TREMoLo : un corpus multi-étiquettes de tweets en français pour la caractérisation des registres de langue

Jade Mekki, Delphine Battistelli, Nicolas Béchet, Gwénolé Lecorvé

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

TREMoLo : un corpus multi-étiquettes de tweets en français pour la caractérisation des registres de langue

Jade Mekki[1, 2] Delphine Battistelli[2] Nicolas Béchet[3] Gwénolé Lecorvé[1]
(1) IRISA, 263 Avenue Général Leclerc, 35 000 Rennes, France
(2) Modyco, 200 Avenue de la République, 92 001 Nanterre, France
(3) IRISA, Campus de Tohannic – Rue Yves Mainguys, 56 000 Vannes, France
prenom.nom@irisa.fr, prenom.nom@parisnanterre.fr

RÉSUMÉ

Des registres tels que familier, courant et soutenu sont un phénomène immédiatement perceptible par tout locuteur d'une langue. Ils restent encore peu étudiés en traitement des langues (TAL), en particulier en dehors de l'anglais. Cet article présente un large corpus de tweets en français annotés en registres de langue. L'annotation intègre des marqueurs propres à ce type de textes (tels que les émoticônes ou les hashtags) et habituellement évincés dans les travaux en TAL. À partir d'une graine annotée manuellement en proportion d'appartenance aux registres, un classifieur de type CamemBERT est appris et appliqué sur un large ensemble de tweets. Le corpus annoté en résultant compte 228 505 tweets pour un total de 6 millions de mots. Des premières analyses statistiques sont menées et permettent de conclure à la qualité du corpus présenté. Le corpus ainsi que son guide d'annotation sont mis à la disposition de la communauté scientifique.

ABSTRACT

TREMoLo : a Multi-Label Corpus of French Tweets for Language Register Characterization

The casual, neutral, and formal language registers are a highly perceptible characteristic of discourse productions. However, they are still poorly studied in natural language processing, especially outside English, and for new textual types like tweets. To stimulate this line of research, this paper introduces a large corpus of French tweets annotated in language registers. It has been built on a preliminary detailed linguistic analysis of tweets. After training a multi-label CamemBERT classifier on a manually annotated subset, the whole corpus of tweets has been automatically labeled. The final corpus counts 228 505 tweets for a total of 6M words. Initial statistical analyses are conducted and allow us to conclude that the corpus presented is of good quality. The corpus and its annotation guide are available to the scientific community.

MOTS-CLÉS : registres de langue, CamemBERT, corpus annoté, tweets.

KEYWORDS: language registers, CamemBERT, annotated corpus, tweets.

1 Introduction

Le registre de langue dans lequel se situe un texte (à l'oral comme à l'écrit) apparaît comme un trait saillant. Il renvoie au contexte d'énonciation dans lequel il est — ou a été — produit (et qui comprend notamment la relation du locuteur avec ses interlocuteurs). Parmi les manifestations possibles de ce phénomène sociolinguistique, le partitionnement en registres tels que familier, courant et soutenu est

probablement le plus répandu. Si des corpus comme GYAFC (Rao & Tetreault, 2018) — où ce type de variations est appelé « niveau de formalité » — ont récemment popularisé le domaine, celui-ci est encore globalement peu étudié en traitement automatique des langues (TAL), et particulièrement en dehors de l'anglais. Par ailleurs, bien que de nouveaux types de textes aient émergé depuis les deux dernières décennies — tels que les tweets, et plus généralement ceux que l'on range sous le terme *Communications Médiées par Ordinateurs (CMO)* —, les travaux sur les registres de langue traitent surtout des types plus classiques de textes dont les caractéristiques sont plus ou moins connues de la littérature linguistique (on associera ainsi généralement par exemple les insultes au registre familier et la diversité de connecteurs logiques à du registre soutenu). Dès lors, les analyses de corpus CMO en termes de registres de langue constituent un défi tant pour la linguistique descriptive que pour les différentes applications en TAL. Pour répondre à ces enjeux, cet article présente le corpus TREMoLo [1], constitué de 228 505 tweets en français (6M mots), annotés en registres de langue familier, courant et soutenu. À partir d'une graine annotée manuellement, les annotations ont été généralisées à l'ensemble du corpus en utilisant un classifieur fondé sur CamemBERT. Après un état de l'art en section 2, la composition du corpus est présentée en section 3, les protocoles d'annotation manuelle et automatique sont décrits en sections 4 et 5. Enfin, la qualité du corpus et quelques premiers résultats statistiques sur la caractérisation des registres sont exposés en section 5.

2 État de l'art

En sociolinguistique, la notion de registre de langue fait globalement référence à la perception de variétés linguistiques associées à des situations de communication particulières (Todorov, 2013) et il est admis qu'un registre de langue peut être caractérisé par des motifs spécifiques (Ferguson, 1982). L'utilisation des termes *« niveau »*, *« style »* ou *« genre »* coexistent (Gadet, 1996; Bourquin, 1965; Joos, 1967) pour désigner ce phénomène, mais le terme *« registre »* semble tout de même tendre à prévaloir, du moins dans la littérature anglo-saxonne (Biber, 1991; Sanders, 1993; Ure, 1982). En linguistique de corpus, c'est ce dernier qui est retenu en particulier dans les travaux de (Biber & Conrad, 2019; Biber, 1991). Ils étudient quantitativement la présence de traits linguistiques définis *a priori* sur un corpus [2] selon différents axes : oral/écrit, formel/informel, etc. L'objectif est d'identifier les cooccurrences de traits selon ces axes. En TAL, pour l'anglais, (Peterson *et al.*, 2011; Pavlick & Tetreault, 2016) proposent des techniques de classification de textes en formel vs. informel à partir d'un corpus de courriers électroniques tandis que (Sheikha & Inkpen, 2010) utilise une régression pour prédire un niveau de formalité à partir d'un corpus de textes formel [3]/informel [4]. Pour le français, dans (Lecorvé *et al.*, 2019), les auteurs étudient conjointement une tâche de classification et de construction d'un corpus de données web [5] annoté en utilisant une approche semi-supervisée. Ces différents travaux présentent plusieurs limites : la composition des corpus montre différents biais en mélangeant les types de textes, les annotations manuelles ne suivent pas de guide d'annotation, et aucune des techniques de prédiction ne prend en compte le fait qu'un même texte puisse être perçu comme appartenant simultanément à plusieurs registres. De ce fait, la qualité de ces corpus annotés peut être discutée. L'approche que nous présentons dans cet article répond à ce besoin en proposant à la communauté un large corpus de données annotées en registres de langue pour le

1. https://gitlab.inria.fr/jmekki/tremolo_corpus
2. Les corpus *T2K-SWAL*, *LSWE* et *ARCHER*.
3. Les *Late Modern English Corpus*, *Enron Email Corpus* et *Open American National Corpus*.
4. Les *Reuters Corpus* et *Open American National Corpus*.
5. 400 000 pages web ont été collectées à partir de requêtes composées de lexiques familier, courant et soutenu.

français dont la qualité est assurée par : la suppression des biais liés à la présence de multiples types de textes, l'élaboration d'un guide d'annotation [6] pour l'annotation manuelle, et l'annotation automatique multi-étiquettes plus mimétique de la réalité.

3 Constitution du corpus

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 type de textes issu des CMO 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 (*TT* pour « Trending Topics ») dans une zone géographique donnée, la diversité des productions devrait être représentative des différentes fonctions du langage et de différents registres de langue. Aussi, si (Sinclair, 2005) avance que la constitution d'un corpus doit se fonder sur des « *critères externes* », c'est à dire les fonctions de communication des textes, cet article tend justement à en découvrir de nouvelles parmi les CMO : afin de ne pas poser d'*a priori* sur ces dernières elles n'ont pas été utilisées pour la constitution du corpus qui s'est basée sur les TT. L'API de Twitter [7] permet, à partir d'un identifiant de lieu (dans notre cas celui de Paris), de récupérer automatiquement les 50 TT. Pour chaque TT, 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 dates différentes. Elles couvrent une durée totale d'un mois. Les tweets non français ont été repérés grâce à un module python [8] qui, pour un texte donné, prédit une langue à une certaine probabilité P. Si $P \leq 0.90$ pour le français, alors le texte est conservé dans le corpus. La valeur de P est fixée afin de garder des textes avec la présence de quelques termes non français intéressants tels que « *lol* », « *dead* », « *stan* »... Quant aux tweets tronqués, ils ont été repérés grâce à un signe de ponctuation spécifique de Twitter : trois points de suspension resserrés différents des « ... » classiques. Une règle symbolique a écarté les tweets qui se terminaient par ce signe de ponctuation particulier. Finalement, après l'exclusion des tweets non français ou tronqués, le corpus compte 228 505 tweets (6 201 339 mots).

6. https://hal.archives-ouvertes.fr/hal-03218217
7. https://developer.twitter.com/en/docs
8. https://pypi.org/project/langdetect/

4 Annotation manuelle

Une analyse linguistique du corpus est faite afin de proposer un guide d'annotation qui inclut certains éléments linguistiques spécifiques aux tweets. Une de nos contributions est de les intégrer au lieu de les écarter comme dans (Agarwal *et al.*, 2011; Pak & Paroubek, 2010; Go *et al.*, 2009).

Descripteurs linguistiques pour l'analyse des CMO (Paveau, 2013) utilise le terme « *technomorphèmes* » pour désigner les formes qui découlent des discours numériques. Parmi elles, les hashtags qui 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 & Vidak, 2014). En plus d'intégrer à notre analyse ce type de fonction pour les hashtags car nous pensons qu'elle joue un rôle dans la perception de registres dans les tweets, nous proposons d'intégrer en outre le degré d'intégration syntaxique plus ou moins fort des hashtags. Un autre type de *technomorphèmes* est le pictogramme qui se réfère à la fois à un « émoticône » [9] et à un « emoji » [10]. Nous utilisons les trois fonctions de la typologie proposée par (Magué *et al.*, 2020) en les adaptant à l'analyse de notre corpus : la fonction de remplacement (quand un pictogramme remplace un syntagme) ; la fonction d'illustration (quand il a une fonction référentielle) ; la fonction de modalisation (quand il indique l'émotion ou l'attitude énonciative de l'auteur). Nous ajoutons une autre fonction : la fonction d'encadrement/structuration (lorsqu'il entoure ou pointe vers une information). 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 *et al.*, 2018) avec ces traits spécifiques aux technomorphèmes, notre annotation prend au final en compte un ensemble de 52 descripteurs experts (dont certains sont présentés tables 1 et 2).

Protocole d'annotation Sur l'ensemble du corpus, 4 000 tweets (ou textes) ont été sélectionnés au hasard pour être annotés manuellement en proportion de registres de langue. Puisque nous divisons l'espace linguistique en 3 registres (familier, courant et soutenu), les catégories utilisées pour l'annotation sont les mêmes. Une catégorie « poubelle » est ajoutée pour les tweets mal encodés ou incompréhensibles. L'annotateur doit ordonner les registres en fonction de leur prédominance dans un texte en leur attribuant un rang [11] qui doit être justifié par la présence d'au moins un descripteur de la liste issue de l'analyse linguistique préliminaire. Chaque rang est ensuite transformé en proportion de registre. Soit :

- R un ensemble de registres ayant obtenu un rang et $Card(R)$ son nombre d'éléments,
- $r_i \in R$ un registre de R ayant obtenu le rang i,
- $inv(i)$ le rang inversé du rang i défini par $inv(i) = Card(R) - i + 1$
- srg la somme des rangs définie par $srg = \sum_{i=1}^{Card(R)} i$

La proportion du registre r_i est alors définie par $Prop_{r_i} = \frac{inv(i)}{srg}$

Ainsi, pour un texte annoté comme ceci, r_1=familier, r_2=soutenu et r_3=courant, on obtient :

- $R = \{r_1, r_2, r_3\}$ et $Card(R) = 3$,
- $inv(1) = 3$, $inv(2) = 2$, $inv(3) = 1$
- $srg = 6$

9. Un émoticône est un signe graphique ressemblant à une émotion (Beccucci, 2018).
10. Un emoji est un symbole répertorié dans une base de données (ibid.).
11. À noter que ne pas mettre de rang signifie que le registre n'est pas présent dans le texte selon l'annotateur.

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 1 – Top 3 des descripteurs (* : issus de notre analyse linguistique pour les CMO) qui caractérisent les registres dans le corpus annoté 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 1.

Les proportions en registre sur cet exemple sont donc familier 50% ($\frac{3}{6}$), soutenu 33% ($\frac{2}{6}$) et courant 17% ($\frac{1}{6}$).

Chaque texte est annoté par 2 annotateurs experts [12]. Ne sont conservées que les étiquettes présentes dans l'intersection des 2 annotations. Pour chaque étiquette sa moyenne est calculée. Sur les 4 000 tweets annotés manuellement par 4 annotateurs, 976 textes ne sont pas dans l'intersection entre les 2 annotations. Une seconde annotation est alors faite par un 5ème annotateur expert. Après cette seconde annotation, 976 tweets sont dans l'intersection. Au final, 3 269 tweets annotés manuellement sont conservés, soit 81,73% des textes totaux de la graine.

Résultats de l'annotation manuelle Les résultats de l'annotation manuelle sont dominés par le registre courant (50,5% de la graine), puis par le familier (38,7%) et enfin le soutenu (10,2%). Afin de caractériser un registre de langue R_1 par rapport aux autres registres (notés R_2), l'importance de chaque descripteur D vu dans R_1 est mesurée en calculant son taux de croissance (TC) calculé à partir de sa fréquence relative, notée $Freq$, dans R_1 et R_2 :

$$TC(D_{R_1|R_2}) = \begin{cases} \infty, & si Freq_{R_2}(D) = 0 \\ \frac{Freq_{R_1}(D)}{Freq_{R_2}(D)}, & sinon \end{cases} \quad (1)$$

Si $TC(D_{R_1|R_2}) > 1$, alors D est émergent. 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 (table 1). 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.

5 Annotation automatique

Notre objectif est d'obtenir un corpus entièrement annoté en registres, or nous ne disposons que de 3 269 textes annotés (graine). Notre approche consiste alors à augmenter l'ensemble de données

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

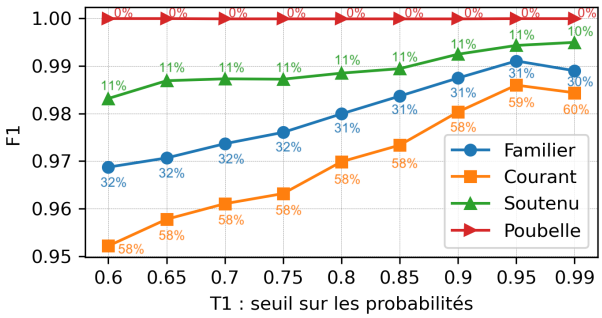

FIGURE 1 – $F1$ pour chaque registre et leur % dans l'ensemble des textes à ajouter à chaque valeur de $T1$

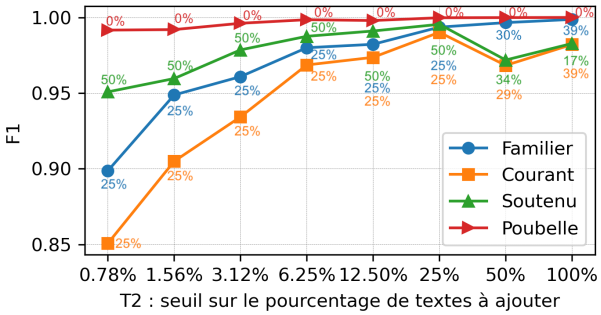

FIGURE 2 – $F1$ pour chaque registre et leur % de textes avec une probabilité ≥ 0.90 pour 1 des 4 registres à chaque valeur de $T2$

d'entraînement afin d'améliorer la qualité de prédiction finales des textes restants.

Méthodologie Tout d'abord, une phase de discrétisation des proportions est appliquée à chaque annotation A : $A \geq 50\% = 1$, et $A < 50\% = 0$. Puis, un modèle en deux étapes est adopté : l'apprentissage du classifieur multi-étiquettes est initialisé à partir de la graine. Un seuil est appliqué pour sélectionner les textes dont la prédiction est jugée fiable. Deux seuils indépendants sont introduits : soit sur la probabilité d'appartenance à un registre (noté $T1$), soit sur le nombre maximum de textes à ajouter (noté $T2$). Pour $T1$, tous les textes ayant une probabilité qui dépasse le seuil sont ajoutés. $T1$ garde la même valeur pour les quatre registres de langue. Pour $T2$, toutes les probabilités sont triées par ordre décroissant et pour chacun des 4 registres les $\frac{T2}{4}$ premiers textes sont ajoutés. Après avoir filtré les textes, le classifieur commence un second apprentissage à partir de la graine augmentée. Ses prédictions sont les prédictions finales. La *F-mesure* (notée $F1$) est utilisée pour évaluer les performances du modèle.

Expériences Pour les expériences, le modèle « base » de CamemBERT est utilisé comme classifieur (Martin *et al.*, 2020). 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. La figure 1 indique 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). De manière générale, le classifieur est sûr de lui : plus de la moitié du corpus est ajouté à la graine (76%) avec le seuil le plus strict : $T1 = 0,99$. Une échelle logarithmique est prise pour faire varier $T2$. Les meilleurs scores sont obtenus lorsque $T2$ est fixé à 25% (figure 2). La légère dégradation des $F1$ à partir de 25% peut être due à la baisse du pourcentage de textes ayant une probabilité ≥ 0.90 pour 1 des 4 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.

Analyse des résultats La distribution des registres est relativement similaire à celle de la graine annotée manuellement : 30,58% familier, 58,81% courant et 10,61% soutenu. Ces résultats suivent la tendance des annotations manuelles et indiquent la qualité des prédictions finales. De plus, pour les registres courant et soutenu, plusieurs traits émergents sont des *technomorphèmes* (table 2). Le registre courant montre un usage commercial des tweets qui met à profit la fonction d'indexation

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	6.7	2.2% / 0.3%	🌲 [ECOLOGIE] 🌳 À #Montréal,
ou de structuration du pictogramme*			🔴 #X banni de #Facebook
			📣 [Webinar] J-1 « Le bilan à 6 ans »
Phrase sans 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 2 – Top 3 des descripteurs (* : issus de notre analyse linguistique pour les CMO) qui caractérisent les registres dans le corpus annoté automatiquement. Chaque descripteur est donné avec son taux de croissance et fréquences relatives associées, ainsi qu'un exemple.

des hashtags afin de les rendre investigables, par exemple : « Le jeu *#MonstrumGame* de X sort ». Le phénomène d'agglutination pour des noms propres sert également à créer de nouveaux mots qui réfèrent à de nouveaux produits : « *#PokemonGO*, une nouvelle application ». Le registre soutenu quant à lui, intègre syntaxiquement les hashtags : « Les violences vécues en *#France* ne sont pas des *#incivilites* » ; et les pictogrammes comme du lexique : « les habitudes de #consommation en 🇫🇷 ». Enfin, le registre familier est utilisé pour dialoguer entre utilisateurs avec des marqueurs de l'oral : « Et *bim* dans tes dents », « *Eh* X il est timide *ou quoi* ? ». Ainsi, ces premières analyses confirment que les *technomorphèmes* ont été intégrés à la norme grammaticale et qu'ils ne sont plus uniquement caractéristiques du registre familier. Ils peuvent au contraire marquer un discours soutenu.

6 Conclusion

Dans cet article, nous avons présenté le corpus TREMoLo qui rassemble 228 505 tweets annotés en registres familier, courant et soutenu. Pour cela, une graine a été annotée manuellement en multiples étiquettes, en suivant un guide d'annotation issu d'une analyse linguistique du corpus. En utilisant le modèle CamemBERT, un enrichissement des données d'apprentissage a ensuite été réalisé afin finalement de prédire des registres pour l'ensemble du corpus. Nos expérimentations ont montré l'excellente qualité du corpus proposé. Suffisamment volumineux pour ouvrir la voie à de futurs travaux statistiques, il permettra de découvrir de nouvelles connaissances fondamentales sur les registres de langue.

Remerciements

Ce travail a bénéficié du soutien du projet TREMoLo [13] (ANR-16-CE23-0019) de l'Agence Nationale de la Recherche (ANR).

13. https://tremolo.irisa.fr/

Références

AGARWAL A., XIE B., VOVSHA I., RAMBOW O. & PASSONNEAU R. J. (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, p. 30–38.

BECCUCCI L. (2018). Pierre halté, les émoticônes et les interjections dans le tchat. limoges : Éditions lambert lucas, 2018. *Communication et organisation*, (54), 253–255.

BIBER D. (1991). *Variation across speech and writing*. Cambridge University Press.

BIBER D. & CONRAD S. (2019). *Register, genre, and style*. Cambridge University Press.

BOURQUIN G. (1965). Niveaux, aspects et registres de langage. remarques à propos de quelques problèmes théoriques et pédagogiques. *Linguistics*, **3**(13), 5–15.

FERGUSON C. A. (1982). Simplified registers and linguistic theory. *Exceptional language and linguistics*, p. 49–66.

GADET F. (1996). Niveaux de langue et variation intrinsèque. *Palimpsestes*, **10**.

GADET F. (2000). Français de référence et syntaxe. *Cahiers de l'Institut de Linguistique de Louvain*, **26**(1-4), 265–283.

GO A., BHAYANI R. & HUANG L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, **1**(12), 2009.

JACKIEWICZ A. & VIDAK M. (2014). Étude sur les mots-dièse. In *shs Web of Conferences*, volume 8, p. 2033–2050 : EDP Sciences.

JACQUES A. (1999). Internet, communication et langue française.

JOOS M. (1967). *The five clocks*, volume 58. New York : Harcourt, Brace & World.

LECORVÉ G., AYATS H., FOURNIER B., MEKKI J., CHEVELU J., BATTISTELLI D. & BÉCHET N. (2019). Towards the automatic processing of language registers : Semi-supervisedly built corpus and classifier for french.

MAGUÉ J.-P., ROSSI-GENSANE N. & HALTÉ P. (2020). De la segmentation dans les tweets : signes de ponctuation, connecteurs, émoticônes et émojis. *Corpus*, (20).

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

MCQUAIL D. (2010). *McQuail's mass communication theory*. Sage publications.

MEKKI J., BATTISTELLI D., LECORVÉ G. & BÉCHET N. (2018). Identification de descripteurs pour la caractérisation de registres. In *Proceedings of Rencontres Jeunes Chercheurs (RJC) of the CORIA-TALN conference*.

PAK A. & PAROUBEK P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, p. 1320–1326.

PAVEAU M.-A. (2013). Genre de discours et technologie discursive. tweet, twittécriture et twittérature. *Pratiques. Linguistique, littérature, didactique*, (157-158), 7–30.

PAVLICK E. & TETREAULT J. (2016). An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics*, **4**(1).

PETERSON K., HOHENSEE M. & XIA F. (2011). Email formality in the workplace : A case study on the enron corpus. In *Proceedings of the Workshop on Languages in Social Media*.

RAO S. & TETREAULT J. (2018). Dear sir or madam, may i introduce the gyafc dataset : Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv :1803.06535*.

REBOURCET S. (2008). Le français standard et la norme : l'histoire d'une «nationalisme linguistique et littéraire» à la française. *Communication, lettres et sciences du langage*, **2**(1), 107–118.

SANDERS C. (1993). *French today : language in its social context*. Cambridge University Press.

SHEIKHA F. A. & INKPEN D. (2010). Automatic classification of documents by formality. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*.

SINCLAIR J. (2005). Corpus and text-basic principles. *Developing linguistic corpora : A guide to good practice*, **92**, 1–16.

TODOROV T. (2013). *Mikhaïl Bakhtine. Le principe dialogique. Suivi de : Ecrits du Cercle de Bakhtine*. Le Seuil.

URE J. (1982). Introduction : approaches to the study of register range. *International Journal of the Sociology of Language*, **1982**(35).