Construction conjointe d'un corpus et d'un classifieur pour les registres de langue en français

Gwénolé Lecorvé[1] Hugo Ayats[1] Benoît Fournier[1] Jade Mekki[1, 2]
Jonathan Chevelu[1] Delphine Battistelli[2] Nicolas Béchet[3]
(1) Univ Rennes, CNRS, IRISA, 6, rue de Kerampont, 22305 Lannion Cedex, France
(2) Universite Paris-Ouest-Nanterre, MoDyCo, 200, avenue de la République 92001 Nanterre Cedex, France
(3) Universite de Bretagne Sud, IRISA, Campus de Tohannic, rue Yves Mainguy, 56017 Vannes Cedex, France
prenom.nom@irisa.fr, delphine.battistelli@u-paris10.fr

RÉSUMÉ

Les registres de langue sont un trait stylistique marquant dans l'appréciation d'un texte ou d'un discours. Cependant, il sont encore peu étudiés en traitement automatique des langues. Dans cet article, nous présentons une approche semi-supervisée permettant la construction conjointe d'un corpus de textes étiquetés en registres et d'un classifieur associé. Cette approche s'appuie sur un ensemble initial et restreint de données expertes. Via une collecte automatique et massive de pages web, l'approche procède par itérations en alternant l'apprentissage d'un classifieur intermédiaire et l'annotation de nouveaux textes pour augmenter le corpus étiqueté. Nous appliquons cette approche aux registres familier, courant et soutenu. À l'issue du processus de construction, le corpus étiqueté regroupe 800 000 textes et le classifieur, un réseau de neurones, présente un taux de bonne classification de 87 %.

ABSTRACT

Joint building of a corpus and a classifier for language registers in French.

Language registers are an observable stylistic trait of texts and speeches. However, they are still poorly studied in natural language processing. In this paper, we present a semi-supervised approach which jointly builds a corpus of texts labeled in registers and an associated classifier. This approach is based on an initial and limited set of expert data. Using an massive automatically retrieved collection of web pages, it iteratively proceeds by alternating the learning of an intermediate classifier and the annotation of new texts to augment the labeled corpus. We apply this approach to formal, neutral, and informal registers. At the end of the process, the labeled corpus gathers 800, 000 texts, and the classifier, a neural network, has an accuracy of 87 %.

MOTS-CLÉS : Registres de langue, apprentissage semi-supervisé, construction de corpus, classification automatique.

KEYWORDS: Language registers, classification, semi-supervised learning, corpus building.

1 Introduction

Les registres de langue fournissent de nombreuses informations sur un locuteur et sa relation avec les destinataires du message. Il s'agit d'un sujet cependant encore peu étudié en traitement automatique des langues (TAL), notamment en raison du manque de données d'apprentissage. Pour pallier ce

problème, cet article présente une approche semi-supervisée de construction d'un corpus textuel étiqueté en registres de langue.

L'approche proposée s'appuie sur un ensemble restreint de données manuellement étiquetées et une vaste collection de pages web automatiquement collectées mais non étiquetées. Le principe de construction tient alors dans l'apprentissage conjoint et itératif d'un classifieur, un réseau de neurones, sur les données étiquetées. Pour une itération donnée, le classifieur permet de catégoriser les données web, de sélectionner celles dont la classification semble fiable, puis de raffiner le classifieur sur la base de l'ensemble des données étiquetées augmenté de celles sélectionnées. Par ce procédé, nous visons une convergence de l'apprentissage du classifieur et de la construction du corpus vers un compromis entre taux de bonne classification et taille du corpus étiqueté. En pratique, nous appliquons ce processus sur un ensemble de $400\,000$ pages web et obtenons un corpus étiqueté en registres familier, courant et soutenu d'environ 750 millions de mots, ainsi qu'un réseau de neurones avec un taux de classification de $87\,\%$. Le jeu de descripteurs utilisé regroupe 46 caractéristiques de natures variées (lexicales, morphologiques, syntaxiques…) issues d'une analyse experte préalable.

Dans cet article, nous présentons tout d'abord en section 2 un état de l'art lié aux registres de langue et à leur traitement en TAL. Les sections 3, 4 et 5 introduisent ensuite les détails respectifs de notre approche, des données utilisées et du classifieur. Enfin, les résultats sont présentés en section 6.

2 État de l'art et positionnement

La notion de registre renvoie à la manière dont les productions linguistiques sont évaluées et catégorisées au sein d'une même communauté linguistique (celle du français par exemple) (Ure, 1982; Biber & Conrad, 2009). C'est ainsi que l'on distingue différents registres caractérisés selon de multiples traits spécifiques (termes plus ou moins complexes, ordre des mots, temps des verbes, longueur des phrases…) et souvent considérés sur une échelle de niveaux (par exemple, soutenu, littéraire, courant, familier, populaire, vulgaire…). Le partitionnement en catégories peut couvrir différents spectres selon la définition retenue de registre – le terme « registre » étant lui-même source de discussion – et traduire des finesses d'analyse variables (Sanders, 1993; Biber & Finegan, 1994; Gadet, 1996). Le sujet peut ainsi recouvrir, par exemple, l'influence du média de communication (Charaudeau, 1997) ou du degré de spécialisation (Borzeix & Fraenkel, 2005; Moirand, 2007) sur le discours. Dans notre travail, nous adoptons une vision plus traditionnelle avec un découpage en 3 registres : familier, courant et soutenu. Ce choix est avant tout motivé par le pragmatisme, ce découpage étant en effet relativement consensuel et peu sujet à ambiguïté pour l'étiquetage manuel d'un ensemble de données initial, tout en n'interdisant pas d'éventuels raffinements pour l'avenir. À défaut de caractérisation expérimentale détaillée – puisque c'est précisément l'objectif du projet dans lequel s'inscrit ce travail, nos 3 registres considérés se définissent par contraste vis-à-vis d'un emploi central, neutre, de la langue, c'est-à-dire la langue telle qu'employée lorsque le destinataire du message n'est pas connu. Pour assurer la bonne compréhension de ce message, cet emploi implique un ensemble minimal d'hypothèses quant aux connaissances spécifiques du destinataire et se calque donc sur la grammaire et le vocabulaire de la langue, sans pour autant en exploiter les tournures ou termes les plus rares. Ce périmètre d'usage définit le registre courant. Le registre soutenu peut alors être considéré comme l'ajout d'une hypothèse sur un haut degré de maîtrise de la langue de la part du lecteur ou interlocuteur. À l'inverse, le registre familier relâche les contraintes de respect de la norme en autorisant des écarts (volontaires ou fautifs) à différents niveaux (grammaire, vocabulaire mais aussi orthographe,

typographie…). Le registre familier fait alors également l'hypothèse d'une certaine compréhension de ces écarts comme autant de codes spécifiques. C'est à travers cette notion récurrente de connaissances partagées, et de donc de communauté, que les registres de langue s'enracinent dans le domaine de la sociolinguistique. Nous n'intégrons cependant pas cette dimension dans cet article.

À notre connaissance, les registres ont été peu étudiés en TAL, voire pas du tout sous l'angle que nous adoptons. Pour autant, tout un pan de travaux s'intéresse à l'utilisation du langage dans des situations particulières, cherchant à caractériser des « sous-langages », à les identifier, à les classer ou à les imiter. Ces sous-langages peuvent être portés par les notions de thème, de type documentaire, de style phonologique, de polarité en termes d'opinion, d'émotion… À notre connaissance cependant, aucun travail ne s'intéresse à la notion de registres de langue mais beaucoup de travaux en traitement de style apportent une base solide en termes de méthodologie et d'outils théoriques. Sans être exhaustif, l'étude des registres de langue partagent des similitudes avec ceux en attribution d'auteur (Stamatatos, 2009; Iqbal *et al.*, 2013), analyse des nouveaux médias (Schler *et al.*, 2006; Kobus *et al.*, 2008; Gianfortoni *et al.*, 2011; Eisenstein, 2013; Cougnon & Fairon, 2014). Différents corpus de référence ont d'ailleurs été publiés pour ces différents médias. Notre travail vise à combler le manque d'équivalent pour la notion de registre.

Les méthodes de traitement de style automatique sont toutes fondées sur un ensemble de descripteurs pertinents dérivés des textes à traiter. En raison de son importance historique, le travaux en attribution d'auteur permettent d'identifier un large éventail de descripteurs. Comme l'indique Stamatatos (2009), les préférences ou les choix d'écriture d'un auteur sont reflétés à plusieurs niveaux de langage. Le plus évident et le plus étudié est le niveau lexical, par exemple à travers la longueur des mots et des phrases d'un texte, la richesse de son vocabulaire ou les fréquences des mots et des n-grammes de mots (De Vel *et al.*, 2001; Sanderson & Guenter, 2006). À cet égard, il est généralement admis dans la communauté que les mots-outils (prépositions, articles, auxiliaires, verbes modaux…) sont d'intérêt notable alors que d'autres mots (noms, adjectifs…) doivent être évités pour le traitement de la style (Koppel & Schler, 2003; Argamon *et al.*, 2007), selon un principe d'orthogonalité entre le style et la signification d'un texte. Ce principe souligne l'importance d'abstraire certains éléments de sens pour l'analyse du style, faute de quoi l'analyse risque d'être biaisée par le thème des textes traités. Malgré tout, quelques descripteurs sémantiques peuvent se révêler utiles, par exemple les fréquences de recours à des synonymes et hyperonymes ou les relations fonctionnelles entre propositions (clarification d'une proposition par une autre, mise en oppositiondots) (McCarthy *et al.*, 2006; Argamon *et al.*, 2007). Par ailleurs, quelque soit leur sens, l'emploi de certains mots témoigne explicitement de l'appartenance du texte à un style précis (Tambouratzis *et al.*, 2004), en particulier dans le cas des registres de langue. Sur le plan syntaxique, l'emploi de descripteurs issus d'analyses morphosyntaxiques et syntaxiques est très largement répandu pour caractériser le style (Koppel & Schler, 2003; Hirst & Feiguina, 2007; Sidorov *et al.*, 2014). Enfin, d'autres travaux se sont intéressés à l'information graphémique en considérant des n-grammes de caractères, les types des graphèmes (lettre, chiffre, ponctuation, majusculedots) ou encore des mesures de compression de l'information (Koppel & Schler, 2003; Marton *et al.*, 2005; Escalante *et al.*, 2011). Dans notre travail, une étude linguistique préliminaire a été menée en ce sens (Mekki *et al.*, 2017, 2018), conduisant à un ensemble de descripteurs pour les 3 registres considérés.

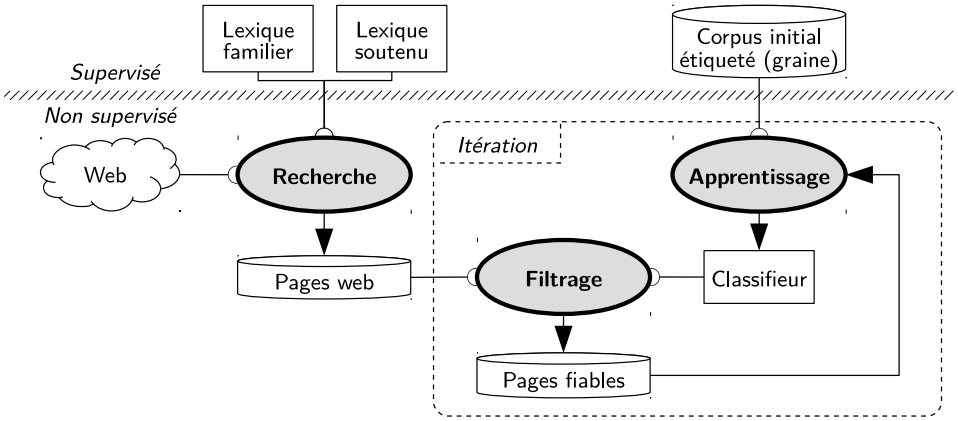


FIGURE 1 – Vue d'ensemble du processus semi-supervisé.

3 Approche proposée

Cette section décrit le processus semi-supervisé de construction d'un corpus étiqueté en registre. Comme illustré sur la figure 1, le processus est amorcé par une étape de collecte de données sur Internet. Cette collecte s'appuie sur deux lexiques spécialisés, l'un pour le registre familier, l'autre pour le soutenu, à partir desquels des requêtes familières ou soutenues sont formées, puis soumises à un moteur de recherche. Après nettoyage automatique, les pages récupérées sont regroupées au sein d'un unique corpus dont on cherche à extraire les plus pertinentes pour chaque registre. Cette extraction se fait par le biais d'un classifieur probabiliste (un réseau de neurones) prédisant la probabilité d'appartenance à chaque registre. Pour résoudre l'interdépendance selon laquelle le classifieur nécessite des données d'entraînement étiquetées et l'étiquetage des données nécessite un classifieur, l'approche procède par itérations. Ainsi, un premier classifieur est initialement entraîné sur une graine, c'est-à-dire un faible ensemble initial de données annotées manuellement et indépendant des pages web récupérés. Ce classifieur permet de sélectionner les textes dont l'appartenance à l'un des registres est considérée comme fiable, dans notre cas si la probabilité d'appartenance à un registre est supérieure à un seuil donné. Ces textes sont ensuite ajoutés à ceux déjà étiquetés, puis une nouvelle itération démarre. Ce processus semi-supervisé permet en fin de processus d'obtenir conjointement un ensemble de textes catégorisés et un classifieur. Notons que le recours à Internet n'est pas une originalité de notre travail puisque de nombreux exemples analogues existent dans littérature, par exemple (Baroni & Bernardini, 2004) (bien que notre processus de collecte ne soit pas itératif ici) ou encore (Lecorvé *et al.*, 2008) pour la collecte de pages thématiques.

Les classes considérées sont « familier », « soutenu » et « courant ». La considération du registre courant se justifie par le fait que les deux premiers registres se définissent par leurs variations respectives à ce troisième. Ainsi, le registre courant, parfois qualifié de neutre, rassemble les textes qui présentent peu d'écart à la norme. Pour compléter ce partitionnement des textes, nous faisons également l'hypothèse que certains textes récupérés n'appartiennent à aucun des 3 registres, soit car le texte est mal formé (langue étrangère, style SMS, texte non naturel…), soit car le registre n'y est pas homogène (par exemple, dans une liste de commentaires). Notre condition de fiabilité d'appartenance permet de modéliser cela.

Requêtes familières	Requêtes soutenues
Exemples positifs	
nain porte quoi	couronne de myrte
roublardise foutre la paix	dioscurisme argutieusement
croquenot se cuiter	hic et nunc
avaler sa chique	géronte séductible
Exemples négatifs	
montrer le chemin	relation sexuelle

TABLE 1 – Exemples de requêtes issues des lexiques familier et soutenu.

4 Données

En pratique, les lexiques sur lesquels s'appuie la collecte de pages web sont constitués de mots et expressions automatiquement récupérés à partir d'une sauvegarde de la version française de Wiktionary [1]. Pour un registre donné, seuls les mots sans ambiguïté d'appartenance à un registre sont considérés, c'est-à-dire les termes ayant toutes leurs acceptions annotées comme appartenant à un même registre. Précisément, les termes annotés comme argotiques, familiers, populaires et vulgaires ont été regroupés au sein du lexique familier et ceux catégorisés comme littéraires et soutenus au sein du lexique soutenu, chacun totalisant ainsi respectivement $6\,000$ et 500 entrées. Les requêtes sont construites registre par registre en combinant au hasard des éléments choisis du lexique associé. Le nombre de requêtes ainsi formées pour chaque lexique est identique afin d'aboutir à un ensemble des pages récupérées sensément équilibré en terme de registre. La longueur des requêtes est empiriquement limitée à un minimum de 2 mots et un maximum de 6 mots afin de garantir une pertinence minimale pour les pages retournées et un nombre de résultats non nul. Les requêtes web sont effectuée à l'aide de l'API Bing [2]. Au total, $12\,000$ requêtes sont soumises, chacune conduisant à un maximum de 50 pages Web. Quelques exemples de ces requêtes sont listés dans la table 1. Bien que certaines requêtes ne fassent *a priori* pas sens (par exemple, « croquenot se cuiter »), $76\,\%$ des requêtes renvoient au moins un résultat et $49\,\%$ en renvoient plus de 50, ces pourcentages étant comparables pour les requêtes familières et soutenues. Par ailleurs, la proportion de requêtes associées à tort à un registre (exemples négatifs dans table 1) est minime. Ces exemples sont en partie dus au fait que quiconque peut éditer Wiktionary, y compris des non-spécialistes. Enfin, signalons que certains dictionnaires en ligne apparaissant régulièrement dans les listes de résultats ont été exclus au moment de la requête afin de ne récolter que des pages où les termes recherchés sont bien en contexte et non isolées dans une définition ou un exemple.

Le contenu textuel des pages web est extrait automatiquement grâce à un outil de nettoyage [3]. Cet outil cherche le corps textuel de la page et ne s'intéresse qu'aux portions de texte « rédigées ». Il exclut ainsi les titres, menus, mentions légales, annonces, etc. mais inclut les commentaires si ceux-ci ont suffisamment de matière linguistique et se conforment au style rédactionnel normé (ponctuation, non abréviation des mots…). Enfin, pour éviter un manque d'homogénéité au sein de pages web longues (par exemple des forums) et de ne pas introduire de biais d'apprentissage liés aux disparités

1. http://fr.wiktionary.org
2. https://docs.microsoft.com/en-us/rest/api/cognitiveservices/bing-web-api-v5-reference
3. http://github.com/glecorve/web-cleaner

de longueur de textes, les textes nettoyés ont été segmentés sur les frontières de paragraphes de sorte à obtenir des segments de 5 000 caractères environ. À partir d'un total de 400 000 pages web et après filtrage des pages n'étant pas en français, le corpus de départ pour nos itérations consiste en environ 825 000 segments textuels, représentant 750 millions de mots. Un effet de ce découpage est d'atténuer l'hypothèse selon laquelle tous les textes contiennent au moins un terme très marqué en matière de registre. Cela apporte de la diversité au corpus mais pourrait également empêcher d'apprendre certaines corrélations entre ces indicateurs saillants et d'autres potentiellement plus discrets.

Enfin, notre ensemble de textes manuellement étiquetés rassemblent des segments issus de romans [4], journaux [5] et pages web. Ces pages web ne proviennent pas de l'ensemble collectés automatiquement pour la construction du corpus et elles ne contiennent ainsi pas nécessairement de termes listés dans nos lexiques spécialisés. Ce constat s'applique également aux textes provenant d'autres sources. L'étiquetage des pages s'est fait par 2 annotateurs sur la base des éléments de définition et de caractérisation relevés par notre étude linguistique préalable (différences entre registres, traits linguistiques à observer, exemples…). Au total, 435 segments textuels, soit environ 440 000 mots, sont considérés, équitablement répartis entre les registres familier, courant et soutenu.

5 Apprentissage du classifieur

Le classifieur s'appuie sur un ensemble de 46 caractéristiques listées par la table 2 et extraites automatiquement à partir de chaque texte. Celles-ci sont issue d'une expertise linguistique préliminaire (étude de l'état de l'art et analyse en corpus) dont les détails peuvent être trouvés dans (Mekki *et al.*, 2017) et (Mekki *et al.*, 2018). Elles couvrent de multiples niveaux d'abstraction de la langue, y compris des aspects liés à l'oral car le registre familier partagent des liens avec cette pratique de la langue (retranscription de certaines élisions de phonèmes, allongement de certaines syllabes…). Ces descripteurs sont tous des fréquences relatives globales à chaque texte (par exemple, le nombre de mots avec des répétitions de voyelles rapporté au nombre de mot dans le texte). Les ressources utilisées pour les descripteurs lexicaux ont été extraites de Wiktionnaire. Les analyses orthographiques et grammaticales (morphosyntaxe et syntaxe) ont été produites grâce à l'outil LangageTool [6]. Le reste du travail est réalisé par un ensemble de scripts Python *ad hoc*.

Diverses remarques sont à formuler concernant l'appartenance de certains mots ou expressions (plusieurs mots) au lexique d'un registre particulier. Tout d'abord, notons qu'aucun lexique du registre courant n'existe. Ensuite, certains mots peuvent être ambigus quant à leur appartenance à un registre, en fonction de leur contexte d'usage. Par exemple, le mot « caisse » peut, certes, faire référence à une voiture en argot mais il peut également porter le simple sens d'un contenant. Ainsi, deux variantes de descripteurs sont considérés pour les fréquences de mots propres à un registre. La première pondère la fréquence d'un mot par le nombre d'acceptions identifiées comme appartenant au registre considéré divisée par le nombre total de ses acceptions. Dans notre exemple, l'observation du mot « caisse » ne compter que pour moitié. L'autre variante est plus stricte. Elle ne comptabilise un mot que si toutes ses acceptions sont identifiées comme appartenant au registre. Le cas des expressions ne nécessite pas cette dualité car celles-ci sont généralement moins ambiguës. Enfin, nous soulignons que la richesse lexicale du registre familier est bien plus grande que celle du registre soutenu. Il s'agit d'un

4. Parmi lesquels Kiffe kiffe demain de Faïza Guène, Albertine disparue de Marcel Proust, Les Mohicans de Paris d'Alexandre Dumas, Les bâtiments de ponts de Rudyard Kipling, Les misérables de Victor Hugo…

5. Une sélection d'articles de L'Humanité.

6. https://languagetool.org/

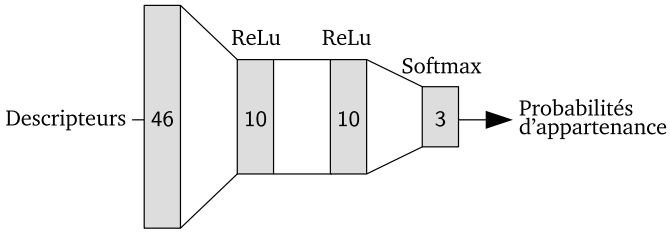


FIGURE 2 – Architecture du réseau de neurones.

phénomème bien connu ayant trait au fait qu'il n'existe qu'une norme du langage mais une infinité de s'en écarter. La richesse du registre familier reflète ces multiples écarts possibles.

Le classifieur est un réseau de neurones multi-couches. Le choix de cet outil d'apprentissage automatique n'est pas une revendication de notre travail. Ce choix se justifie avant tout par la facilité actuelle à construire des réseaux de neurones grâce aux multiples boîtes à outils disponibles. Par ailleurs, les possibilités d'interconnexions entre neurones et les multiples fonctions d'activation existantes permettent de modéliser par des réseaux de neurones d'autres techniques comme des classifieurs naïfs de Bayes ou des modèles de type exponentiel. Enfin, les réseaux de neurones sont connus pour être liés à la propriété de prolongement (Mikolov *et al.*, 2013; Le & Mikolov, 2014). Bien que l'objectif du présent article soit la construction d'un corpus et d'un premier classifieur. Des perspectives futures pourraient être d'observer les similarités entre documents tels présentes dans l'espaces des embeddings produits par notre modèle. Tel que l'illustre la figure 2, le réseau de neurones que nous considérons prend en entrée le vecteur des 46 valeurs représentant un texte. Les valeurs en sortie sont les probabilités d'appartenance à chaque registre. Toutes les couches du réseau sont des couches denses. Les 2 premières sont composées de 10 neurones, la première avec une fonction d'activation de type *leaky ReLU* [7], l'autre avec la fonction *tanh*. La dernière couche est composée de 3 neurones avec une fonction *softmax* afin de produire une distribution de probabilités. Cette architecture est issue de quelques tests sur un ensemble de développement mais n'a pas fait l'objet d'une étude approfondie. Une fois le réseau appris, les probabilités d'appartenance pour un texte fourni en entrée sont directement interprétées comme le niveau de confiance du réseau. Un seuil est alors appliqué pour déterminer s'il faut classer le texte ou non.

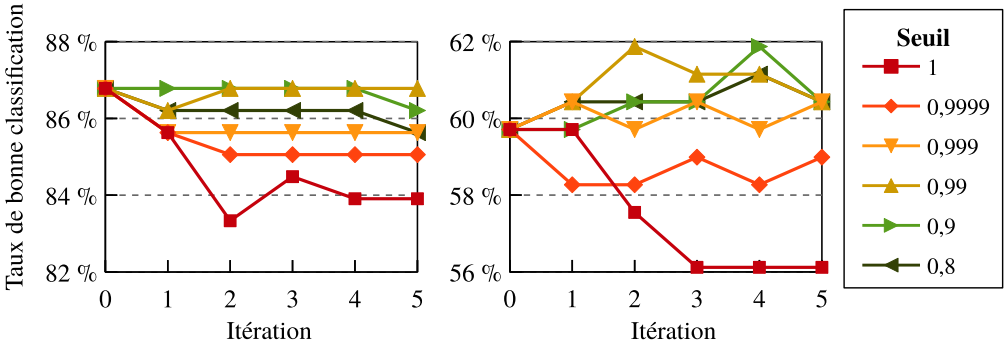
6 Résultats

Les expériences ont été menées en utilisant les bibliothèques Keras [8] et TensorFlow [9]. Hormis lors de l'apprentissage du premier modèle sur la graine, les classifieurs successifs sont appris par lot de 100 instances sur 20 époques en utilisant l'algorithme d'optimisation *rmsprop* et l'erreur absolue moyenne comme fonction objectif. Les 435 segments initialement annotés sont répartis en un ensemble d'apprentissage (40%, soit 174 segments), de développement (20%) et de test (40%). À chaque itération, les segments nouvellement sélectionnés parmi les données web sont injectés dans l'ensemble d'apprentissage pour 80% et l'ensemble de développement pour le reste. L'ensemble de test n'est jamais modifié afin de pouvoir mesurer l'évolution du classifieur tout au long du processus.

7. Paramètre α fixé à $0,1$.
8. `https://keras.io/`
9. `https://www.tensorflow.org/`

Lexique
– Mots familiers pondérés par leur nombre d'acceptions familières : 7 828 éléments
– Mots soutenus pondérés par leur nombre d'acceptions soutenues : 565 éléments
– Mots strictement familiers (toutes les acceptions sont familières) : 3 075 éléments
– Mots strictement soutenus (toutes les acceptions sont soutenues) : 166 éléments
– Expressions familières : 3 453 éléments
– Expressions soutenues : 143 éléments
– Noms d'animaux : 78 éléments
– Onomatopées (« ah », « pff »…) : 125 éléments
– Termes du langage SMS (« slt », « lol », « tkt »…) : 540 éléments
– Anglicisme (lexique et syntaxe)
– Mots inconnus
– Emploi de « ça »
– Emploi de « ce »
– Emploi de « cela »
– Emploi de « des fois »
– Emploi de « là »
– Emploi de « parfois »
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 »
– Emploi de mots terminant en « -iotte »
– Emploi de mots terminant en « -o »
– Emploi de mots terminant en « -ou »
– Emploi de mots terminant en « -ouze »
Morphosyntaxe
– Emplois des temps : impératif présent, indicatif futur, indicatif imparfait, indicatif passé simple, indicatif présent, conditionnel présent, subjonctif imparfait, subjonctif présent
– 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 « est-ce que »
– Emploi de la conjonction « et »
– Négations sans « ne »
– Autres fautes de syntaxe

TABLE 2 – Liste des descripteurs utilisés par le classifieur



(a) Test **(b) Pages collectées annotées**

FIGURE 3 – Taux de bonne classification pour chaque itération sur le test (a) et le sous-ensemble annoté des pages récupérées (b).

Par ailleurs, un sous-ensemble des 139 pages web collectées a été tiré aléatoirement et annoté manuellement. Ces pages proviennent équitablement des requêtes familières et soutenues. Parmi ces pages, 27 sont étiquetées pour le registre familier (19 % des pages), 69 pour le registre courant (50 %), 38 pour le registre soutenu (27 %) et 5 comme non étiquetables (4 %) car équivoques [10]. Ce deuxième ensemble est complémentaire de l'ensemble de test car il est constitué de pages qui contiennent des mots connus comme appartenant à un registre, hypothèse absente pour l'ensemble de test. De plus, nous pouvons dores et déjà souligner que les proportions respectives de chaque registre diffèrent entre l'ensemble de test et le sous-ensemble annoté des pages web collectées.

Nous étudions tout d'abord les résultats du classifieur, puis le corpus produit en parallèle.

6.1 Classifieur

La figure 3 montre l'évolution du taux de bonne classification du modèle au fur et à mesure des itérations sur l'ensemble de test (a) et sur le sous-ensemble annoté des pages collectés. Les résultats sont présentés pour différentes valeurs du seuil de sélection des pages, allant de $0,8$ (probabilité d'appartenance) à 1 (c.-à-d. que le classifieur est sûr de lui). Ces valeurs élevées se justifient par le taux élevé de bonne classification de 87 % dès l'initialisation du processus. Sur l'ensemble de test, nous pouvons constater que le classifieur est très stable en dépit des apports en nouvelles données, quelque soit l'ensemble de données. Cela semble signifier que ces nouvelles données sont cohérentes avec notre graine mais qu'elle n'apporte pas d'éléments supplémentaires permettant d'améliorer les performances. Parmi les seuils testés, la consigne de sélection la plus stricte (seuil = 1) conduit à une nette dégradation des résultats au cours du processus. Les seuil $0,9$ et $0,99$ produisent les meilleurs résultats. Sur le sous-ensemble des pages collectées, les résultats sont nettement moindres, bien que toujours largement au-dessus d'une classification aléatoire ou naïve [11]. Cette difficulté accrue sur les données collectées automatiquement provient probablement d'éléments que le nettoyage automatique

10. Souvent à cause d'un mélange de registres entre des propos rapportés et des passage de narration.

11. Taux de bonne classification de 36 % dans le cas d'un tirage aléatoire informé sur la distribution des classe, 50 % dans le cas du vote majoritaire (classe « courant »).

	(a) Test			(b) Pages collectées annotées		
	Familier	Courant	Soutenu	Familier	Courant	Soutenu
Rappel	$0,90$	$0,78$	$0,93$	$0,53$	$0,72$	$0,45$
Précision	$0,84$	$0,90$	$0,87$	$0,52$	$0,64$	$0,61$
F-mesure	$0,87$	$0,83$	$0,90$	$0,52$	$0,68$	$0,52$

TABLE 3 – Rappel, précision et F-mesure pour chaque registre en fin de processus (seuil = $0,99$) sur l'ensemble de test (a) et sur le sous-ensemble manuellement annoté des pages collectées (b).

n'a pas réussi à enlever [12]. Les conclusions générale sur l'impact des différents seuils sont identiques. En complément, il est intéressant de noter que le seuil de $0,99$ produit une augmentation du taux de bonne classification lors des 2 premières étapes de sélection, puis baisse progressivement. Ce comportement pose la question de la détermination automatique d'un critère d'arrêt des itérations. En l'état, des analyses plus approfondies sont nécessaires pour mieux comprendre les phénomènes observés et élaborer un critère de qualité globale du corpus étiqueté après chaque étape.

La table 3 présente les taux de rappel et précision ainsi que la F-mesure en fin de processus pour le seuil $0,99$, sur l'ensemble de test et sur notre extrait annoté des pages collectées. Sur le premier ensemble, il apparaît que ces mesures sont relativement homogènes entre registres. La F-mesure la plus basse est celle du registre courant. Elle s'explique notamment par un rappel plus bas que pour les autres registres. Les résultats sont à l'inverse sur l'extrait annoté des pages web puisque ils sont globalement beaucoup plus faibles (conformément aux résultats de la figure 3) et que le registre courant est celui le mieux reconnu. Les registres familier et soutenu présentent eux une F-mesure très faible mais pour des raisons différentes. Pour le premier, le modèle semble retourner avoir des difficultés globales avec un rappel une précision à peine supérieurs à la moyenne alors que, pour le second, les faibles résultats semblent davantage liés à un forte proportion de faux négatifs (rappel faible). Dans l'optique d'une utilisation fiable du corpus construit, il apparaît donc nécessaire d'améliorer encore ces résultats. Pour cela, une attention particulière devra notamment être porté sur la limitation des faux positifs car ceux-ci tendent probablement à freiner ou fausser la convergence du processus semi-supervisé.

6.2 Corpus étiqueté automatiquement

Les figures 4 et 5 illustrent la construction du corpus étiqueté. La première présente l'évolution de la taille de ce corpus pour les différents seuils étudiés, la seconde la proportion de chaque registre dans celui-ci dans l'unique meilleur cas d'un seuil de sélection fixé à $0,99$. En terme de taille, il apparaît, d'une part, que la totalité ou quasi totalité des pages collectées termine le processus avec une étiquette. Étant donné le bruit déjà évoqué dans les données, ce constat semble à nouveau indiquer la nécessité d'un critère d'arrêt du processus. D'autre part, il apparaît que l'étiquetage intégral des données se produit rapidement, c'est-à-dire en peu d'itérations. Par exemple, 89% des étiquettes sont validées à l'issue de la première passe pour un seuil de sélection de $0,8$. Ceci témoigne de la grande confiance du modèle dans ses prédictions. Nous pensons que ceci peut s'expliquer par une importance trop grande donnée à certaines descripteurs (par exemple, l'apparition ou l'absence de termes d'un lexique spécifique) pour prédire un registre. Une solution pourrait être d'introduire un mécanisme d'abandon

12. Par exemple, dans le cas de forums où de nombreux éléments textuels sont à supprimer pour n'isoler que le corps des réponses.

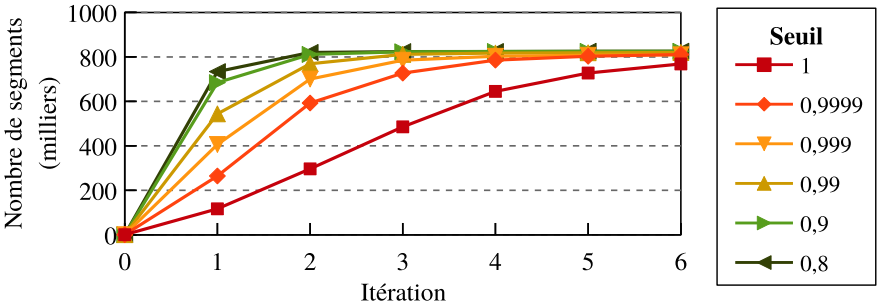


FIGURE 4 – Taille du corpus pour chaque itération.

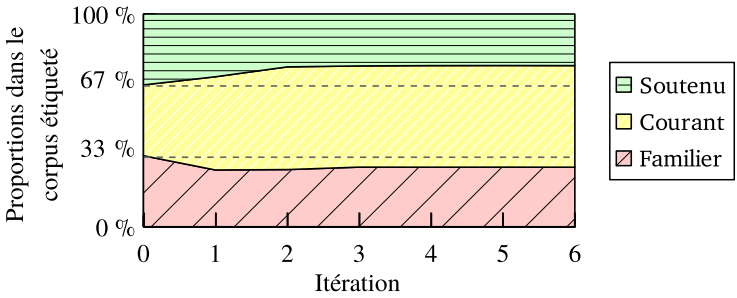


FIGURE 5 – Évolution de la proportion de chaque registre dans le corpus étiqueté (pourcentage sur le nombre de textes, seuil = $0,99$).

(*dropout*) lors de l'apprentissage du réseau de neurones pour l'amener à des prédictions s'appuyant sur un spectre plus large d'informations. De premières expériences ont été conduites dans cette direction mais il apparaît que cette stratégie peut conduire à des dégradation du taux de classification lorque le taux d'abandon des informations est mal configuré. Enfin, l'évolution de la répartition des classes est également intéressante à observer puisque nous avons montré à travers l'annotation de quelques pages collectées que la répartition des registres diffère entre notre graine, volontairement équilibrée, et les pages collectée, avec un forte dominance du registre courant. La figure 5 montre que l'étiquetage semi-supervisée de notre approche corrige de lui-même cette différence. Dans le corpus final, le registre courant représente $48\,\%$ des étiquettes, contre respectivement $28\,\%$ et $24\,\%$ pour les registres familier et soutenu. Ces nombres coïncident globalement avec ceux de notre étiquetage manuel sur un petit échantillon aléatoire.

En fin de processus pour le seuil de $0,99$, les textes annotés comme familiers viennent pour $68\,\%$ des requêtes construites sur le lexique familier et, donc, pour $32\,\%$ de celles sur le lexique soutenu. Ces rapports sont respectivement de $47\,\%$ / $53\,\%$ et $37\,\%$ / $63\,\%$ pour le corpus des registres courant et soutenu. La table 4 montre ainsi 2 extraits de pages issues de requêtes soutenues mais l'une ayant été étiquetée comme du registre courant et l'autre soutenu. D'une part, ces proportions montrent que les requêtes formées à partir des lexiques facilitent la construction du corpus par un amorçage approprié du processus puisqu'elles n'enferment pas les pages récoltées dans le registre de leur requête d'origine. Une conclusion intéressante est donc que la présence de termes discriminants pour un registre n'est pas un indice suffisant pour catégoriser un texte dans ledit registre. En cela, cette

Courant	Soutenu
Oui, Monsieur Adrien Richard, si vous aimez mieux, le directeur de l'usine, mais nous, nous ne l'appelons que Monsieur Adrien, parce qu'on a été à l'école ensemble et qu'il nous appelle aussi par notre prénom.	D'ailleurs, nous retrouvons la même distinction dédaigneuse à l'égard des professionnels et de leur " vil salaire " qui ne les empêche pas de mourir " ès hôpitaux ", chez le docte Muret.

TABLE 4 – Rappel, précision et F-mesure pour chaque registre en fin de processus (seuil = $0,99$) sur l'ensemble de test (a) et sur le sous-ensemble manuellement annoté des pages collectées (b).

souplesse justifie également le recours à une classification des pages récoltées. D'autre part, l'analyse montre que certains phénomènes sont encore mal compris et que la méthode devrait être affinée. En particulier, il apparaît que la plupart des textes classés comme familiers à partir de requêtes soutenues (et réciproquement) ne devrait pas l'être. Hormis une règle stricte qui interdirait ces situations, il semble nécessaire d'observer les traits qui contribuent à ces erreurs et d'affiner la méthode actuel, par exemple en considérant des descripteurs complémentaires ou plus précis. Par exemple, il apparaît que les mots vulgaires, confondus dans l'ensemble des termes familiers, ont un rôle ambigü. De même, les fréquences de ponctuation ou la longueur des phrases devraient être prises en compte.

7 Conclusion

Dans cet article, nous avons présenté un processus semi-supervisé qui construit conjointement un corpus textuel étiqueté en registres de langue et un classifieur associé. En s'appuyant sur un très large ensemble de textes et quelques ressources expertes de départ, le résultat de cette approche est un corpus constitué de $800\,000$ segments textuels représentant un total d'environ 750 millions de mots. Le classifieur parallèlement obtenu atteint un taux de bonne classification de $87\,\%$ sur l'ensemble de test mais des résultats plus modestes sur un sous-ensemble étiqueté manuellement des pages collectées. Ces résultats semblent démontrer la validité de l'approche et d'une majorité des annotations produites mais ils démontrent également le besoin d'affiner ses différents aspects.

Parmi les pistes de travail pour l'avenir, une analyse linguistique poussée sur la qualité des annotations automatiques doit être poursuivie, ainsi qu'une étude des importances de chaque descripteur dans le réseau de neurones et de la robustesse du réseau si certains descripteurs venaient à être absents ou anormaux. Par ailleurs, le choix du seuil de sélection ne semble pas critique ni même réellement propice à éviter la présence trop importante de faux positifs. Il serait instructif de comprendre pourquoi et d'essayer d'autres stratégies (par exemple, en limitant la sélection des textes à un nombre fixe par itération). À plus long terme, le corpus ouvre de multiples pistes d'utilisation, comme, dans le cas qui nous intéressent, la transposition automatique d'un texte d'un registre vers un autre.

Remerciements

Ce travail a bénéficié du soutien financier de l'Agence Nationale de la Recherche (ANR) dans le cadre du projet TREMoLo (ANR-16-CE23-0019).

Références

ARGAMON S., WHITELAW C., CHASE P., HOTA S. R., GARG N. & LEVITAN S. (2007). Stylistic text classification using functional lexical features. *Journal of the Association for Information Science and Technology*, **58**(6), 802–822.

BARONI M. & BERNARDINI S. (2004). Bootcat : Bootstrapping corpora and terms from the web. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, p. 1313–1316.

BIBER D. & CONRAD S. (2009). *Register, genre, and style*. Cambridge University Press.

BIBER D. & FINEGAN E. (1994). *Sociolinguistic perspectives on register*. Oxford University Press on Demand.

BORZEIX A. & FRAENKEL B. (2005). *Langage et travail (communication, cognition, action)*. CNRS éd.

CHARAUDEAU P. (1997). *Le discours d'information médiatique : la construction du miroir social*. Nathan.

COUGNON L.-A. & FAIRON C. (2014). *SMS Communication : A linguistic approach*, volume 61. John Benjamins Publishing Company.

DE VEL O., ANDERSON A., CORNEY M. & MOHAY G. (2001). Mining e-mail content for author identification forensics. *ACM Sigmod Record*, **30**(4), 55–64.

EISENSTEIN J. (2013). What to do about bad language on the internet. In *Proceedings of North American Chapter of the Association for Computational Linguistics : Human Language Technologies (HLT-NAACL)*, p. 359–369.

ESCALANTE H. J., SOLORIO T. & MONTES-Y GÓMEZ M. (2011). Local histograms of character n-grams for authorship attribution. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (HTL-ACL)*, p. 288–298 : Association for Computational Linguistics.

GADET F. (1996). Niveaux de langue et variation intrinsèque. *Palimpsestes*, **10**, 17–40.

GIANFORTONI P., ADAMSON D. & ROSÉ C. P. (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*, p. 49–59 : Association for Computational Linguistics.

HIRST G. & FEIGUINA O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, **22**(4), 405–417.

IQBAL F., BINSALLEEH H., FUNG B. C. & DEBBABI M. (2013). A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, **231**, 98–112.

KOBUS C., YVON F. & DAMNATI G. (2008). Normalizing sms : are two metaphors better than one ? In *Proceedings of the International Conference on Computational Linguistics (COLING)*, p. 441–448 : Association for Computational Linguistics.

KOPPEL M. & SCHLER J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, p. 72–80.

LE Q. & MIKOLOV T. (2014). Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning (ICML)*, p. 1188–1196.

LECORVÉ G., GRAVIER G. & SÉBILLOT P. (2008). On the use of web resources and natural language processing techniques to improve automatic speech recognition systems. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, p. 592–599.

MARTON Y., WU N. & HELLERSTEIN L. (2005). On compression-based text classification. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, volume 3408, p. 300–314 : Springer.

MCCARTHY P. M., LEWIS G. A., DUFTY D. F. & MCNAMARA D. S. (2006). Analyzing writing styles with coh-metrix. In *Proceedings of the FLAIRS Conference*, p. 764–769.

MEKKI J., BATTISTELLI D., BÉCHET N. & LECORVÉ G. (2017). *« Nous nous arrachâmes promptement avec ma caisse » : quels descripteurs linguistiques caractérisent les registres de langue ?* Technical report, IRISA, équipe EXPRESSION ; MoDyCo.

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

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (HLT-NAACL)*, p. 746–751.

MOIRAND S. (2007). *Les discours de la presse quotidienne. Observer, analyser, comprendre*. Puf.

SANDERS C. (1993). *Sociosituational variation*. Cambridge : Cambridge University Press.

SANDERSON C. & GUENTER S. (2006). Short text authorship attribution via sequence kernels, markov chains and author unmasking : An investigation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 482–491 : Association for Computational Linguistics.

SCHLER J., KOPPEL M., ARGAMON S. & PENNEBAKER J. W. (2006). Effects of age and gender on blogging. In *Proceedings of the AAAI spring symposium : Computational approaches to analyzing weblogs*, volume 6, p. 199–205.

SIDOROV G., VELASQUEZ F., STAMATATOS E., GELBUKH A. & CHANONA-HERNÁNDEZ L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, **41**(3), 853–860.

STAMATATOS E. (2009). A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, **60**(3), 538–556.

TAMBOURATZIS G., MARKANTONATOU S., HAIRETAKIS N., VASSILIOU M., CARAYANNIS G. & TAMBOURATZIS D. (2004). Discriminating the registers and styles in the modern greek language-part 2 : Extending the feature vector to optimize author discrimination. *Literary and Linguistic Computing*, **19**(2), 221–242.

URE J. (1982). Introduction : approaches to the study of register range. *International Journal of the Sociology of Language*, **1982**(35), 5–24.