple de registres, nous avons obtenu 100 $(20 * 5)$ motifs représentants, qui ont été répartis dans 1 000 $(20 * 5 * 10)$ tweets représentants à évaluer. Comme le test considérait 3 couples de registres, au final, 3 000 $(3 * 1\,000)$ tweets ont été pris en compte pour l'évaluation perceptuelle. Afin de rendre la tâche d'évaluation raisonnable en termes de temps, chaque examinateur devait départager 30 paires de tweets (environ 20 minutes). L'outil (FlexEval [11]), utilisé pour mettre en place la plate-forme d'évaluation, a géré automatiquement le tirage des paires à faire évaluer, afin de couvrir la totalité des cas à évaluer.

Site Web mis en place pour la tâche d'évaluation FlexEval nous a permis de rendre accessible la tâche d'évaluation aux examinateurs en créant un site Web [12]. FlexE-

11. FlexEval est accessible via ce lien : `https://gitlab.inria.fr/expression/tools/FlexEval`.
12. Le site d'évaluation de nos résultats est accessible via ce lien : `http://expression.enssat.fr/test_tremolo/stage/page:visitor/debut/`

FIGURE 6.5 – Exemple de page test pour le registre cible Soutenu

val propose des briques préfabriquées aux utilisateurs pour l'aider à créer une plate-forme d'évaluation perceptuelle. À partir de templates de pages types (telles qu'une page de connexion, une page avec un formulaire pour le visiteur, ou une page avec le test à effectuer), l'utilisateur peut construire le parcours, qu'il souhaite pour son site web, et adapter les pages à ses besoins. Dans notre étude, nous avons structuré notre site en 5 pages permettant différents parcours. La page 1 est la page d'accueil ; la page 2 est la page où le visiteur rentre son adresse mail ; la page 3 est un formulaire qui s'affiche uniquement lors de la première connexion du visiteur ; la page 4 présente la tâche d'évaluation, et donne quelques informations sur les registres de langue ainsi que les tweets ; la page 5 affiche le test. Le visiteur pouvait se déconnecter à tout moment : lorsqu'il se reconnectait en rentrant son adresse mail, il se retrouvait à l'étape du test à laquelle il s'était déconnecté. Le formulaire (page 3) demande au visiteur son prénom et son nom, son âge, si sa langue maternelle est le français, s'il travaille dans le domaine de la linguistique ou bien du TAL, et enfin la fréquence (tous les jours / de temps en temps / jamais) à laquelle il lit et écrit des tweets. La figure 6.5 montre un exemple d'une page de test où il est demandé à l'examinateur de sélectionner le tweet lui semblant le plus soutenu des deux. Si aucun

des tweets proposés ne lui apparaissait comme soutenu, alors il pouvait spécifier qu'il n'arrivait pas à les départager pour passer au couple de tweets suivant.

6.3.2.2 Résultats expérimentaux

Pour recruter des examinateurs, nous avons diffusé le lien de notre plate-forme d'évaluation sur diverses listes de diffusion de la communauté scientifique (linguistique et TAL). Au total, 28 personnes ont effectué la tâche d'évaluation.

Profils des examinateurs Sur les 28 examinateurs, 21 examinateurs avaient pour langue maternelle le français ; 17 sur les 28 examinateurs ont précisé travailler dans le domaine de la linguistique ou bien du TAL. Parmi les 28 examinateurs, seulement 4 examinateurs nous ont dit être des lecteurs de tweets au quotidien, 12 nous ont informés qu'ils n'en lisaient que de temps en temps et les 12 autres jamais. Seulement 1 examinateur écrivait des tweets tous les jours, 5 de temps en temps et la majorité (22 examinateurs) n'en écrivait jamais. De manière globale, les examinateurs lisaient plus souvent des tweets qu'ils n'en rédigeaient.

Registre cible : familier Lorsque les examinateurs devaient sélectionner le tweet qui leur semblait le plus familier (registre cible A), ils ont sélectionné le tweet cible T_A dans 96% des cas. Dans seulement 3 cas, les examinateurs ont sélectionné le tweet T_B n'étant pas du registre cible. Pour ces 3 cas, le motif représentant contenu dans le tweet T_A avait un taux de croissance bas : 1.30, 1.09 et 1.42. Cela signifie donc que lorsque le motif représentant est faiblement caractéristique du registre cible, il ne permet pas clairement à le rendre perceptible. En cela, le taux de croissance est confirmé comme pertinent puisque sa faible valeur renvoie à la manifestation plus faible du registre cible dans le tweet.

Registre cible : courant Le taux de réussite a été plus bas lorsque les examinateurs ont du sélectionner le tweet qui leur semblait le plus courant : dans 68% des cas seulement ils ont sélectionné le tweet T_A effectivement du registre cible A (le registre courant). Lorsque l'examinateur a sélectionné le tweet T_B, dans la majorité des cas (c'est à dire dans 86% des cas), les tweets du registre courant comportaient des *MSE* avec des taux de croissance bas : entre 2 et 1.

Registre cible : soutenu La tâche d'évaluation a été mieux réussie pour le registre cible soutenu, que celui du courant, avec 80% des cas réussis. Dans 99% des cas, où l'examinateur a sélectionné le tweet T_B, le *MSE* du registre cible A avait un taux de croissance bas (c'est à dire entre 1 et 2). Par exemple, le motif suivant ⟨ (*lemme:que*), (*lemme:le, nombre:singulier*) ⟩ avait un taux de croissance de 2.53. Dès lors, le tweet qui illustrait ce motif (cf. exemple (130)), présentait un degré de caractérisation du registre cible soutenu plutôt faible. Cela peut expliquer pourquoi l'examinateur ne l'a pas sélectionné.

(130) Manuel Valls se confie dans le #JDD ... le choix de "Laurent Blanc" est une erreur de casting pour le #FCBarcelone - "De manière immodeste, je pense que le #Barca a besoin de moi" ! url_path

6.4 Conclusion

Dans ce chapitre nous avons présenté les travaux conduits pour obtenir un sous-ensemble de MSE. Pour cela, nous avons procédé en deux temps. Le premier réduit la redondance des MSE grâce à des approches de regroupement automatique de données ; le second sélectionne pour chaque groupe un motif représentant. L'ensemble des motifs représentants constitue un sous-ensemble de MSE donc moins redondant. Les expériences détaillées ont montré d'un côté l'efficacité de notre algorithme de regroupement, d'un autre côté sa capacité à ne pas perdre des MSE intéressants peu fréquents. Ensuite, lors des sélections de motifs représentants à partir des sept méthodes proposées nous n'avons pas posé d'*a priori* sur un meilleur type de motif représentant. C'est pourquoi, nous avons des résultats très similaires à partir desquels un expert humain ne pourrait trancher. Une manière de répondre à ce problème est d'évaluer automatiquement les motifs représentants.

Deux évaluations ont été présentées : une évaluation automatique et une évaluation perceptuelle. Le but de l'évaluation automatique était double. Tout d'abord, vérifier la qualité des sous-ensembles de motifs représentants. Puis, découvrir si une des sept manières de sélectionner ces représentants était meilleure. Le protocole mis en place, pour évaluer la qualité du sous-ensemble de motifs séquentiels, a reposé sur l'hypothèse qu'un bon motif représentant ferait un descripteur d'apprentissage pertinent. Ces descripteurs devaient permettre au classifieur de faire des prédictions de qualité. Aussi, pour vérifier la qualité des motifs représentants, nous avons implémenté des classifieurs binaires dont les tâches étaient de prédire si un texte était du registre A, ou du registre B. Ces classifieurs

ont appris à distinguer les deux registres en se basant sur des descripteurs d'apprentissage composés des motifs représentants. Les résultats ont validé la qualité des motifs représentants, dès lors qu'ils ont permis des prédictions de qualité. Bien que les différences n'ont pas été pas significatives, entre les sept méthodes de sélection, nous avons pu noter des prédictions légèrement meilleures lorsque les motifs avaient été sélectionnés avec la méthode notée *ItemFreqGR*.

La seconde évaluation a validé la pertinence des motifs représentants, en terme de caractérisation des registres de langue. En effet, l'évaluation manuelle des motifs a mis en lumière le fait qu'un tweet T_A, qui contenait un motif représentant du registre cible A, est majoritairement sélectionné comme étant de ce registre. De plus, le taux de croissance a été confirmé comme pertinent pour notre tâche puisque dans la grande majorité des cas où l'examinateur a sélectionné le tweet T_B, qui n'était pas du registre cible A, les taux de croissance des motifs étaient très bas. Autrement dit, plus le taux de croissance d'un motif est haut, plus il contribue à la caractérisation du registre cible ; et inversement, plus le taux de croissance d'un motif est bas, moins il contribue à la caractérisation du registre cible.

Ainsi, ce chapitre a exposé les expériences validant la qualité des sous-ensembles de motifs représentants. En cela, il valide également la totalité de notre chaîne de traitement pour caractériser les registres de langue dont la principale contrainte était de ne pas poser d'*a priori* linguistique sur les motifs à obtenir. Une dernière validation testant sa robustesse face à un autre phénomène linguistique est présentée dans le chapitre suivant.

Chapitre 7

Généralisation de notre chaîne de traitement

Sommaire

7.1 Introduction

Ce chapitre présente les travaux testant notre chaîne de traitement pour un autre cas d'usage. Ils sont motivés par deux principales raisons : tester sa robustesse pour caractériser un autre phénomène linguistique que les registres de langue ; vérifier qu'elle lèverait également les verrous algorithmiques de la fouille de MSE et de leur partitionnement en traitant un autre jeu de données. Avec ces expériences, nous avons obtenu un ensemble de motifs exploitable et interprétable, c'est-à-dire d'une taille raisonnable et sans redondance. Ces résultats ont validé :

1. notre méthodologie de fouille de MSE, en vérifiant son passage à l'échelle face à un autre jeu de données réelles ;
2. la robustesse de l'algorithme RGMSE , en l'appliquant à un second ensemble de MSE à regrouper.

Nos expériences montrent la réussite de notre démarche ne contraignant pas la découverte des MSE, en posant des *a priori* linguistiques sur ce que devraient être des MSE intéres-

sants, car généralisable à des cas d'usage analysant d'autres phénomènes linguistiques à partir de données contrastées.

Nous avons choisi de tester notre approche pour caractériser les genres de textes adressés aux enfants. Par exemple, caractériser le genre encyclopédique par rapport au genre journalistique avec un ensemble de MSE. Ce phénomène linguistique est choisi car :

1. notre absence de connaissance sur le sujet limite le risque de biaiser l'extraction de MSE par des attentes linguistiques ;
2. un corpus de textes déjà constitué et segmenté proprement en phrases est disponible : le corpus TextToKids.

Dès lors, notre chaîne de traitement a été appliquée au corpus TextToKids pour découvrir des MSE caractéristiques d'un genre de textes cible G_c, par rapport à un genre de textes source G_s. Pour rendre les résultats obtenus à partir du corpus TextToKids comparables à ceux obtenus à partir du corpus TREMoLo-Tweets, nous avons conservé les mêmes paramètres expérimentaux sans changer leurs valeurs fixées pour les expériences à partir du corpus TREMoLo-Tweets. Les résultats ont validé notre approche en constituant un ensemble de MSE pertinents pour caractériser les genres de textes enfantins.

Nous débutons ce chapitre en présentant le corpus TextToKids, à partir duquel nous avons cherché l'ensemble complet de MSE dont la fouille est introduite dans la deuxième section. Puis, nous montrons comment nous avons sélectionné un sous-ensemble de MSE que nous avons évalués. Les résultats de cette évaluation sont exposés dans la dernière section.

7.2 Corpus TextToKids

Contexte Le projet TextToKids est un projet ANR [1] dont le but est de « faciliter l'écriture et le filtrage de textes pour les enfants, notamment mais pas seulement pour leur parler de l'actualité (ex : élections présidentielles, Brexit, accueil des migrants en France, etc.) dans le respect de leurs compétences linguistiques. La tranche d'âge visée est celle des jeunes lecteurs, c'est-à-dire les 7-12 ans. » [2]. Le projet regroupe divers domaines scientifiques avec des linguistes et des informations travaillant avec des journalistes spécialisés. Une de ses finalités est de « proposer des outils d'assistance (analyse textuelle automatisée, moteur de recherche, reformulation, bonnes pratiques) » (Ibid.).

1. ANR-19-CE38-0014
2. `https://texttokids.irisa.fr/`

ID	Genre	Phrase
1	Encyclopédique	L'histoire du Portugal est partagée avec celle de la péninsule Ibérique.
2		La région a été peuplée par des pré celtes et des celtes, elle est à l'origine de peuples comme les Gallaeci, les Lusitaniens et les Cinii.
3		Elle a été visitée par les Phéniciens et les Carthaginois.
4	Journalistique	Le Premier ministre travaille dans un lieu qui s'appelle Matignon.
5		Chaque ministre travaille dans un domaine différent (ça va de la sécurité à la culture) avec sa propre équipe.
6		Par exemple, le ministre de l' Education nationale peut changer ce qu'on apprend à l'école.
7	Fiction	J'ai senti que c'était très important.
8		Qu'il fallait que je comprenne très vite.
9		Je devais absolument dire quelque chose à Frrintek, mais pour ça, il fallait que j'analyse la situation.

TABLE 7.1 – Trois exemples de phrases sont donnés pour les trois genres de textes.

Corpus Le corpus est constitué de trois genres de textes : des textes encyclopédiques, journalistiques et des fictions. Au total il compte 1 499 textes pour 157 720 phrases. Tous les textes ont été collectés et répartis manuellement dans différentes tranches d'âge. Dans notre cas, nous avons choisi de considérer une seule tranche d'âge, celle des [7-13] ans, pour les trois genres de textes afin d'éviter toute variation causée par des contrastes dus aux différentes tranches d'âge. Au final, le sous-corpus de TextToKids que nous avons utilisé est composé de 17 186 phrases pour le genre encyclopédique, 32 880 phrases pour le genre journalistique et 19 360 phrases pour le genre fiction. Pour donner une idée de ce que contient ce sous-corpus de TextToKids, nous donnons avec la table 7.1 trois exemples de phrases pour chaque genre de textes. À partir de ce sous-corpus divisé en trois sous-corpus illustrant les genres encyclopédique, journalistique et fiction, nous avons cherché des descripteurs linguistiques distinguant ces genres de textes entre eux. Pour cela, nous avons ré-utilisé le protocole expérimentale cherchant des MSE caractéristiques d'un registre de langue cible face à un registre de langue source.

ID	G_c	G_s
1	Encyclopédique	Journalistique
2		Fiction
3	Journalistique	Encyclopédique
4		Fiction
5	Fiction	Encyclopédique
6		Journalistique

TABLE 7.2 – Les six paires de genres considérées pour la fouille de MSE.

7.3 Caractérisation des genres de textes à partir du corpus TextToKids

Transformation des phrases en séquences Une transformation des données textuelles en données séquentielles est nécessaire pour exécuter une fouille de MSE. Dans ces travaux, nous avons choisi la phrase comme segment textuel représenté par une séquence. Nous avons fait ce choix car les textes entiers sont trop longs pour être représentés par une séquence, nous aurions obtenu des fréquences de MSE donnant une vision imprécise de leur présence dans les sous-corpus. En outre, le corpus TextToKids est déjà segmenté en phrases propres. L'unité textuelle représentée par un itemset est le mot. Ce dernier est décrit par un ensemble de traits linguistiques portés par des items. Nous avons repris les mêmes traits que pour la fouille de MSE à partir de TREMoLo-Tweets, c'est à dire les lemmes, les parties grammaticales, les caractéristiques morphologiques, les fonctions syntaxiques et les sous-mots.

Découverte de l'ensemble complet de MSE À partir de ces données séquentielles, la fouille de MSE a pour objectif de découvrir des motifs distinguant G_c de G_s en comparant leurs fréquences dans les textes de G_c et de G_s via leurs taux de croissance TC. Nous rappelons que les MSE sont des motifs dont TC dépasse le seuil ρ fixé par l'utilisateur. Au total, six paires de genres ont été considérées lors des fouilles de MSE. Comme expliqué lors de l'introduction, nous n'avons pas modifié les paramètres fixés dans les expériences de fouilles de MSE précédentes à partir du corpus TREMoLo-Tweets. Nous rappelons également que les paramètres du protocole de fouille de MSE ont été fixés pour réduire le risque de passer à côté de MSE intéressants. C'est pourquoi ils ne contraignent pas fortement la fouille avec des seuils élevés :

— le seuil $minsup_c$ et $minsup_s$ ont tous les deux été fixés à 1% ;

ID	Genre de textes	motifs fréquents		motifs clos	
		Nbr de MS	**Taille du fichier de résultats**	**Nbr de MS**	**Taille du fichier de résultats**
1	Encyclopédique	30 455 400	1 GO	1 967 167	74 MO
2	Journalistique	389 214 968	25 GO	1 361 571	49 MO
3	Fiction	10 877 210	424 MO	999 076	37 MO

TABLE 7.3 – Nombre de motifs fréquents et clos découverts pour les trois genres de textes avec le poids des fichiers les stockant.

— la contrainte de gap a été fixée à $P[1, 1]$ pour avoir des MS dont les itemsets sont contigus ;

— le seuil sur le taux de croissance ρ a été fixé à 1 afin d'avoir tous les MSE possibles, même ceux ayant un faible taux de croissance.

La principale validation de cette étape est le fait que la fouille ait abouti, cela veut dire que la complexité algorithmique n'a pas empêché la fouille de MSE.

Résultats de la fouille de MSE La fouille de MSE résulte de deux étapes : la première est la fouille de motifs clos et fréquents pour tous les genres de textes ; la seconde est la sélection des MSE à partir des motifs clos et fréquents. La table 7.3 donne le nombre de motifs fréquents et clos découverts pour les trois genres de textes. Chaque ligne correspond à un genre, les résultats de la fouille de motifs fréquents sont présentés avant ceux de la fouille de motifs clos. Pour chaque ensemble de MS découvert, le poids du fichier les stockant est précisé. Le genre de texte journalistique est celui à partir duquel le plus grand nombre de motifs fréquents ont été découverts. En revanche, l'ensemble de motifs clos est 286 fois moins volumineux que celui des motifs fréquents. Cette différence de quantité de motifs indique qu'il y a une forte redondance dans les données séquentielles représentant le genre journalistique, puisque que la majorité des motifs fréquents est incluse dans une sur-séquence ayant la même fréquence. Peut être que le genre textuel journalistique emploie plus de routines linguistiques, notamment parce que les publications, quotidiennes ou hebdomadaires, sont itérées très régulièrement. La seconde fouille de motifs est celle des MSE à partir des motifs clos de G_c et des motifs fréquents de G_s. La table 7.4 précise pour chaque paire de genres considérée le nombre de MSE découverts. Par rapport aux résultats obtenus en caractérisant les registres de langue, nous avons avec les genres de textes des ensembles plus volumineux avec une moyenne de 1 045 553 de MSE contre une

ID	G_c	G_s	**Nbr de MSE**
1	Encyclopédique	Journalistique	1 722 859
2		Fiction	1 823 956
3	Journalistique	Encyclopédique	584 081
4		Fiction	1 117 584
5	Fiction	Encyclopédique	492 766
6		Journalistique	538 070

TABLE 7.4 – Nombre de MSE découverts pour chaque paire de genres de textes.

moyenne de 599 585 MSE pour les registres. Ces résultats indiquent qu'il y aurait plus de formes linguistiques spécifiques à certains genres, qu'il n'y aurait de formes linguistiques spécifiques à certains registres de langue, puisque la fouille de MSE caractérisant les genres de texte découvre plus de motifs avec des fréquences supérieures dans G_c à celles dans G_s, que la fouille de MS caractérisant les registres de langue. La section suivante montre comment nous avons réduit les ensembles de MSE caractérisant les genres de textes à des sous-ensembles moins importants.

7.4 Constitution d'un sous-ensemble de résultats

Un sous-ensemble de MSE doit être moins redondant et moins volumineux pour qu'il soit jugé plus exploitable que l'ensemble complet de MSE. Pour cela, nous reprenons les deux étapes précédemment utilisées dans le chapitre 6 : le regroupement des MSE selon leur similarité et le tirage d'un motif par groupe pour constituer un ensemble de motifs représentants. Cette approche est validée par les résultats obtenus lors de l'évaluation des ensembles de motifs représentants. Ces résultats sont présentés à la fin de cette section.

Réduction de la redondance des MSE Toujours dans l'optique de pouvoir comparer nos résultats, nous avons utilisé exactement le même outil automatique : l'algorithme RGMSE dont les paramètres sont les mêmes que pour la caractérisation des registres. Le premier résultat est la capacité de RGMSE à regrouper de larges ensembles d'objets puisque les ensembles de MSE sont plus importants que ceux issus de TREMoLo-Tweets. La table détaille, pour les six couples de genres de textes, le nombre de groupes de MSE lors des trois partitions créées durant l'exécution de RGMSE. La même tendance que celle constatée lors des regroupements des MSE issus de TREMoLo-Tweet est présente : le nombre de groupes baisse légèrement à chaque nouvelle partition indiquant une cohérence

ID	Gc	Gs	**Partition 1**	**Partition 2**	**Partition 3**
1	Encyclopédique	Journalistique	2 524	2 512	2 464
2		Fiction	2 880	2 868	2 813
3	Journalistique	Encyclopédique	2 401	2 394	2 370
4		Fiction	2 974	2 961	2 914
5	Fiction	Encyclopédique	1 868	1 862	1 838
6		Journalistique	1 717	1 713	1 698

TABLE 7.5 – Nombre de groupes de MSE pour les trois partitions obtenues lors de l'exécution de RGMSE et pour les six paires de genres de textes.

dans la redistribution des MSE lors des itérations de RGMSE. Ces redistributions des MSE semblent bien répondre aux limites de la première itération tirant sans remise les MSE rendant les groupes de MSE très dépendants de l'ordre dans lequel les MSE sont tirés. En effet, lorsque nous regardons la taille du plus grand groupe de MSE, nous observons une nette diminution de sa taille entre la première et la seconde itération. Par exemple, pour le couple de genres fiction et encyclopédique : il contient 35 543 MSE lors de la première itération et 11 769 MSE lors de la deuxième. De même, pour le couple de genres journalistique et fiction : il contient 39 940 MSE puis 9 707 lors de, respectivement, la première et la deuxième itération.

Réduction de la quantité des MSE À partir des partitions obtenues, nous avons sélectionné pour chaque groupe un motif représentant. Les sept méthodes de sélection ont été utilisées. En un premier lieu nous montrons que les motifs représentants sont de bonne qualité, dans un second lieu qu'il est toujours difficile pour un expert de départager quelle est la meilleure méthode de sélection de ces motifs représentants.

Tout d'abord, pour montrer la qualité des représentants nous donnons dans la table 7.6 20 exemples de motifs représentants sélectionnés avec *ItemFreqGR*. Ces exemples caractérisent le genre fiction par rapport aux genres encyclopédique et journalistique, puis les genres encyclopédiques et journalistiques par rapport au genre fiction. Ces quelques exemples montrent déjà la qualité des motifs. Les motifs 1 à 10 marquent la dimension narrative de la fiction par rapport aux genres encyclopédique et journalistique. Par exemple, le motif 1 renvoie à des structures syntaxiques souvent répétées dans les phrases qui contribuent à la structuration temporelle du récit. Ci-dessous des exemples [3] de deux phrases se suivant dans le texte dont elles sont issues et contenant toutes les deux le

3. Tous les exemples sont issus du corpus TextToKids.

ID	Motif représentant
Fiction vs. Encyclopédique	
1	{sous-mot :_quand, lemme :quand} {nombre=singulier}
2	{sous-mot :_puis}
3	{lemme :aimer}
4	{lemme :–}, {personne :première}
5	{sous-mot :_monsieur}
Fiction vs. Journalistique	
6	{lemme :croire}
7	{lemme :entendre}
8	{syntaxe :modifieur, sous-mot :_si}{nombre :singulier}
9	{syntaxe :modifieur} {sous-mot :_lui}
10	{lemme :ton, sous-mot :_ton}
Encyclopédique vs. Fiction	
11	{lemme :français}
12	{lemme :américain}
13	{sous-mot :_On}{pos :verbe, nombre :singulier, syntaxe :racine}
14	{sous-mot :_Cette}
15	{pos :participe présent}
Journalistique vs. Fiction	
16	{sous-mot :_travail}
17	{sous-mot :_lieu}
18	{sous-mot :_explique}
19	{lemme :politique}
20	{lemme :pourquoi, pos :ADVB de questionnement, sous-mot :_Pourquoi}

TABLE 7.6 – Cinq exemples de motifs représentants sélectionnés avec *ItemFreqGR* pour quatre couples de genres.

motif 1 :

(131) Il parlait encore de tante Briqueboeuf, il me racontait ce qu'il ferait avec elle quand elle serait là, ce qu'il lui dirait, ce qu' elle lui répondrait .

(132) Mais il ne me demandait plus jamais quand elle viendrait.

Peut être est-ce une manière de marquer l'introduction de différents temps verbaux pour apprendre aux enfants quand les utiliser. Le motif 4, illustré par les exemples (133) et (134), montre la présence de dialogues avec un locuteur qui parle à la première personne soit du singulier soit du pluriel.

(133) – Nous on mange les pâtes tous les soirs, dit Emma en regardant la soupe de potiron qu'elle ne pourrait pas avaler même si on lui payait dix mille

euros !

(134) – Je te dirai demain, je ne sais pas si je verrai mes parents cette semaine et je ne sais pas si Sean sera là samedi soir.

Le motif 6 illustre également la présence du locuteur dans le texte avec le verbe *croire*. Face au registre journalistique se voulant neutre, la présence d'un verbe modalisateur montre que la fiction met plus en avant la présence du locuteur comme l'illustrent les exemples (135) et (136).

(135) Je crois que je vais arrêter là.

(136) - Je crois bien, oui, je crois bien.

Les motifs 11 à 15 distinguent le genre encyclopédique avec des formes montrant les dimensions historique et géographique de certains de ces contenus. Prenons par exemple les motifs 11 et 12 illustrés par les extraits ci-dessous :

(137) Les faubourgs de Lille, quartiers voisins, sont attaqués par l'armée française en 1645.

(138) Ancienne colonie française, sa devise (issue de cette période de son histoire) est Florebo quocumque ferar, qui est traduisible par "je fleurirai partout où je serai plantée/portée".

(139) Plusieurs espèces d'eau douce, comme l'écrevisse américaine, le poisson-chat commun, ou la perche soleil, originaires d'Amériques du Nord, ont été introduits en Europe, où ils étaient élevés dans des bassins.

(140) La morue se trouve aussi de l'autre côté de l'Atlantique, sur les côtes américaines.

Enfin, les motifs 16 à 20 caractérisent le genre journalistique par rapport au genre fiction. Les exemples (141) et (142) montrent que le motif 18 est un verbe de parole fréquemment utilisé pour rapporter le discours d'une personnalité.

(141) "on ne peut soigner que le problème pour lequel le patient a été hospitalisé, sans regarder s'il y en a d' autres à traiter, explique le docteur Renaud Péquignot. résultat : les patients reviennent à l' hôpital ."

(142) "c'est pour nous faire peur ou nous mettre en colère, comme ça on croit plus facilement à leurs théories", explique t il.

Comme illustré par les exemples (143) à (146), le motif 20 quant à lui permet de poser une question rhétorique aux journalistiques pour introduire leurs propos.

(143) Pourquoi les ministres veulent ils tous faire des réformes ?

(144) Pourquoi est il mauvais pour les humains et l' environnement ?

(145) Pourquoi des automobilistes sont ils en colère ?

(146) Pourquoi ce qui se passe en Algérie touche autant en France ?

Ces quelques exemples indiquent que les motifs représentants sont de bonne qualité et ont réussi à distinguer les genres entre eux avec descripteurs linguistiques spécifiques.

Le second point était de trancher sur une meilleure méthode de sélection de ces motifs représentants. Cependant, de même que lors des expériences caractérisant les registres de langue, les motifs représentants obtenus avec des méthodes différentes sont très similaires. Ci-dessous des exemples de motifs représentants pour différentes métriques pour un même groupe de MSE :

- *Médoïde* : ⟨{lemme :certain}, {pos :nom commun}⟩ ;
- *ItemsetFreq* : ⟨{lemme :certain}⟩ ;
- *ItemFreq* : ⟨{lemme :certain, nombre :pluriel, pos : déterminant, syntaxe :déterminant}, {pos :nom commun}⟩ ;
- *ItemsetFreqGR* ⟨{lemme :certain}⟩ ;
- *ItemFreqGR* : ⟨{lemme :certain}⟩ ;
- *ItemsetFreqCov* : ⟨{lemme :certain}⟩ ;
- *ItemFreqCov* : ⟨{lemme :certain}⟩.

Pour les départager, nous avons utilisé l'évaluation automatique afin de comparer quel type de motif représentant conduisait à de meilleures prédictions.

Évaluation automatique des résultats Nous n'avons pas fait d'évaluation perceptuelle pour cette tâche de caractérisation des genres de textes. En effet, nous pensons qu'une phrase décontextualisée, car tirée d'un texte, est insuffisante pour qu'un évaluateur perçoive un genre de texte. Or, l'unité textuelle considérée dans ces expériences est la phrase. L'évaluation automatique reprend le protocole présenté en section 6.3 qui utilise l'algorithme Random Forest comme classifieur binaire. Trois expériences de classification

Couple de genres de textes	**F-mesure**						
	Med(C)	**Itemset Freq**	**Item Freq**	**Itemset FreqGR**	**Item FreqGR**	**Itemset FreqCov**	**Item FreqCov**
journalistique vs. fiction	0.88	<u>0.91</u>	0.90	<u>0.91</u>	<u>0.91</u>	0.90	0.90
journalistique vs. encyclopédique	0.86	0.85	0.87	0.86	<u>0.88</u>	0.85	0.86
encyclopédique vs. fiction	0.86	0.88	0.86	0.88	<u>0.89</u>	0.89	0.88

TABLE 7.7 – Les scores de la f-mesure pour les sept méthodes de sélection de motifs représentants pour les trois couples de genres de textes. Les meilleurs scores sont soulignés.

ont été conduites : les genres journalistique vs. fiction ; les genres journalistique vs. encyclopédique et les genres encyclopédique vs. fiction. La table 7.7 précise pour chaque paire de genres de textes les f-mesures obtenues selon la méthode de sélection d'un motif représentant. Chaque ligne représente un couple de genres et chaque colonne une méthode de sélection. Globalement, tous les résultats sont bons puisqu'ils varient de 0.85 pour le pire score jusqu'à 0.91 pour les meilleurs scores. Ces résultats montrent que les sous-ensembles de motifs représentants sont pertinents pour distinguer les genres de textes entre eux. En outre, bien que tous les scores aient des valeurs proches, les motifs représentants sélectionnés selon la méthode ItemFreqGR ont systématiquement conduit aux meilleurs résultats. Cela indique que les motifs partageant le plus d'items communs avec les autres motifs de son groupe et étant fortement caractéristique du genre cible sont les meilleurs motifs représentants.

7.5 Conclusion

Le principal objectif de ce dernier chapitre était de montrer la possibilité d'utiliser notre chaîne de traitement pour caractériser un autre phénomène linguistique. Nous avons montré qu'une approche ne contraignant pas fortement la fouille de MSE pour répondre aux verrous algorithmiques, liés à la recherche des motifs fréquents, peut être appliquée à un large ensemble de données. Les résultats obtenus ont validé cette approche en montrant sa capacité à être généralisée. Premièrement, les deux étapes avec une complexité algorithmique exponentielle ont réussi à traiter des jeux de données plus volumineux que

ceux issus de TREMoLo-Tweets : cela indique que nos stratégies visant à la limiter, lors de la fouille des MSE et de leur regroupement, sont robustes. Deuxièmement, notre algorithme RGMSE a paru être logique dans sa manière de regrouper les MSE similaires entre eux avec des partitions cohérentes durant ses itérations. Troisièmement, l'évaluation automatique des sous-ensembles de motifs représentants a mis au jour de meilleures performances lorsque les motifs représentants étaient choisis en regardant les fréquences des items du motif pondérées par son taux de croissance. Ainsi, notre méthodologie visant à caractériser un phénomène linguistique à partir de données contrastées est validée une seconde fois avec un cas d'usage différent du notre.

Chapitre 8

Conclusion

Nous concluons ce manuscrit en présentant un bilan des travaux conduits dans le cadre de cette thèse mettant en avant nos contributions. Il est suivi d'une dernière section proposant des pistes de futurs travaux.

8.1 Bilan de la thèse

L'objectif principal de cette thèse est de caractériser les registres de langue par extraction de **M**otifs **S**équentiels **É**mergents (MSE). L'objectif secondaire est de proposer une chaîne de traitements utilisable pour d'autres cas d'usage. Pour répondre au premier objectif, nous avons proposé une chaîne de traitements dans le but d'obtenir un ensemble de MSE distinguant les registres de langue entre eux. Durant l'élaboration de cette chaîne de traitement, nous avons veillé à ne pas poser d'*a priori* linguistiques afin de pouvoir l'appliquer à d'autres phénomènes linguistiques représentés avec des données contrastées. Les résultats obtenus montrent que nous sommes parvenus à remplir le but initial de caractériser les registres de langue grâce à une méthodologie robuste. Ils montrent également que notre second but est atteint avec une chaîne de traitement pouvant être appliquée à d'autres cas d'usages tels que la caractérisation des genres de textes adressés aux enfants.

Notre travail résulte en quatre contributions :

1. un corpus de tweets en français étiquetés automatiquement en registres de langue avec un sous-corpus annoté manuellement et un guide d'annotation en registres de langue pour le français ;
2. une chaîne de traitement comprenant un test de la robustesse de la fouille de MSE pour caractériser les registres de langue ;
3. une méthodologie de fouille de MSE originale contraignant peu la recherche de motifs pour réduire le risque de manquer des motifs intéressants ;

4. un processus de réduction des résultats avec le groupement des MSE entre eux selon leur similarité et la sélection d'un motif représentant par groupe.

Nous détaillons dans cette section ces contributions en commençant par le corpus.

Corpus TREMoLo-Tweet Après avoir délimité notre objet d'étude, nous l'avons illustré avec un corpus de tweets. Ce corpus, TREMoLo-Tweets, rassemble 228 505 tweets étiquetés en registres familier, courant et soutenu. Pour étiqueter l'intégralité de TREMoLo-Tweets, un sous-ensemble de tweets a été annoté manuellement en registres de langue. L'annotation manuelle a été encadrée par un guide d'annotation issu d'une étude linguistique du corpus. Elle a ensuite été généralisée à l'intégralité du corpus de manière itérative : l'ensemble de données d'entraînement a été augmenté à chaque itération. Les résultats obtenus ont montré la qualité des prédictions et la pertinence des descripteurs linguistiques du guide d'annotation.

Test de la robustesse de la fouille de MSE Pour êtres certains de pouvoir nous fier aux MSE, nous avons testé la robustesse des techniques de fouille de MSE à partir d'un ensemble de données artificielles pour connaître en amont les motifs à retrouver. Les extractions de MSE ont validé les techniques de fouille de MSE car elles ont retrouvé les MSE attendus. Elles ont également révélé une complexité algorithmique importante pour l'extraction des motifs clos, et un nombre élevé de motifs fréquents extraits.

Protocole de fouille de MSE Plusieurs stratégies, jouant sur la représentation des données textuelles en données séquentielles, ont été mises en place pour ne pas faire exploser le coût algorithmique lors de l'extraction des MSE à partir de données réelles. Chaque tweet a été représenté par une séquence : la limite de taille imposée par Twitter a garanti une longueur maximale de 280 caractères. Chaque mot des tweets a été décrit par un itemset comprenant un nombre restreint de traits linguistiques : le lemme, les sous-mots, les caractéristiques morphologiques, la partie grammaticale du discours, la fonction syntaxique. Chacun de ces traits a été symbolisé par un item de l'itemset. Pour limiter le nombre de MSE extraits, les motifs clos ont été utilisés et seuls les motifs contiguës ont été fouillés.

Réduction et évaluation des résultats Afin de réduire le nombre de MSE à un sous-ensemble de taille et de redondance raisonnables, c'est-à-dire interprétable par un expert,

les MSE ont été partitionnés selon leurs similarités. L'algorithme RGMSE a été introduit pour regrouper les MSE sans avoir à fixer *a priori* le nombre de groupes à obtenir. Il permet de traiter un grand nombre de motifs à regrouper grâce à une première étape de tirage sans remise des motifs. Une évaluation automatique a assuré la qualité des partitions de MSE, et une exploration des groupes a validé leurs pertinences. Pour chacun des groupes de motifs, un motif représentant est sélectionné. Sept méthodes de sélection de motif représentant ont été proposées dans le but de répondre à divers besoins utilisateurs. Deux évaluations ont vérifié leur capacité à caractériser R_c. La première s'est basée sur la perception humaine des registres de langue, et a regardé un nombre limité de motif représentant. L'examinateur devait sélectionner le tweet qu'il percevait comme appartenant à R_c. La tâche d'évaluation a supposé qu'un tweet contenant un motif représentant caractéristique de R_c serait perceptible comme appartenant à R_c. 28 examinateurs ont complété la tâche d'évaluation. Les résultats ont donné une première validation de la capacité des motifs représentants à caractériser R_c. Une seconde évaluation a été proposée pour vérifier que la totalité des motifs représentants permet bien de distinguer un registre A d'un registre B. Elle s'est servie d'un classifieur binaire pour prédire si le tweet considéré était du registre A ou du registre B. Pour entraîner ce classifieur à distinguer les deux registres, les motifs représentants sont utilisés comme descripteurs d'apprentissage. Les résultats obtenus ont de nouveau prouvé la capacité des motifs représentants à bien distinguer les registres A et B. Ils ont également révélé que les motifs représentants sélectionnés avec la méthode $ItemFreqGR$ conduisaient à des résultats légèrement meilleurs qu'avec les six autres méthodes.

8.2 Perspectives

Plusieurs perspectives peuvent être considérées pour approfondir les travaux présentés dans ce manuscrit. Cette section dresse, pour chaque étape de notre chaîne de traitement, les pistes de travaux complémentaires possibles. Nous concluons en présentant des possibles utilisations de notre approche pour répondre à divers questions de recherche en linguistique de corpus ou cas d'usage en TAL.

8.2.1 Pistes d'approfondissements de notre travail

Tout d'abord, lors de l'annotation manuelle d'une partie du corpus TREMoLo-Tweets les annotateurs devaient justifier leur choix de registres en sélectionnant au moins un descripteur linguistique présent dans le tweet. Un ensemble de descripteurs linguistiques a donc été annoté en proportions de registres de langue. De futurs travaux pourraient explorer cet ensemble pour découvrir si des descripteurs ont été systématiquement sélectionnés ensemble. Cela permettrait de mettre en lumière des ensembles de descripteurs linguistiques caractéristiques d'un registre de langue. Afin de valider leur pertinence, nous pourrions utiliser ces descripteurs linguistiques pour entraîner un classifieur à prédire le registre de langue d'un tweet. Des prédictions de qualité confirmeraient la pertinence des descripteurs linguistiques pour distinguer les registres de langue entre eux. Par ailleurs, l'annotation manuelle n'a pas été validée par un accord inter-annotateurs. De fait, nous n'avons pas trouvé d'accord inter-annotateurs pertinent pour l'annotation en proportions de registres de langue utilisée. La proposition d'un accord inter-annotateurs adapté à ce type d'annotation manuelle constituerait une contribution notable. Enfin, les mesures d'évaluation (précision, rappel et f-mesure) indiquent la performance du classifieur au regard de l'annotation manuelle. L'enjeu est de savoir si les annotations sont pertinentes d'un point de vue linguistique. Étant donné la taille du corpus TREMoLo-Tweet, nous ne pouvons pas explorer manuellement l'ensemble du corpus. Nous pourrions toutefois tirer aléatoirement 500 textes à annoter manuellement pour comparer les annotations manuelles et automatiques.

Ensuite, nous avons vu que l'extraction de MSE est limitée par son coût algorithmique. L'utilisation de techniques d'échantillonage des séquences à fouiller pourrait être une solution. De futurs travaux pourraient implémenter des approches proposées dans la littérature comme celle introduite par (Raissi et Poncelet 2008). En outre, dans notre chaîne de traitement les motifs séquentiels clos ont été utilisés. Un défaut de ce type de motif est qu'un motif clos $M_1 = \langle (a, b, c), (b, c) \rangle$ avec un support de 6 ne pourra pas contenir $M_2 = \langle (a, b, c), (b) \rangle$ dont le support est de 5. Pour réunir M_1 et M_2, nous pourrions utiliser les *δ-motifs* introduits par (Holat, Plantevit et al. 2014). Les *δ-motifs* rapprochent les motifs avec des supports voisins. Dès lors, leur utilisation permettrait de ne pas différencier des motifs proches, et amènerait peut être à un ensemble de MSE moins grand que celui obtenu avec les motifs clos. D'autres types de traits linguistiques pour décrire un mot peuvent également être envisagés. Par exemple, le module Spacy [1] annote

1. L'outil est disponible en ligne : `https://spacy.io/`.

si l'unité textuelle contient des majuscules et des minuscules, se compose de chiffres ou de lettres, est une entité nommée. Tous ces traits pourraient être intégrés aux itemsets pour découvrir s'ils caractérisent un registre de langue. Enfin, pour faire émerger des MSE caractéristiques d'un registre cible, d'autres mesures que celle du taux de croissance pourraient être étudiées.

L'étape de regroupement des MSE selon leur similarité ouvre également des pistes d'amélioration. Parmi elles, citons la possibilité de pondérer différemment la mesure S^2MP. Deux expériences pourraient être menées : donner plus d'importance au contenu commun entre deux séquences, ou bien donner plus d'importance à l'ordre des itemsets des deux séquences. Faire varier ces pondérations conduirait à des partitions de MSE différentes. Afin d'obtenir de nouvelles partitions, nous pourrions aussi laisser l'algorithme RGMSE itérer autant de fois nécessaires à l'obtention d'une partition où plus aucun MSE ne change de groupe. Chercher à obtenir des partitions différentes nous permettrait peut être de découvrir des MSE caractéristiques des registres de langue différents.

Enfin, plusieurs manières de sélectionner des motifs représentants pourraient être testées. Par exemple, nous pourrions pondérer les types de traits contenus dans les MSE. Nos travaux ont montré que les sous-mots portent des informations intéressantes comme la position du mot dans la phrase avec la présence d'une majuscule. Dès lors, nous pourrions donner plus de poids à un MSE qui comporte des sous-mots. Cela ferait remonter des MSE plus explicites et permettrait peut être d'avoir des motifs caractéristiques des registres de langue encore plus interprétables.

8.2.2 Pistes d'ouverture à d'autres questions de recherche et applications

Linguistique de corpus Notre approche pourrait être adaptée pour caractériser tous types de variations linguistiques illustrées par un corpus de textes composés de sous-corpus contrastés. Par rapport à une approche classique de linguistique de corpus, elle présente deux avantages : elle est capable d'examiner un large jeux de données dont l'exploration manuelle serait chronophage et de réduire le risque d'introduire des biais avec des *a priori* sur les résultats à obtenir. Pour donner une idée de ses applications possibles en linguistique de corpus, nous donnons dans ce paragraphe un exemple de cas d'usage pour chaque type de variation issu de la typologie proposée par (Gadet 1996). La première est la variation diachronique, elle regarde l'évolution des productions linguistiques à travers

le temps. Par exemple, (Guillot 2017) a étudié l'évolution des pronoms démonstratifs du 9ème au 15ème siècle. À partir du corpus divisé en sous-corpus représentant chacun un siècle, nous pourrions découvrir des MSE distinguant les pronoms démonstratifs selon les siècles. Pour cela, notre méthodologie pourrait s'adapter en contraignant la fouille de MSE pour qu'elle écarte les MS ne comportant pas de pronom démonstratif. La deuxième variation est la variation diatopique analysant la diversité des productions linguistiques selon l'espace géographique. les travaux de (Sekaninová 2012) ont analysé la variété des formes issues du verlan [2] dans le rap selon la région des rappeurs. Son corpus est composé d'un sous-corpus représentant les rappeurs du nord et un d'un sous-corpus représentant ceux du sud. En utilisant les sous-mots pour décrire chaque mot, les MSE seraient à même de découvrir des formes linguistiques distinguant le verlan du nord de celui du sud et inversement. Enfin, la variation diastratique observe les usages linguistiques selon un angle social. L'étude de (Armstrong et al. 2001) a regardé les différences d'emplois des nouvelles formes de féminisations entre des locuteurs hommes et des locutrices femmes. Ici, la fouille de MSE serait à même de se concentrer sur les propriétés morphologiques des mots en les croisant avec les terminaisons portées par les sous-mots afin de découvrir des féminisations propres aux locuteurs et celles propres aux locutrices.

TAL Par rapport à d'autres techniques de TAL (comme les réseaux neuronaux) notre chaîne de traitement présente deux avantages pour son utilisation en entreprise. Le premier est sa capacité à extraire des connaissances à partir d'un jeu de données, réduit ou volumineux, et sans poser d'*a priori*, le second est son explicabilité. Selon leur domaine, les entreprises utilisent des techniques de classifications automatiques pour catégoriser automatiquement des données textuelles : étiqueter des commentaires selon la satisfaction des clients (en considérant les classes « satisfait », « neutre », « insatisfait ») ; ou bien des comptes rendus d'évènements selon leurs types (« incident » vs. « non-incident »), ou encore des emails (« remboursement », « échange », « autre »). Ces classifieurs ont besoin d'une grande quantité de données annotées manuellement pour s'entraîner : par exemple, nous avons annoté 4 000 tweets pour étiqueter notre corpus en registres de langue. Or, annoter manuellement des données est un processus chronophage et une tâche coûteuse lorsque des employés doivent y dédier leur temps. Notre chaîne de traitement peut réduire ce coût en permettant de découvrir des descripteurs d'apprentissages qui distinguent les

2. Le verlan est défini comme l'« Argot codé qui procède par inversion des syllabes à l'intérieur du mot (par exemple *ripou*, *pourri*). » par le Larousse en ligne.

classes entre elles en fouillant un ensemble réduit de textes annotés. En outre, tous ces acteurs du privé sont contraints par la confiance des utilisateurs, les exigences des régulateurs et la responsabilité envers les autres parties prenantes. Comme notre approche est explicable car maîtrisée de bout en bout par l'utilisateur, elle peut être utilisée pour extraire des connaissances à partir de jeux de données tout en permettant à l'utilisateur de maîtriser le processus. Par exemple, notre approche pourrait être utilisée afin de découvrir automatiquement des ensembles de descripteurs linguistiques évoluant dans le temps permettant d'adapter les descripteurs d'apprentissage pour des classifieurs étiquetant des textes comme haineux ou non. En effet, les utilisateurs des réseaux sociaux voulant déjouer la censure adaptent leur discours et emploient de nouvelles formes linguistiques. Notre approche serait à même de fouiller les données textuelles à travers le temps afin de découvrir les MSE illustrant ces stratégies d'évitement car nouvelles et émergentes par rapport à l'historique de ces utilisateurs. Au-delà du traitement du langage naturel, nos travaux peuvent s'appliquer à des domaines tels que le traitement des séries temporelles en aidant les industries à réduire leur coût de maintenance en analysant leur logs matériels ; ou la segmentation de consommateurs en utilisant les MSE pour caractériser des groupes de clients décrivant leurs habitudes d'achats croisées avec leurs profils.

Ces quelques exemples ont montré, de manière non exhaustive, le champ d'applications ouvert à notre chaîne de traitement dans le domaine de la linguistique de corpus ou celui du TAL.

Chapitre 9

Annexes

I - Glossaire

CMO : Communications Médiées par Ordinateurs

ME : Motif Ensembliste

MR : Motif Représentant

MS : Motif Séquentiel

MSE : Motif Séquentiel Émergent

SA : Sujet d'Actualité

TAL : Traitement Automatique des Langues

TC : Taux de Croissance

II - Liste des 72 descripteurs issus de (Mekki, Battistelli, Lecorvé et al. 2018)

Lorsque les descripteurs n'ont pas de référence bibliographique, cela veut dire qu'ils sont issus de notre analyse linguistique.

Niveau lexical

1. Liste non-exhaustive de termes discriminants (Branca-Rosoff 1999 ; Bilger et Cappeau 2004)
2. Emprunts étrangers (Gadet 2003)
3. Onomatopées (Ilmola 2012)
4. Adverbes spécifiques (Gadet 2003)
5. Verbes familier (Ilmola 2012)
6. Préférence pour les verbes du premier groupe
7. Locutions verbales (Ilmola 2012)
8. Décumule des comparatifs synthétiques (Gadet 1997)
9. Sur-présence de *là* (Gadet 1997)
10. Sur-présence de *et*
11. Position de *Et* en début de phrase
12. Position de *Et puis* en début de phrase
13. *faire* utilisé comme verbe de parole
14. utilisation de *des fois/parfois* selon les registres

Niveau morphologique

15. contraction de *cela* en *ça* et de *cela est* en *c'est* (Gadet 1997)
16. sujet *on* transposé en *nous* (Bilger et Cappeau 2004)
17. Terminaisons en -*asse* (Ilmola 2012)
18. Terminaisons en -*o* (Ilmola 2012)
19. Dérivation d'un nom commun ou bien d'un adjectif en adverbe (Ilmola 2012)
20. Répétition de voyelles (Branca-Rosoff 1999)
21. Utilisation des majuscules (Branca-Rosoff 1999)
22. Répétitions des signes de ponctuation (Branca-Rosoff 1999)
23. Passage du tutoiement au vouvoiement (Bilger et Cappeau 2004)
24. Augmentation du nombre de syllabes en */oe/* sous l'effet de mots verlanisés (Gadet 2003)

25. Troncation finale des mots (Gadet 1997)
26. Troncation initiale des motifs (Gadet 1997)
27. Élision du *e* (Favart 2010)
28. Apocope dans le pronom *tu* (Favart 2010)
29. Apocope du *r* (Favart 2010)
30. Liaisons parasites avec *z* (Favart 2010 ; Gadet 2003)
31. Écriture phonétique avec une utilisation des chiffres (SOMMANT 2005)
32. Mots plus courts (SOMMANT 2005)
33. Lettre unique (SOMMANT 2005)
34. Élision du pronom *qui* devant une voyelle (Gadet 1997)

Niveau syntaxique

35. Subjonctif qui aligne sur le présent (Gadet 2003)
36. Phrase interrogative sans inversion sujet verbe (Gadet 2003)
37. Syntaxe neutre avec un élément lexical discriminant (Ilmola 2012)
38. Construction de la négation complète et incomplète (Bilger et Cappeau 2004)
39. *ça* suivi d'un verbe
40. Construction du futur avec le verbe *aller*
41. Utilisation non standard des prépositions
42. « Relative populaire » (Gadet 2003)
43. Utilisation du passé composé plutôt que du passé simple
44. *c'est* devant un syntagme nominal pluriel (Favart 2010)
45. Métaphores animales (Ilmola 2012)
46. Effacement du pronom sujet *il* remplacé par *y* dans les constructions impersonnelles (Favart 2010)
47. Sujet et Verbe en un seul mot (SOMMANT 2005)
48. Transposition de la structure *qu'est-ce que* en *qu'est-ce*
49. Sur-utilisation du présent indicatif
50. Verbe *être* au singulier suivi d'un syntagme nominal pluriel (Bilger et Cappeau 2004)

51. Complément d'objet direct précédant le verbe sans être accordé
52. Participe passé du verbe *faire* suivi d'un verbe infinitif (Gadet 2003)
53. *Si* suivi d'un verbe au conditionnel
54. Fréquence basse des phrases passives
55. Sur-utilisation d'*est ce que* pour les phrases interrogatives (Ilmola 2012)
56. Fréquence des mots de liaisons (Gadet 2003)
57. Des éléments ponctuants (Gadet 2003)
58. Outils de « planification » du discours (Branca-Rosoff 1999 ; Bilger et Cappeau 2004)
59. Utilisations de phrase adverbiale
60. Substantivation et nominalisation (Bilger et Cappeau 2004)
61. Emploi de relatifs (Gadet 2003)
62. Maintient de *des* devant un adjectif
63. Pronominalisation est faite avec une préposition et non un pronom
64. Les pronoms *y* et *en* pour faire référence à un être humain (Favart 2010)
65. Sur-utilisation de *que* (Favart 2010)
66. Inversement de l'ordre des mots
67. Phrase impérative avec le complément d'objet indirect devant le complément d'objet direct
68. Phrase impérative avec le pronom *en* devant le complément d'objet indirect
69. Sur-utilisation des constructions *c'est... qui* et *c'est... que*
70. Rajout de *à lui* ou *à elle* après les pronoms *son* ou *sa* (Gadet 2003)
71. Utilisation de *que* plutôt que de *où*
72. Utilisation du pronom personnel *il* ou *elle* après un nom propre

III - Grammaires hors-contexte probabilistes pour générer des langages artificiels

Grammaire pour le registre familier

S $\longrightarrow$ S1 [0.20] | S2 [0.40] | S3 [0.40]

S1 ⟶ Pro SV PC_i [0.7] | V Pro_perso SN PC_i [0.3]
S2 ⟶ Pro neg1 V neg2 SN PC [0.3] | Pro V neg2 SN PC [0.7]
S3 ⟶ Pro_perso SV PC [1]
SV ⟶ V SN [1]
SN ⟶ DET N [0.7] | DET ADJ N [0.3]
Pro ⟶ Pro_dem [0.80] | Pro_perso [0.20]
Pro_perso ⟶ 'elle' [0.25] | 'on' [0.50] | 'il' [0.25]
Pro_dem ⟶ 'ça' [1]
neg1 ⟶ 'ne' [1]
neg2 ⟶ 'pas' [1]
V ⟶ 'chante' [0.25] | 'répète' [0.25] |'radote' [0.50]
DET ⟶ 'une' [0.34] | 'la' [0.33] | 'cette' [0.33]
ADJ ⟶ 'mélodieuse' [0.33] | 'belle' [0.34] | 'longue' [0.33]
N ⟶ 'chanson' [0.35] | 'chansonnette' [0.1] | 'musique' [0.35] | 'romance' [0.1] | 'ballade' [0.1]
PC_i ⟶ ' ?' [1]
PC ⟶ PC_a [1]
PC_a ⟶ "." [1]

Grammaire pour le registre courant

S ⟶ S1 [0.20] | S2 [0.40] | S3 [0.40]
S1 ⟶ Pro SV PC_i [0.5] | V Pro_perso SN PC_i [0.5]
S2 ⟶ Pro neg1 V neg2 SN PC [0.5] | Pro V neg2 SN PC [0.5]
S3 ⟶ Pro_perso SV PC [1]
SV ⟶ V SN [1]
SN ⟶ DET N [0.5] | DET ADJ N [0.5]
Pro ⟶ Pro_dem [0.10] | Pro_perso [0.90]
Pro_perso ⟶ 'elle' [0.33] | 'on' [0.33] | 'il' [0.34]
Pro_dem ⟶ 'ça' [0.10] | "il" [0.90]
neg1 ⟶ 'ne' [1]
neg2 ⟶ 'pas' [1]
V ⟶ 'chante' [0.33] | 'répète' [0.33] |'radote' [0.34]
DET ⟶ 'une' [0.34] | 'la' [0.33] | 'cette' [0.33]

ADJ ⟶ 'mélodieuse' [0.33] | 'belle' [0.34] | 'longue' [0.33]
N ⟶ 'chanson' [0.35] | 'chansonnette' [0.1] | 'musique' [0.35] | 'romance' [0.1] | 'ballade' [0.1]
PC_i ⟶ ' ?' [1]
PC ⟶ PC_a [1]
PC_a ⟶ "." [1]

Grammaire pour le registre soutenu
S ⟶ S1 [0.20] | S2 [0.40] | S3 [0.40]
S1 ⟶ Pro SV PC_i [0.3] | V Pro_perso SN PC_i [0.7]
S2 ⟶ Pro neg1 V neg2 SN PC [0.7] | Pro V neg2 SN PC [0.3]
S3 ⟶ Pro_perso SV PC [1]
SV ⟶ V SN [1]
SN ⟶ DET N [0.3] | DET ADJ N [0.7]
Pro ⟶ Pro_dem [0.05] | Pro_perso [0.95]
Pro_perso ⟶ 'elle' [0.45] | 'on' [0.05] | 'il' [0.50]
Pro_dem ⟶ 'ça' [0.20] | 'il' [0.80]
neg1 ⟶ 'ne' [1]
neg2 ⟶ 'pas' [1]
V ⟶ 'chante' [0.40] | 'répète' [0.40] |'radote' [0.20]
DET ⟶ 'une' [0.34] | 'la' [0.33] | 'cette' [0.33]
ADJ ⟶ 'mélodieuse' [0.33] | 'belle' [0.34] | 'longue' [0.33]
N ⟶ 'chanson' [0.1] | 'chansonnette' [0.1] | 'musique' [0.1] | 'romance' [0.35] | 'ballade' [0.35]
PC_i ⟶ ' ?' [1]
PC ⟶ PC_a [1]
PC_a ⟶ "." [1]

IV - Exemples de langages artificiels générés

Registre familier

on chante la belle musique ?

ça répète pas la mélodieuse musique .
ça radote pas cette romance .
ça répète pas une musique .
on radote pas la musique .
il répète une mélodieuse musique .
on répète la chansonnette .
on chante une musique .
ça radote pas cette longue chansonnette .
ça radote pas la chanson .
ça radote pas la chanson .
ça radote pas la chanson .
ça ne radote pas une musique .
ça radote pas une longue chanson .
elle répète une musique .
chante on une chanson ?
on chante cette chanson .
elle radote une romance .
on radote une longue musique .
ça ne radote pas la chanson .
ça répète pas une musique .
il répète cette belle chanson .
ça radote pas cette romance .
ça ne répète pas cette mélodieuse musique .

Registre courant

il répète cette belle chanson .
chante il la belle chansonnette ?
elle radote pas la musique .
elle chante la mélodieuse chanson .
il répète la chanson .
il répète la romance ?
on répète pas la musique .
on répète une belle musique .
elle radote pas une mélodieuse musique .
on ne radote pas la longue musique .
il ne chante pas la longue chanson .
il radote la belle chanson .
elle ne répète pas une ballade .
répète on cette chanson ?
répète on cette mélodieuse chanson ?
elle radote la belle musique .

radote on une chanson ?
elle chante une musique .
elle chante pas une musique .
on radote la longue ballade .
on chante pas la chanson .
il répète pas cette chanson .
elle répète pas la longue chanson .
il ne chante pas la belle chanson .
répète elle la chanson ?
elle chante pas une mélodieuse musique .

Registre soutenu

il ne chante pas une belle ballade .
elle répète une belle romance .
chante il la belle chansonnette ?
elle répète pas une belle romance .
elle ne chante pas la romance .
répète elle la mélodieuse ballade ?
elle répète une belle ballade ?
chante il la longue ballade ?
il chante la chansonnette .
elle ne chante pas la mélodieuse ballade .
il répète cette longue musique .
il répète cette ballade ?
elle ne répète pas cette mélodieuse ballade .
chante il la belle musique ?
elle chante pas la belle romance .
elle ne chante pas la romance .
il chante la ballade .
chante il une ballade ?
il radote cette mélodieuse romance .
répète il la longue ballade ?
elle chante pas cette mélodieuse ballade .
on ne chante pas cette belle ballade .
il chante une belle chanson .
elle ne répète pas une longue romance .
elle ne répète pas la mélodieuse romance .
il chante pas cette mélodieuse chansonnette .
elle répète une belle ballade .
il ne radote pas la ballade .
il répète une belle musique .

Bibliographie personnelle

— Mekki, Jade, Delphine Battistelli, Gwénolé Lecorvé et al. (2018), « Identification de descripteurs pour la caractérisation de registres », in : *Rencontre des jeunes chercheurs en traitement automatique du langage naturel et recherche d'information (CORIA-TALN-RJC).*

— Mekki, Jade, Nicolas Béchet et al. (2020), « Caractérisation de registres de langue par extraction de motifs séquentiels émergents », in : *JADT 2020 : 15èmes Journées Internationales d'Analyse statistique des Données Textuelles.*

— Mekki, Jade, Gwénolé Lecorvé et al. (2021), « TREMoLo-Tweets : a Multi-Label Corpus of French Tweets for Language Register Characterization », in : *RANLP 2021-Recent Advances in Natural Language Processing*

— Mekki, Jade, Delphine Battistelli, Nicolas Béchet et al. (2021), « TREMoLo : un corpus multi-étiquettes de tweets en français pour la caractérisation des registres de langue », in : *Traitement Automatique des Langues Naturelles.*

— Lecorvé, Gwénolé, Hugo Ayats, Benoît Fournier, Jade Mekki et al. (2018), « Construction conjointe d'un corpus et d'un classifieur pour les registres de langue en français » in : *Traitement automatique du langage naturel (TALN).* 2018.

— Lecorvé, Gwénolé, Hugo Ayats, Benoît Fournier, Jade Mekki et al. (2019), « Towards the automatic processing of language registers : Semi-supervisedly built corpus and classifier for french » in : *International Conference on Computational Linguistics and Intelligent Text Processing (CICLing).*

BIBLIOGRAPHIE

ADAM, Jean-Michel (2002), « Le style dans la langue et dans les textes », in : *Langue française*, p. 71-94.

AGARWAL, Apoorv et al. (2011), « Sentiment analysis of twitter data », in : *Proceedings of the workshop on language in social media (LSM 2011)*, p. 30-38.

AGRAWAL, Rakesh, Ramakrishnan SRIKANT et al. (1994), « Fast algorithms for mining association rules », in : *Proc. 20th int. conf. very large data bases, VLDB*, t. 1215, Citeseer, p. 487-499.

AGRAWAL, Rakesh et Ramakrishnan SRIKANT (1995), « Mining sequential patterns », in : *Proceedings of the eleventh international conference on data engineering*, IEEE, p. 3-14.

ALTSCHUL, Stephen F et al. (1990), « Basic local alignment search tool », in : *Journal of molecular biology* 215.*3*, p. 403-410.

AMO, Sandra de et al. (2004), « An Apriori-based Approach for First-Order Temporal Pattern Mining. », in : *SBBD*, p. 48-62.

ARGAMON, Shlomo, Moshe KOPPEL, Jonathan FINE et al. (2003), « Gender, genre, and writing style in formal written texts », in : *Text & Talk* 23.*3*, p. 321-346.

ARGAMON, Shlomo, Moshe KOPPEL, James W PENNEBAKER et al. (2007), « Mining the blogosphere : Age, gender and the varieties of self-expression », in : *First Monday* 12.*9*.

ARGAMON, Shlomo, Casey WHITELAW et al. (2007), « Stylistic text classification using functional lexical features », in : *Journal of the Association for Information Science and Technology* 58.*6*.

ARGAMON, Shlomo Engelson (2019), « Register in computational language research », in : *Register Studies* 1.*1*, p. 100-135.

ARMSTRONG, Nigel et al. (2001), « La langue française au féminin : le sexe et le genre affectent-ils la variation linguistique ? », in : *La langue française au féminin*, p. 1-238.

ASEERVATHAM, Sujeevan, Aomar OSMANI et Emmanuel VIENNET (2006), « bitSPADE : A lattice-based sequential pattern mining algorithm using bitmap representation », in : *Sixth International Conference on Data Mining (ICDM'06)*, IEEE, p. 792-797.

Auer, Peter (2008), *Style and social identities : Alternative approaches to linguistic heterogeneity*, t. 18, Walter de Gruyter.

Authier, Jacqueline et André Meunier (1972), « Norme, grammaticalité et niveaux de langue », in : *Langue française 16*, p. 49-62.

Ayres, Jay et al. (2002), « Sequential pattern mining using a bitmap representation », in : *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 429-435.

Baggioni, Daniel et Marie-Louise Moreau (1997), « Norme », in : *M.-L. Moreau, Sociolinguistiqueconcepts de base, Mardaga*, p. 217-223.

Bailey, James, Thomas Manoukian et Kotagiri Ramamohanarao (2002), « Fast algorithms for mining emerging patterns », in : *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, p. 39-50.

Bally, Charles (1909), *Traité de stylistique française*, t. 3, C. Winter.

Bastide, Yves et al. (2002), « Pascal : un algorithme d'extraction des motifs fréquents », in : *Technique et science informatiques* 21.*1*, p. 65-95.

Bayardo Jr, Roberto J (1998), « Efficiently mining long patterns from databases », in : *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, p. 85-93.

Beaufort, Richard, Sophie Roekhaut et Cédrick Fairon (2008), « Définition d'un système d'alignement SMS/français standard à l'aide d'un filtre de composition », in : *Proceedings of JADT 2008*, p. 155-166.

Beccucci, Laurène (2018), « Pierre HALTÉ, Les émoticônes et les interjections dans le tchat. Limoges : Éditions Lambert Lucas, 2018 », in : *Communication et organisation. Revue scientifique francophone en Communication organisationnelle 54*, p. 253-255.

Béchet, Nicolas et al. (2012), « Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares », in : *23es Journées Francophones d'Ingénierie des Connaissances-IC 2012*, p. 149-164.

— (2015), « Sequence mining under multiple constraints », in : *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, p. 908-914.

Belvisi, Nicole Mariah Sharon, Naveed Muhammad et Fernando Alonso-Fernandez (2020), « Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features », in : *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, IEEE, p. 1-6.

Benyoucef, Sana (2011), « Analyse linguistique des figures de style dans (les fleurs du mal) de », thèse de doct., Batna.

Bernstein, Basil (1977), « Langage et classes sociales : codes socio-linguistiques et contrôle social », in.

BERTOCCHINI, P et E Costanzo (2010), « La notion de registre de langue », in : *Le français dans le monde 371*, p. 26-27.

Biber, Douglas (1991), *Variation across speech and writing*, Cambridge University Press.

— (1994), « An analytical framework for register studies », in : *Sociolinguistic perspectives on register*, p. 31-56.

— (1995), *Dimensions of register variation : A cross-linguistic comparison*, Cambridge University Press.

— (2006), *University language : A corpus-based study of spoken and written registers*, t. 23, John Benjamins Publishing.

— (2012), « Register as a predictor of linguistic variation », in : *Corpus Linguistics and Linguistic Theory* 8.*1*, p. 9-37.

Biber, Douglas et Susan Conrad (2019), *Register, genre, and style*, Cambridge University Press.

Biber, Douglas et Jesse Egbert (2018), *Register variation online*, Cambridge University Press.

Biber, Douglas et Edward Finegan (1990), « ARCHER (A Representative Corpus of Historical English Registers) », in.

— (1994), *Sociolinguistic perspectives on register*, Oxford University Press on Demand.

Biber, Douglas, Stig Johansson et al. (2002), « Longman grammar of spoken and written English. London : Longman, 1999. Hard-back Ǧ69. Pp. xii+ 1,204. ISBN 0 582 23725 4. Reviewed by Manfred Krug, University of Freiburg », in : *English Language and Linguistics* 6.*379ą416*.

Biber Douglas et Conrad, Susan (2001), « Register variation : A corpus approach », in : *The handbook of discourse analysis*, p. 175-196.

Bilger, Mireille et Paul Cappeau (2004), « L'oral ou la multiplication des styles », in : *Langage et société 3*, p. 13-30.

Blanche-Benveniste, Claire et al. (1990), « Le français parlé(études grammaticales) », in : *Sciences du langage*.

Bonnard, Henri (1981), *Code du français courant*, Magnard.

Borde, Davy (2018), *Tirons la langue : plaidoyer contre le sexisme dans la langue française*, Les Éditions Utopia.

Borzeix, Anni et Béatrice Fraenkel (2005), « Langage et travail(communication, cognition, action) », in : *CNRS communication*.

Bourdieu, Pierre (2014), *Langage et pouvoir symbolique*, Édition du Seuil.

Bourquin, Guy (1965), « Niveaux, aspects et registres de langage. Remarques à propos de quelques problèmes théoriques et pédagogiques », in.

Branca-Rosoff, Sonia (1999), « Des innovations et des fonctionnements de langue rapportés à des genres », in : *Langage & société* 87.*1*, p. 115-129.

Breiman, Leo (2001), « Random forests », in : *Machine learning* 45.*1*, p. 5-32.

Brooke, Julian, Tong Wang et Graeme Hirst (2010), « Automatic acquisition of lexical formality », in : *Coling 2010 : Posters*, p. 90-98.

Buhmann, Martin D (2000), « Radial basis functions », in : *Acta numerica* 9, p. 1-38.

Burfoot, Clint et Timothy Baldwin (2009), « Automatic satire detection : Are you having a laugh ? », in : *Proceedings of the ACL-IJCNLP 2009 conference short papers*, p. 161-164.

Caillies, Stéphanie (2009), « Descriptions de 300 expressions idiomatiques : familiarité, connaissance de leur signification, plausibilité littérale,«décomposabilité» et «prédictibilité» », in : *LAnnee psychologique* 109.*3*, p. 463-508.

Cajolet-Laganière, Hélène et Pierre Martel (2011), « Les marques de niveaux de langue dans les dictionnaires : synthèse et proposition », in : *Français du Canada-Français de France*, Max Niemeyer Verlag, p. 185-198.

Caliński, Tadeusz et Jerzy Harabasz (1974), « A dendrite method for cluster analysis », in : *Communications in Statistics-theory and Methods* 3.*1*, p. 1-27.

Calvet, Louis-Jean (2021), *Charles Bally*, `https://www.universalis.fr/encyclopedie/charles-bally/`, Accessed : 2021-12-16.

Capelle, Matthieu, Cyrille Masson et Jean-François Boulicaut (2002), « Mining frequent sequential patterns under a similarity constraint », in : *International conference on intelligent data engineering and automated learning*, Springer, p. 1-6.

Cellier, Peggy, Thierry Charnois et Marc Plantevit (2010), « Sequential patterns to discover and characterise biological relations », in : *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, p. 537-548.

Červenková, Marie (2014), *L'Enrichissement du français contemporain des sources argotiques. Etude diachronique des passages d'un corpus lexical entre les registres de langue*, Presses Académiques Francophones.

Charaudeau, Patrick et Dominique Maingueneau (2002), *Dictionnaire d'analyse du discours*, Seuil.

Charnet, Chantal (1994), « Lecture de : Les Ponctuants de la langue de D. Vincent », in : *Cahiers de praxématique 23*, p. 153-154.

Charolles, Michel et Béatrice Lamiroy (2001), « Syntaxe phrastique et transphrastique du but au résultat », in : *Macrosyntaxe et macrosémantique, Actes du colloque international dAarhus, 17-19 mai 2001*, Peter Lang, p. 383-419.

Chen, Tao et Min-Yen Kan (2013), « Creating a live, public short message service corpus : the NUS SMS corpus », in : *Language Resources and Evaluation* 47.*2*, p. 299-335.

Chevalier, Jean-Claude (2007), « Rey, Alain, Frédéric Duval & Gilles Siouffi, Mille ans de langue française : histoire d'une passion », in : *Histoire Épistémologie Langage* 29.*2*, p. 228-231.

Chuquet, Hélène (1990), *Pratique de la traduction : anglais-français*, Editions OPHRYS.

Chuquet, Hélène et Michel Paillard (1987), *Approche linguistique des problèmes de traduction anglais-français*, Editions Ophrys.

Cougnon, Louise-Amélie et Cédrick Fairon (2014), *SMS Communication : A linguistic approach*, t. 61, John Benjamins Publishing Company.

Craig, Hugh, Arthur F Kinney et al. (2009), *Shakespeare, computers, and the mystery of authorship*, Cambridge University Press.

Crémilleux, Bruno et al. (2009), « Discovering knowledge from local patterns in sage data », in : *Data Mining and Medical Knowledge Management : Cases and Applications*, IGI Global, p. 251-267.

Crystal, David (2008), *Txtng : The gr8 db8*, OUP Oxford.

— (2011), *Internet linguistics : A student guide*, Routledge.

Damourette, Jacques et Edouard Pichon (1930), *Essai de grammaire de langue française*, d'Artrey.

Dangel, Jacqueline (2007), « Le registre des voix rhétoriques et théâtrales romaines : de la Republique à l'Empire », in : *Le registre des voix rhétoriques et théâtrales romaines : de la Republique à l'Empire*, p. 109-136.

Davies, David L et Donald W Bouldin (1979), « A cluster separation measure », in : *IEEE transactions on pattern analysis and machine intelligence 2*, p. 224-227.

De Saussure, Ferdinand (1972), « Cours de linguistique générale (1916) », in : *Edition critique préparée par T. de mauro. paris : payothèque*.

De Vel, Olivier, Alison Anderson et al. (2001a), « Mining e-mail content for author identification forensics », in : *ACM Sigmod Record* 30.*4*, p. 55-64.

— (2001b), « Mining e-mail content for author identification forensics », in : *ACM Sigmod Record* 30.*4*.

De Vel, Olivier, Malcolm Corney et al. (2002), « Language and gender author cohort analysis of e-mail for computer forensics », in : *Proceedings of Digital Forensics Research Workshop*, p. 1-16.

Dehaspe, Luc et Hannu Toivonen (1999), « Discovery of frequent datalog patterns », in : *Data Mining and knowledge discovery* 3.*1*, p. 7-36.

Demaizière, Colette (1989), « Les niveaux de langue dans le roman québécois : Michel Tremblay et Réjean Ducharme », in : *Cahiers de l'AIEF* 41.*1*, p. 81-98.

Demanuelli, Claude et Jean Demanuelli (1991), *Lire et traduire : anglais-français*, Masson.

Develotte, Christine et Marie-Anne Paveau (2017), « Pratiques discursives et interactionnelles en contexte numérique. Questionnements linguistiques », in : *Langage et société 2*, p. 199-215.

Devolder, Laurence-Lola (2007), « Langue et registre (s) : illustration par l'indice d'usage familier », thèse de doct., Paris 5.

Dewaele, Jean-Marc (2001), « Une distinction mesurable : corpus oraux et écrits sur le continuum de la deixis », in : *Journal of French Language Studies* 11.*2*, p. 179-199.

Dong, Guozhu et Jinyan Li (1999), « Efficient mining of emerging patterns : Discovering trends and differences », in : *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 43-52.

Dong, Guozhu et Jian Pei (2007), *Sequence data mining*, t. 33, Springer Science & Business Media.

Dostie, Gaétane et Claus D Pusch (2007), « Présentation. Les marqueurs discursifs. Sens et variation », in : *Langue française 2*, p. 3-12.

Dunn, Joseph C (1973), « A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters », in.

Durand, Jacques et Chantal Lyche (1999), « Regard sur les glissantes en français : français standard, français du Midi », in : *Cahiers de grammaire* 24, p. 39-65.

Eckert, Penelope (2003), *The meaning of style*, na.

Eisenstein, Jacob (2013), « What to do about bad language on the internet. », in : *Proceedings of HLT-NAACL*.

El Bouziki, Mohammed (2021), « Marques des niveaux de langue et subjectivité des lexicographes », in : *Langues, Cultures et Communication* 5.*1*, p. 1-9.

Escalante, Hugo Jair, Thamar Solorio et Manuel Montes-y-Gómez (2011), « Local histograms of character n-grams for authorship attribution », in : *Proceedings of HTL-ACL*.

Fan, Hongjian et Kotagiri Ramamohanarao (2003), « Efficiently mining interesting emerging patterns », in : *International Conference on Web-Age Information Management*, Springer, p. 189-201.

Fauconnier, Jean-Philippe (2015), *French Word Embeddings*, url : `http://fauconnier.github.io`.

Favart, Françoise (2010), « Le stéréotype de registre de langue populaire dans le roman du second XXe siècle (1966-2006) », in : *Textes et contextes 5*.

Ferguson, Charles A (1982), « Simplified registers and linguistic theory », in : *Exceptional language and linguistics*, p. 49-66.

Ferreira, Pedro Gabriel et Paulo J Azevedo (2005), « Protein sequence pattern mining with constraints », in : *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, p. 96-107.

Fiot, Celine, Anne Laurent et Maguelonne Teisseire (2005), « Motifs séquentiels flous : un peu, beaucoup, passionnément », in : *EGC : Extraction et Gestion des Connaissances*, p. 507-518.

Fiot, Céline, Gérard Dray et al. (2004), « A la recherche des motifs séquentiels flous », in : *12èmes Rencontres Francophones sur la Logique Floue et ses Applications (LFA)*, Cépadues Editions, p. 131-138.

Fiot, Céline, Anne Laurent et Maguelonne Teisseire (2006), « Des motifs séquentiels généralisés aux contraintes de temps étendues », in : *EGC : Extraction et Gestion des Connaissances*, p. 603-614.

Fournier-Viger, Philippe, Antonio Gomariz et al. (2014), « Fast vertical mining of sequential patterns using co-occurrence information », in : *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, p. 40-52.

Fournier-Viger, Philippe, Cheng-Wei Wu, Antonio Gomariz et al. (2014), « VMSP : Efficient vertical mining of maximal sequential patterns », in : *Canadian conference on artificial intelligence*, Springer, p. 83-94.

Fournier-Viger, Philippe, Cheng-Wei Wu et Vincent S Tseng (2013), « Mining maximal sequential patterns without candidate maintenance », in : *International Conference on Advanced Data Mining and Applications*, Springer, p. 169-180.

Francis, W Nelson et Henry Kucera (1979), « Brown corpus manual », in : *Letters to the Editor* 5.*2*, p. 7.

Frei, Henri (1971), *La grammaire des fautes : introduction à la linguistique fonctionnelle, assimilation et différenciation, brièveté et invariabilité, expressivité*, t. 1, Slatkine.

Fuchs, Catherine (2021a), *LANGUE REGISTRES DE*, `http://www.universalis-edu.com.passerelle.univ-rennes1.fr/encyclopedie/registres-de-langue/`, Accessed : 2021-12-14.

— (2021b), *Norme linguistique*, `http://www.universalis-edu.com.passerelle.univ-rennes1.fr/encyclopedie/norme-et-usage/`, Accessed : 2021-12-17.

Gadet, Françoise (1996), « Niveaux de langue et variation intrinsèque », in : *Palimpsestes. Revue de traduction 10*, p. 17-40.

— (1997), « La variation, plus qu'une écume », in : *Langue française*, p. 5-18.

— (2000), « Français de référence et syntaxe », in : *Cahiers de l'Institut de Linguistique de Louvain* 26.*1-4*, p. 265-283.

— (2003), « Is there a French theory of variation ? », in.

— (2007), *La variation sociale en français*, Editions Ophrys.

Gallardo, Catherine Camugli (2005), « Niveaux de langue et variations linguistiques dans la comparaison interlangue de tables du Lexique-Grammaire », in : *Lingvisticae investigationes* 28.*2*, p. 169-188.

Garcia-Hernandez, Rene Arnulfo, Jose Francisco Martinez-Trinidad et Jesus Ariel Carrasco-Ochoa (2006), « A new algorithm for fast discovery of maximal sequential patterns in a document collection », in : *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, p. 514-523.

Garmadi, Juliette (1981), *La sociolinguistique*, FeniXX.

Gautier, Antoine (2014), « Phrase et syntaxe : sur quelques aspects de l'intégration », in : *Langue française 2*, p. 27-41.

Gianfortoni, Philip, David Adamson et Carolyn P Rosé (2011), « Modeling of stylistic variation in social media with stretchy patterns », in : *Proceedings of the Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*.

Glasco, Sarah (2011), « Contrastes : grammaire du français courant by Denise Rochat », in : *The French Review* 84.*4*, p. 829-830.

Go, Alec, Richa Bhayani et Lei Huang (2009), « Twitter sentiment classification using distant supervision », in : *CS224N project report, Stanford* 1.*12*, p. 2009.

Goethals, Bart (2003), « Survey on frequent pattern mining », in : *Univ. of Helsinki* 19, p. 840-852.

Golobic, Marjan (1982), « Quatre niveaux de langue tels que reflетés par le Petit Robert », thèse de doct., University of British Columbia.

Golubéva-Monatkina, Nathalia (1991), « Gadet, Françoise. Le français ordinaire. Paris : Armand Colin, 1989 », in : *Canadian Modern Language Review* 47.*4*, p. 800-802.

Gomariz, Antonio et al. (2013), « Clasp : An efficient algorithm for mining frequent closed sequences », in : *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, p. 50-61.

Gonon, Laetitia et al. (2016), « Sur la scène de crime… : Enquête sur les enjeux linguistiques et stylistiques de motifs récurrents dans le thriller contemporain », in : *SHS Web of Conferences*, t. 27, EDP Sciences, p. 06006.

Greimas, Algirdas Julien et Joseph Courtés (1979), « Dictionnaire raisonné de la théorie du langage », in : *Paris : Hachette*.

Guan, En-Zheng et al. (2005), « Mining maximal sequential patterns », in : *2005 International Conference on Neural Networks and Brain*, t. 1, IEEE, p. 525-528.

Guerin, Emmanuelle (2008), « Le'français standard' : une variété située ? », in : *Congrès Mondial de Linguistique Française*, EDP Sciences, p. 200.

Guillot, Céline (2017), *Le démonstratif en français : étude de sémantique grammaticale diachronique (9ème-15ème siècles)*, Peeters.

Guralnik, Valerie et George Karypis (2001), « A scalable algorithm for clustering sequential data », in : *Proceedings 2001 IEEE international conference on data mining*, IEEE, p. 179-186.

Halliday, Michael Alexander Kirkwood (1994), « Language as social semiotic », in : *Language and literacy in social practice*, p. 23-43.

Han, Jiawei, Jian Pei, Behzad Mortazavi-Asl et al. (2000), « FreeSpan : frequent pattern-projected sequential pattern mining », in : *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 355-359.

Han, Jiawei, Jian Pei et Yiwen Yin (2000), « Mining frequent patterns without candidate generation », in : *ACM sigmod record* 29.*2*, p. 1-12.

Han, Jiawei, Jian Pei, Yiwen Yin et Runying Mao (2004), « Mining frequent patterns without candidate generation : A frequent-pattern tree approach », in : *Data mining and knowledge discovery* 8.*1*, p. 53-87.

Han, Meng, Zhihai Wang et Jidong Yuan (2015), « Mining Closed and Multi-Supports-Based Sequential Pattern in High-Dimensional Dataset. », in : *International Arab Journal of Information Technology (IAJIT)* 12.*4*.

He, Yulan et Deyu Zhou (2011), « Self-training from labeled features for sentiment analysis », in : *Information Processing & Management* 47.*4*.

Hellermann, Jacqueline (1969), *Le Français courant II*.

Heylighen, Francis et Jean-Marc Dewaele (1999), « Formality of language : definition, measurement and behavioral determinants », in : *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel* 4.

Hirst, Graeme et Ol'ga Feiguina (2007), « Bigrams of syntactic labels for authorship discrimination of short texts », in : *Literary and Linguistic Computing* 22.*4*.

Hmida, Firas et al. (2018), « Assisted Lexical Simplification for French Native Children with Reading Difficulties », in : *The Workshop of Automatic Text Adaptation, 11th International Conference on Natural Language Generation*.

Ho, Joshua, Lior Lukov et Sanjay Chawla (2005), « Sequential pattern mining with constraints on large protein databases », in : *Proceedings of the 12th international conference on management of data (COMAD)*, Citeseer, p. 89-100.

Holat, Pierre, Marc Plantevit et al. (2014), « Sequence classification based on delta-free sequential patterns », in : *2014 IEEE International Conference on Data Mining*, IEEE, p. 170-179.

Holat, Pierre, Nadi Tomeh et Thierry Charnois (2015), « Classification de texte enrichie à l'aide de motifs séquentiels », in : *TALN 2015*.

Holmes, David I (1994), « Authorship attribution », in : *Computers and the Humanities* 28.*2*, p. 87-106.

Hong, Tzung-Pei, Kuie-Ying Lin et Shyue-Liang Wang (2001), « Mining fuzzy sequential patterns from multiple-item transactions », in : *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*, t. 3, IEEE, p. 1317-1321.

Huang, Minlie et al. (2004), « Discovering patterns to extract protein–protein interactions from full texts », in : *Bioinformatics* 20.*18*, p. 3604-3612.

Hymes, Dell (2013), *Foundations in sociolinguistics : An ethnographic approach*, Routledge.

Ilmola, Maarit (2012), *Les registres familier, populaire et vulgaire dans le canard enchaıné et charlie hebdo : étude comparative*.

Indyk, Piotr, Nick Koudas et Shanmugavelayutham Muthukrishnan (2000), « Identifying representative trends in massive time series data sets using sketches », in : *26th International Conference on Very Large Data Bases, VLDB 2000*, p. 363-372.

Inokuchi, Akihiro, Takashi Washio et Hiroshi Motoda (2003), « Complete mining of frequent patterns from graphs : Mining graph data », in : *Machine Learning* 50.*3*, p. 321-354.

Iqbal, Farkhund et al. (2013), « A unified data mining solution for authorship analysis in anonymous textual communications », in : *Information Sciences* 231.

Irvine, Judith T (1985), « Status and style in language », in : *Annual Review of Anthropology* 14.*1*, p. 557-581.

Ishihara, Shunichi (2011), « A forensic authorship classification in sms messages : A likelihood ratio based approach using n-gram », in : *Proceedings of the Australasian Language Technology Association Workshop 2011*, p. 47-56.

Ivancsy, Renata et Istvan Vajk (2005), « Efficient sequential pattern mining algorithms », in : *WSEAS Transactions on Computers* 4.*2*, p. 96-101.

Jackiewicz, Agata et Marko Vidak (2014), « Étude sur les mots-dièse », in : *shs Web of Conferences*, t. 8, EDP Sciences, p. 2033-2050.

Jacques, Anis (1999), « Internet, communication et langue française », in.

Jaubert, Anna (2007), « La diagonale du style. Étapes d'une appropriation de la langue », in : *Pratiques* 135.*1*, p. 47-62.

Joos, Martin (1967), *The five clocks*, t. 58, New York : Harcourt, Brace & World.

Juan, SUN et PU Zhihong (2018), « Pour une approche sociolinguistique en didactique du français langue étrangère-l'argot français contemporain en classe. », in : *Synergies Chine 13*.

Just, Marcel A et Patricia A Carpenter (1980), « A theory of reading : From eye fixations to comprehension. », in : *Psychological review* 87.*4*, p. 329.

Kalmbach, Jean-Michel (2012), *La grammaire du français langue étrangère pour étudiants finnophones*.

Khalid, Osama et Padmini Srinivasan (2020), « Style matters ! Investigating linguistic style in online communities », in : *Proceedings of the International AAAI Conference on Web and Social Media*, t. 14, p. 360-369.

Kiesling, Scott F (2009), « Style as stance », in : *Stance : sociolinguistic perspectives*, p. 171-194.

Kishida, Kazuaki (2005), *Property of average precision and its generalization : An examination of evaluation indicator for information retrieval experiments*, National Institute of Informatics Tokyo, Japan.

Kobus, Catherine, François Yvon et Géraldine Damnati (2008), « Normalizing SMS : are two metaphors better than one ? », in : *Proceedings of COLING*.

Koch, Peter et Wulf Oesterreicher (2001), « Langage parlé et langage écrit », in.

Koppel, Moshe et Jonathan Schler (2003), « Exploiting stylistic idiosyncrasies for authorship attribution », in : *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, t. 69, p. 72-80.

Koppel, Moshe, Jonathan Schler et Shlomo Argamon (2011), « Authorship attribution in the wild », in : *Language Resources and Evaluation* 45.*1*, p. 83-94.

Kress, Gunther (2009), *What is mode ? I : C. Jewitt (red.)*

Kumar, Pradeep et al. (2007), « Rough clustering of sequential data », in : *Data & Knowledge Engineering* 63.*2*, p. 183-199.

Labov, William (1966), « The effect of social mobility on linguistic behavior », in : *Sociological Inquiry* 36.*2*, p. 186-203.

— (1972), *Language in the inner city : Studies in the Black English vernacular*, 3, University of Pennsylvania Press.

— (1988), « The judicial testing of linguistic theory », in : *Language in Context : Connecting Observation and Understanding, Norwood, Ablex*.

Labrosse, Céline (2021), *Pour une langue sans sexisme : petit traité pratique pour un usage au quotidien*, Groupe Fides Inc.

Lafontaine, Dominique (1986), *Le parti pris des mots*, t. 156, Editions Mardaga.

Lahiri, Shibamouli (2015), « SQUINKY ! A corpus of sentence-level formality, informativeness, and implicature », in : *arXiv preprint arXiv :1506.02306*.

Lebrun, Monique, Nathalie Lacelle et Jean-François Boutin (2012), « Genèse et essor du concept de littératie médiatique multimodale », in : *Mémoires du livre/Studies in Book Culture* 3.*2*.

Lecorvé, Gwénolé et al. (2018), « Construction conjointe d'un corpus et d'un classifieur pour les registres de langue en français », in : *Traitement automatique du langage naturel (TALN)*.

— (2019), « Towards the Automatic Processing of Language Registers : Semi-supervisedly Built Corpus and Classifier for French », in.

Ledegen, Gudrun (2021), *Normes*, Sociolinguistique du contact : Dictionnaire des termes et concepts [en ligne]. Lyon : ENS Éditions, 2013 (généré le 17 décembre 2021) `http://books.openedition.org/enseditions/12480`, Accessed : 2021-12-17.

Ledegen, Gudrun et Isabelle Léglise (2013), *Variations et changements linguistiques*.

Legallois, Dominique, Thierry Charnois et Thierry Poibeau (2016), « Repérer les clichés dans les romans sentimentaux grâce à la méthode des «motifs» », in : *Lidil. Revue de linguistique et de didactique des langues 53*, p. 95-117.

Levenshtein, Vladimir I et al. (1966), « Binary codes capable of correcting deletions, insertions, and reversals », in : *Soviet physics doklady*, t. 10, 8, Soviet Union, p. 707-710.

Li, Jinyan, Guimei Liu et Limsoon Wong (2007), « Mining statistically important equivalence classes and delta-discriminative emerging patterns », in : *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 430-439.

Lin, Jerry Chun-Wei, Ting Li, Philippe Fournier-Viger et Tzung-Pei Hong (2015), « A fast algorithm for mining fuzzy frequent itemsets », in : *Journal of Intelligent & Fuzzy Systems* 29.*6*, p. 2373-2379.

Lin, Jerry Chun-Wei, Ting Li, Philippe Fournier-Viger, Tzung-Pei Hong et al. (2017), « Efficient mining of multiple fuzzy frequent itemsets », in : *International Journal of Fuzzy Systems* 19.*4*, p. 1032-1040.

Lin, Nancy P, Hung-Jen Chen et al. (2007), « Mining negative fuzzy sequential patterns », in : *Proceedings of the 7th WSEAS international conference on simulation, modelling and optimization*, p. 15-17.

Lin, Nancy P, Wei-Hua Hao et al. (2007), « Fast mining maximal sequential patterns », in : *The International Conference on Simulation, Modeling and Optimization*, Citeseer, p. 405-408.

Ling, Rich et Naomi S Baron (2007), « Text messaging and IM : Linguistic comparison of American college data », in : *Journal of language and social psychology* 26.*3*, p. 291-298.

Lu, Eric Hsueh-Chan, Vincent S Tseng et S Yu Philip (2010), « Mining cluster-based temporal mobile sequential patterns in location-based service environments », in : *IEEE transactions on knowledge and data engineering* 23.*6*, p. 914-927.

Lucas, Nadine et Bruno Crémilleux (2004), « Fouille de textes hiérarchisée appliquée à la détection de fautes », in : *Document numérique* 8.*3*, p. 107-133.

Lucas, Nadine, Bruno Crémilleux et Leny Turmel (2003), « Signalling well-written academic articles in an English corpus by text mining techniques », in : *UCREL technical papers* 16, p. 465-474.

Luo, Congnan et Soon M Chung (2005), « Efficient mining of maximal sequential patterns using multiple samples », in : *Proceedings of the 2005 SIAM International Conference on Data Mining*, SIAM, p. 415-426.

MacLeod, Nicci et Tim Grant (2012), « Whose Tweet ? Authorship analysis of microblogs and other short-form messages », in.

MacQueen, James et al. (1967), « Some methods for classification and analysis of multivariate observations », in : *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, t. 1, 14, Oakland, CA, USA, p. 281-297.

Magué, Jean-Philippe, Nathalie Rossi-Gensane et Pierre Halté (2020), « De la segmentation dans les tweets : signes de ponctuation, connecteurs, émoticônes et émojis », in : *Corpus 20*.

Maingueneau, Dominique (2007), « Genres de discours et modes de généricité », in : *Le français aujourd'hui 4*, p. 29-35.

— (2012), *Analyser les textes de communication*, Armand Colin.

Malblanc, Alfred (1944), *Pour une stylistique comparée du français et de l'allemand : essai de représentation linguist. comparée*, Didier.

Manessy, Gabriel (1994), « Normes endogènes et français de référence », in : *M. Beniamino, C.*

Marcellesi, Christiane (1976), « Norme et enseignement du français », in : *Cahiers de linguistique sociale 1*, p. 1-9.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez et al. (2020), « CamemBERT : a Tasty French Language Model », in : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez et al. (2019), « CamemBERT : a tasty French language model », in : *arXiv preprint arXiv :1911.03894*.

Marton, Yuval, Ning Wu et Lisa Hellerstein (2005), « On Compression-Based Text Classification. », in : *Proceedings of the European Conference on Information Retrieval (ECIR)*, t. 3408.

Masseglia, Florent, Pascal Poncelet et Maguelonne Teisseire (1999), « Extraction efficace de motifs séquentiels : le prétraitement des données. », in : *Proc. 15emes Journees Bases de Donnees Avancees, BDA*, p. 341-360.

Masseglia, Florent, Maguelonne Teisseire et Pascal Poncelet (2004), « Extraction de motifs séquentiels. problemes et méthodes », in : *Revue des Sciences et Technologies de l'Information-Série ISI : Ingénierie des Systèmes d'Information* 9.*3/4*, p. 183-210.

Maurais, Jacques (2008), *Les Québécois et la norme : L'évaluation par les Québécois de leurs usages linguistiques*, Office québécois de la langue française Québec.

McCarthy, Philip M. et al. (2006), « Analyzing Writing Styles with Coh-Metrix. », in : *Proceedings of the FLAIRS Conference*.

McClosky, David, Eugene Charniak et Mark Johnson (2006), « Effective self-training for parsing », in : *Proceedings of HLT-NAACL*.

McQuail, Denis (2010), *McQuail's mass communication theory*, Sage publications.

McSherry, Frank et Marc Najork (2008), « Computing information retrieval performance measures efficiently in the presence of tied scores », in : *European conference on information retrieval*, Springer, p. 414-421.

Mekki, Jade, Delphine Battistelli, Nicolas Béchet et al. (2021), « TREMoLo : un corpus multi-étiquettes de tweets en français pour la caractérisation des registres de langue », in : *Traitement Automatique des Langues Naturelles*, ATALA, p. 224-232.

Mekki, Jade, Delphine Battistelli, Gwénolé Lecorvé et al. (2018), « Identification de descripteurs pour la caractérisation de registres », in : *Rencontre des jeunes chercheurs en traitement automatique du langage naturel et recherche d'information (CORIA-TALN-RJC)*.

— (2021), « TREMoLo-Tweets corpus : guide d'annotation pour un corpus annoté en registres de langue pour le français », in.

Mekki, Jade, Nicolas Béchet et al. (2020), « Caractérisation de registres de langue par extraction de motifs séquentiels émergents », in : *JADT 2020 : 15èmes Journées Internationales d'Analyse statistique des Données Textuelles*.

Mekki, Jade, Gwénolé Lecorvé et al. (2021), « TREMoLo-Tweets : a Multi-Label Corpus of French Tweets for Language Register Characterization », in : *RANLP 2021-Recent Advances in Natural Language Processing*.

Mercier, Louis, C Verreault et T Lavoie (2002), « Le français, une langue qui varie selon les contextes », in : *Le français, une langue à apprivoiser. Textes des conférences prononcées au Musée de la civilisation (Québec, 2000-2001) dans le cadre de l'exposition «Une grande langue : le français dans tous ses états*, p. 41-60.

Milroy, Lesley (1986), « Social network and linguistic focusing », in : *Dialect and language variation*, Elsevier, p. 367-380.

Mohan, Ashwin, Ibrahim M Baggili et Marcus K Rogers (2010), « Authorship attribution of SMS messages using an N-grams approach », in : *Proceedings of CERIAS Tech Report* 11, p. 1-12.

Mohtasseb, Haytham, Amr Ahmed et al. (2009), « Mining online diaries for blogger identification », in.

Möller, Steffen, Evgenia V Kriventseva et Rolf Apweiler (2000), « A collection of well characterised integral membrane proteins », in : *Bioinformatics* 16.*12*, p. 1159-1160.

Monneret, Philipe et Fabrice Poli (2020), *Grammaire du français Terminologie grammaticale*, rapp. tech.

Moreau, Marie-Louise (1999), « La pluralité des normes dans la francophonie », in : *DiversCité Langues* 4.

Muller, Bodo (1985), *Le français d'aujourd'hui*, t. 47, Klincksieck.

Narkhede, Sarang (2018), « Understanding AUC-ROC Curve », in : *Towards Data Science* 26.

Nespoulous, J-L et André Borrell (1979), « De la diversité des usages linguistiques : quelle (s) langue (s) enseigner ? », in : *Langues (Les) Modernes Lavausseau* 73.*2-3*, p. 260-271.

Okuno, Syunya, Hiroki Asai et Hayato Yamana (2014), « A challenge of authorship identification for ten-thousand-scale microblog users », in : *2014 IEEE International Conference on Big Data (Big Data)*, IEEE, p. 52-54.

Osthus, Dietmar (2004), « Le bon usage d'Internet : le discours normatif sur la Toile », in : *disponible sur http ://www. dietmar-osthus. de/norme. htm.[Page consultée le 26 mars 2014.]*

Pak, Alexander et Patrick Paroubek (2010), « Twitter as a corpus for sentiment analysis and opinion mining. », in : *LREc*, t. 10, 2010, p. 1320-1326.

Panuccio, Antonello, Manuele Bicego et Vittorio Murino (2002), « A Hidden Markov Model-based approach to sequential data clustering », in : *Joint IAPR international*

workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR), Springer, p. 734-743.

Paquette, Jean-Marcel (1983), « Procès de normalisation et niveaux/registres de langue », in : *La norme linguistique, Québec/Paris : Gouvernement du Québec, CLF et Le Robert*, p. 367-381.

Pasquier, Nicolas (1999), « Closed sets based discovery of association rules », in : *GDR-I3'1999 Meeting*, p. 55-64.

— (2000), « Extraction de Bases pour les Règles d'Association à partir des Itemsets Fermés Fréquents », in : *Inforsid'2000 Congress*, p. 56-77.

Pasquier, Nicolas et al. (1999), « Efficient mining of association rules using closed itemset lattices », in : *Information systems* 24.*1*, p. 25-46.

Paveau, Marie-Anne (2013), « Genre de discours et technologie discursive. Tweet, twittécriture et twittérature », in : *Pratiques. Linguistique, littérature, didactique 157-158*, p. 7-30.

— (2017), *L'analyse du discours numérique. Dictionnaire des formes et des pratiques*, Hermann.

Paveau, Marie-Anne et Laurence Rosier (2008), *La langue française. Passions et polémiques*.

Pavlick, Ellie et Joel Tetreault (2016), « An empirical analysis of formality in online communication », in : *Transactions of the Association of Computational Linguistics* 4.*1*.

Pei, Jian, Jiawei Han, Hongjun Lu et al. (2001), « H-mine : Hyper-structure mining of frequent patterns in large databases », in : *proceedings 2001 IEEE international conference on data mining*, IEEE, p. 441-448.

Pei, Jian, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang et al. (2004), « Mining sequential patterns by pattern-growth : The prefixspan approach », in : *IEEE Transactions on knowledge and data engineering* 16.*11*, p. 1424-1440.

Pei, Jian, Jiawei Han, Behzad Mortazavi-Asl et Hua Zhu (2000), « Mining access patterns efficiently from web logs », in : *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, p. 396-407.

Pennebaker, James W, Martha E Francis et Roger J Booth (2001), « Linguistic inquiry and word count : LIWC 2001 », in : *Mahway : Lawrence Erlbaum Associates* 71.*2001*, p. 2001.

Perrin-Naffakh, Anne-Marie (1985), *Le cliché de style en français moderne : nature linguistique et rhétorique, fonction littéraire*, Presses Univ de Bordeaux.

Peterson, Kelly, Matt Hohensee et Fei Xia (2011), « Email formality in the workplace : A case study on the Enron corpus », in : *Proceedings of the Workshop on Languages in Social Media*, Association for Computational Linguistics, p. 86-95.

Pétillon, Sabine (2006), « Style, critique génétique et modèles rédactionnels : perspectives linguistiques », in : *Corpus 5*.

Pettit, Zoë (2005), « Translating register, style and tone in dubbing and subtitling », in : *Journal of Specialised Translation* 4.*4*, p. 49-65.

Pierozak, Isabelle (2003), « Le "français tchaté" : un objet à géométrie variable ? », in : *Langage et société 2*, p. 123-144.

Pillière, Linda (1997), « Étude linguistique de quelques propriétés du style de Virginia Woolf », thèse de doct., Paris 4.

Pinto, Helen et al. (2001), « Multi-dimensional sequential pattern mining », in : *Proceedings of the tenth international conference on Information and knowledge management*, p. 81-88.

Plantevit, Marc et Bruno Crémilleux (2009), « Condensed representation of sequential patterns according to frequency-based measures », in : *International Symposium on Intelligent Data Analysis*, Springer, p. 155-166.

Plantevit, Marc, Anne Laurent et Maguelonne Teisseire (2006), « HYPE : Prise en compte des hiérarchies lors de l'extraction de motifs séquentiels multidimensionnels », in : *EDA'06 : Entrepôts de Données et Analyse en ligne*, Cépaduès, p. 155-176.

— (2008), « Fouille de données multidimensionnelles : différentes stratégies pour prendre en compte la mesure », in : *EDA : Entrepôts de Données et l'Analyse en ligne*, Cépaduès, p. 61-76.

Plecháč, Petr (2019), « Relative contributions of Shakespeare and Fletcher in Henry VIII : An Analysis Based on Most Frequent Words and Most Frequent Rhythmic Patterns », in : *arXiv preprint arXiv :1911.05652*.

Poisson, Esther (2012), « L'oral, l'écrit et les registres », in : *Correspondance*.

Pöll, Bernhard et Elmar Schafroth (2010), « Normes et hybridation linguistiques en francophonie : actes de la section 6 du Congrès de l'Association des francoromanistes allemands, Augsbourg, 24-26 septembre 2008 », in : *Normes et hybridation linguistiques en francophonie*, p. 1-272.

Poudat, Céline et Frédéric Landragin (2017), *Explorer un corpus textuel : Méthodes-pratiques-outils*, De Boeck Superieur.

Prikhodkine, Alexei (2011), « Dynamique normative du français en usage en Suisse romande : enquête sociolinguistique dans les cantons de Vaud, Genève et Fribourg », in : *Dynamique normative du francais en usage en Suisse romande*, p. 1-339.

Quillard, Virginie (2000), « Interroger en français parlé : études syntaxique, pragmatique et sociolinguistique », thèse de doct., Tours.

Quiniou, Solen et al. (2012), « Fouille de données pour la stylistique : cas des motifs séquentiels émergents », in : *Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12)*, p. 821-833.

Rabatel, Julien (2011), « Extraction de motifs contextuels : Enjeux et applications dans les données séquentielles », in : *Theses, Université MontpellierII Sciences et Techniques du Languedoc*.

Ragel, Roshan, Pramod Herath et Upul Senanayake (2013), « Authorship detection of SMS messages using unigrams », in : *2013 IEEE 8th International Conference on Industrial and Information Systems*, IEEE, p. 387-392.

Raissi, Chedy et Pascal Poncelet (2008), « Échantillonnage pour l'extraction de motifs séquentiels : des basesde données statiques aux flots de données. », in : *EGC*, p. 145-156.

Rao, Sudha et Joel Tetreault (2018), « Dear sir or madam, may i introduce the gyafc dataset : Corpus, benchmarks and metrics for formality style transfer », in : *arXiv preprint arXiv :1803.06535*.

Rastier, François (2005), « Enjeux épistémologiques de la linguistique de corpus », in : *La linguistique de corpus 31-45*.

Rebourcet, Séverine (2008), « Le français standard et la norme : l'histoire d'une «nationalisme linguistique et littéraire» à la française », in : *Communication, lettres et sciences du langage* 2.*1*, p. 107-118.

Rey, Alain (1972), « Usages, jugements et prescriptions linguistiques », in : *Langue française 16*, p. 4-28.

Rigat, Francoise et E Piccardo (2008), « ça marche ! Grammaire du français courant. », in.

Rousseeuw, Peter J (1987), « Silhouettes : a graphical aid to the interpretation and validation of cluster analysis », in : *Journal of computational and applied mathematics* 20, p. 53-65.

Saevanee, Hataichanok, Nathan Clarke et Steven Furnell (2011), « SMS linguistic profiling authentication on mobile device », in : *2011 5th International Conference on Network and System Security*, IEEE, p. 224-228.

Sanders, Carol (1993), *French today : language in its social context*, Cambridge University Press.

Sanderson, Conrad et Simon Guenter (2006), « Short text authorship attribution via sequence kernels, Markov chains and author unmasking : An investigation », in : *Proceedings of the 2006 EMNPL*, Association for Computational Linguistics, p. 482-491.

Saneifar, Hassan et al. (2008), « S2mp : Similarity measure for sequential patterns », in : *AusDM : Australasian Data Mining*, t. 87, ACS, p. 095-104.

Schler, Jonathan et al. (2006), « Effects of Age and Gender on Blogging. », in : *Proceedings of the AAAI spring symposium : Computational approaches to analyzing weblogs*, t. 6.

Schmid, Helmut (1994), « TreeTagger-a language independent part-of-speech tagger », in : *http ://www. ims. uni-stuttgart. de/projekte/corplex/TreeTagger/*.

Sekaninová, Tereza (2012), « Stéréotypes liés au verlan : variation diatopique dans le rap français », in : *Mémoire de DEA, Université Masaryk*.

Sequeira, Karlton et Mohammed Zaki (2002), « Admit : anomaly-based data mining for intrusions », in : *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 386-395.

Sheikha, Fadi Abu et Diana Inkpen (2010), « Automatic classification of documents by formality », in : *IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*.

Sidorov, Grigori et al. (2014), « Syntactic n-grams as machine learning features for natural language processing », in : *Expert Systems with Applications* 41.*3*.

Simonin, Jacky et Sylvie Wharton (2013), *Sociolinguistique du contact. Dictionnaire des termes et concepts*, ENS éditions.

Sinclair, John (2005), « Corpus and text-basic principles », in : *Developing linguistic corpora : A guide to good practice* 92, p. 1-16.

Smith, DJ, S Spencer et T Grant (2009), « Authorship analysis for counter terrorism », in : *Unpublished Research Report, QinetiQ/Aston University*.

Sohn, Dae-Neung, Jung-Tae Lee et Hae Chang Rim (2009), « The contribution of stylistic information to content-based mobile spam filtering », in : *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, p. 321-324.

Sohn, Dae-Neung, Joong-Hwi Shin et al. (2008), « Contents-based korean sms spam filtering using morpheme unit features », in : *Annual Conference on Human and Language Technology*, Human et Language Technology, p. 195-200.

Songram, Panida et Veera Boonjing (2008), « Closed multidimensional sequential pattern mining », in : *International Journal of Knowledge Management Studies* 2.*4*, p. 460-479.

Soulet, Arnaud et Bruno Crémilleux (2008), « Adequate condensed representations of patterns », in : *Data mining and knowledge discovery* 17.*1*, p. 94-110.

Soulet, Arnaud et François Rioult (2014), « Extraire les motifs minimaux efficacement et en profondeur », in : *Journées Francophones Extraction et Gestion des Connaissances, EGC 2014*, Hermann-Éditions, pp-383.

Srikant, Ramakrishnan et Rakesh Agrawal (1996), « Mining sequential patterns : Generalizations and performance improvements », in : *International conference on extending database technology*, Springer, p. 1-17.

Srinivasan, Ashwin, Ross D King et Douglas W Bristol (1999), « An assessment of submissions made to the predictive toxicology evaluation challenge », in : *IJCAI*, t. 99, Citeseer, p. 270-275.

Stamatatos, Efstathios (2009), « A survey of modern authorship attribution methods », in : *Journal of the American Society for information Science and Technology* 60.*3*, p. 538-556.

Stourdzé, Colette et May Collet-Hassan (1969), « Les niveaux de langue », in : *Le français dans le monde* 65, p. 18-21.

Styler, Will (2011), « The enronsent corpus », in.

Tanasa, Doru et Brigitte Trousse (2003), « Le prétraitement des fichiers logs web dans le "Web Usage Mining" multi-sites », in : *Journées Francophones de la Toile (JFT'2003)*, p. 113-122.

Taylor, Paul J et Sally Thomas (2008), « Linguistic style matching and negotiation outcome », in : *Negotiation and Conflict Management Research* 1.*3*, p. 263-281.

Theodoridis, Sergios et Konstantinos Koutroumbas (2006), *Pattern recognition*, Elsevier.

Thomson, Rob et Tamar Murachver (2001), « Predicting gender from electronic discourse », in : *British Journal of Social Psychology* 40.*2*, p. 193-208.

Thurlow, Crispin et Alex Brown (2003), « Generation Txt ? The sociolinguistics of young people's text-messaging », in : *Discourse analysis online* 1.*1*, p. 30.

Todorov, Tzvetan (2013), *Mikhaıl Bakhtine. Le principe dialogique. Suivi de : Ecrits du Cercle de Bakhtine*, Le Seuil.

Tognini-Bonelli, Elena (2001), *Corpus linguistics at work*, t. 6, John Benjamins Publishing.

Trudgill, Peter et al. (1974), *The social differentiation of English in Norwich*, t. 13, CUP archive.

Tseng, Vincent Shin-Mu et Kawuu WC Lin (2005), « Mining sequential mobile access patterns efficiently in mobile web systems », in : *19th International Conference on Advanced Information Networking and Applications (AINA'05) Volume 1 (AINA papers)*, t. 2, IEEE, p. 762-767.

Tutin, Agnès et Olivier Kraif (2016), « Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents », in : *Lidil. Revue de linguistique et de didactique des langues 53*, p. 119-141.

Tyne, Henry (2012), « La variation dans l'enseignement-apprentissage d'une langue 2 », in : *Le français aujourd'hui 1*, p. 103-112.

Uno, Takeaki, Tatsuya Asai et al. (2003), « LCM : An Efficient Algorithm for Enumerating Frequent Closed Item Sets. », in : *Fimi*, t. 90, Citeseer.

Uno, Takeaki, Masashi Kiyomi, Hiroki Arimura et al. (2004), « LCM ver. 2 : Efficient mining algorithms for frequent/closed/maximal itemsets », in : *Fimi*, t. 126.

Uno, Takeaki, Masashi Kiyomi et Hiroki Arimura (2005), « Lcm ver. 3 : Collaboration of array, bitmap and prefix tree for frequent itemset mining », in : *Proceedings of the 1st international workshop on open source data mining : frequent pattern mining implementations*, p. 77-86.

Ure, Jean (1982), « Introduction : approaches to the study of register range », in : *International Journal of the Sociology of Language* 1982.*35*, p. 5-24.

Urieli, Assaf et Ludovic Tanguy (2013), « L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane », in : *20e conférence du Traitement Automatique du Langage Naturel (TALN)*, publication-en.

Valdman, Albert (1982), « Français standard et français populaire : sociolectes ou fictions ? », in : *French Review*, p. 218-227.

Vickers, Brian (2011), « Shakespeare and authorship studies in the twenty-first century », in : *Shakespeare Quarterly* 62.*1*, p. 106-142.

Vijayalakshmi, S, V Mohan et S Suresh Raja (2010), « Mining of users access behavior for frequent sequential pattern from web logs », in : *International Journal of Database Management System (IJDM)* 2.

Villeneuve, Anne-José (s. d.), « Langue de ville et langue de soirée : variation stylistique et maintien des contraintes en français québécois soutenu Anne-José Villeneuve David Rosychuk Davy Bigot Université de l'Alberta Université de l'Alberta Université Concordia Plusieurs études ont analysé le français québécois (FQ) familier (Sankoff, Sankoff, Laberge et », in : ().

Vincent, Diane (1993), *Les ponctuants de la langue et autres mots du discours*, Nuit blanche.

Wagner, Robert-Léon et Bernard Quemada (1969), « Pour une analyse des français contemporains », in : *Le français dans le monde* 69, p. 61-73.

Wang, Jianyong et Jiawei Han (2004), « BIDE : Efficient mining of frequent closed sequences », in : *Proceedings. 20th international conference on data engineering*, IEEE, p. 79-90.

Wang, Qian, Darryl N Davis et Jiadong Ren (2016), « Mining frequent biological sequences based on bitmap without candidate sequence generation », in : *Computers in biology and medicine* 69, p. 152-157.

Weber, Corinne (2017), *Le nouveau visage de la pluralité langagière : repères et questionnement à l'heure du numérique.*

— (2019), « Interrogations épistémologiques autour de l'oralité. Quel paradigme pour la didactique de la prononciation de demain ? », in : *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle* 16.*16-1.*

Wolfram, Walter A (1969), « A Sociolinguistic Description of Detroit Negro Speech. Urban Language Series, No. 5. », in.

Wu, Xiaozhu et Ximei Zhang (2019), « An efficient pixel clustering-based method for mining spatial sequential patterns from serial remote sensing images », in : *Computers & Geosciences* 124, p. 128-139.

Xatara, Claudia (2002), « La traduction phraséologique », in : *Meta : journal des traducteurs/Meta : Translators' Journal* 47.*3*, p. 441-444.

Yan, Xifeng, Jiawei Han et Ramin Afshar (2003), « Clospan : Mining : Closed sequential patterns in large datasets », in : *Proceedings of the 2003 SIAM international conference on data mining*, SIAM, p. 166-177.

Yang, Albert C-C et al. (2003), « Information categorization approach to literary authorship disputes », in : *Physica A : Statistical Mechanics and its Applications* 329.*3-4*, p. 473-483.

Yang, Min et Kam-Pui Chow (2014), « Authorship attribution for forensic investigation with thousands of authors », in : *IFIP International Information Security Conference*, Springer, p. 339-350.

Yu, Chung-Ching et Yen-Liang Chen (2005), « Mining sequential patterns from multidimensional sequence data », in : *IEEE Transactions on Knowledge and Data Engineering* 17.*1*, p. 136-140.

Yun, Xiong et Zhu Yangyong (2007), « BioPM : An efficient algorithm for protein motif mining », in : *2007 1st International Conference on Bioinformatics and Biomedical Engineering*, IEEE, p. 394-397.

Zadeh, Lotfi A (1996), « Fuzzy sets », in : *Fuzzy sets, fuzzy logic, and fuzzy systems : selected papers by Lotfi A Zadeh*, World Scientific, p. 394-432.

Zaki, Mohammed J (2001), « SPADE : An efficient algorithm for mining frequent sequences », in : *Machine learning* 42.*1*, p. 31-60.

Zaki, Mohammed Javeed (2000), « Scalable algorithms for association mining », in : *IEEE transactions on knowledge and data engineering* 12.*3*, p. 372-390.

Zhang, Xiuzhen, Guozu Dong et Ramamohanarao Kotagiri (2000), « Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets », in : *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 310-314.

Zhao, Ying et Justin Zobel (2007), « Searching with style : Authorship attribution in classic literature », in : *ACM International Conference Proceeding Series*, t. 244, Citeseer, p. 59-68.

Zhou, Baoyao, Siu Cheung Hui et Alvis Cheuk M Fong (2004), « CS-mine : an efficient WAP-tree mining for web access patterns », in : *Asia-Pacific Web Conference*, Springer, p. 523-532.

Zribi-Hertz, Anne (2011), « Pour un modèle diglossique de description du français : quelques implications théoriques, didactiques et méthodologiques », in : *Journal of French Language Studies* 21.*2*, p. 231-256.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : Caractérisation des registres de langues par extraction de motifs séquentiels émergents

Mot clés : registres de langues, traitement automatique des langues, motifs séquentiels

Résumé : Cette thèse s'intéresse à la caractérisation automatique des registres de langue. Sur le plan linguistique, notre contribution est d'étudier les apports des techniques de traitement automatique des langues pour extraire de nouvelles connaissances à propos des registres familier, courant et soutenu. Sur le plan informatique, nous avons proposé une méthode suffisamment générique et non supervisée pour caractériser tout type de variation linguistique, les registres s'apparentant alors à un cas d'usage. Dans le manuscrit, nous dressons tout d'abord un état des lieux des multiples différentes définitions présentes dans la littérature, par rapport auquel nous positionnons nos travaux. Nous présentons alors la constitution linguistiquement motivée d'un large corpus de tweets en français annotés en registres. Les annotations résultent d'un procédé semi-supervisé fondé sur une graine annotée manuellement en registres et un classifieur qui généraliste les annotations à l'ensemble des tweets. À partir de ce corpus annoté, nous montrons ensuite que l'emploi de techniques d'extraction de motifs séquentiels émergents permet d'extraire des traits linguistiques caractéristiques des registres étudiés. Enfin, nous détaillons notre approche pour réduire le nombre de motifs extraits en vue d'une meilleure interprétabilité des caractérisations produites.

Title: Characterisation of language registers using emerging sequential pattern extraction

Keywords: Language registers, natural language processing, sequential patterns

Abstract: This PhD thesis aims at automatically characterising language registers. From a linguistic point of view, our contribution is to study the potential of natural language processing techniques to extract new knowledge about the casual, neutral, and formal registers. On the computational side, we have proposed a sufficiently generic and unsupervised method to characterise any type of linguistic variation, the registers then being similar to a use case. The manuscript first draws up an inventory of the many different definitions present in the literature, against which we position our work. Second, the constitution of a large lingustically-motivated corpus of French tweets annotated in registers is presented. The annotations result from a semi-supervised process based on a seed manually annotated in registers and a classifier that generalizes the annotations to all the tweets. Based on this annotated corpus, we then show that the use of emergent sequential pattern extraction techniques enables the extraction of linguistic peculiarities of the registers under study. Finally, we detail our approach for reducing the number of extracted patterns, which allows a better interpretability of the characterizations produced.